Graph approximations to geodesics on embedded manifolds

Mira Bernstein, Vin de Silva,
John C. Langford, Joshua B. Tenenbaum

July 5, 2002

Abstract

We analyze the problem of extracting a low-dimensional manifold from a high dimensional space and (incidentally) prove low-distortion low-dimension embedding by the Isomap algorithm [5] when 1) samples form a cover on an embedded low-dimensional Euclidean manifold, 2) the embedded manifold has a large branch separation, and 3) a large radius of curvature.

1 Introduction

Given a set of points in a high dimensional space, how can we extract a mapping from the high dimensional space to a low dimensional space which preserves the relative geometry of the points? In other words, how can we extract a low-distortion low-dimensional representation of points?

Some special case algorithms for extracting one or two dimensional surfaces from two or three dimensional spaces have been analyzed [1], [2] with sufficient conditions dependent upon a notion of local feature size. The analysis here covers the more general case of extracting a low dimensional space from a high dimensional space.

This problem has been studied and various approaches exist. The approaches can be roughly sorted by strength and types of assumptions which are necessary in order to successfully extract a low-distortion low-dimensional representation.

On one end of the spectrum are the assumption free approaches of which the Johnson-Lindenstrauss lemma [9] is typical. The implication of this lemma is that m points in a high dimensional Euclidean space can be mapped into an $O(\log m/\epsilon^2)$ Euclidean space in such a way that distances are distorted by at most a factor of $1 \pm \epsilon$. This approach is general and can be applied to other metrics and other notions of optimality. See [10] for a recent survey.

The principle inadequacy of this assumption-free approach is that $O(\log m/\epsilon^2)$ is simply too large for many practical uses. In many cases we really do have reason to believe that some low-dimensional structure exists and want to extract. Successful extraction is important for human understanding of the dataset and because many geometric algorithms apply to only 2 or 3 dimensional spaces.

A much more restrictive notion comes from principle component analysis (PCA) which can extract a low-distortion low-dimensional representation when the points are on (or near) a low dimensional affine subspace in the high dimensional space. The largest drawback of PCA is that the assumption that there exists a low dimensional affine subspace is too restrictive. In particular, there are many situations where the points might lie on a low dimensional manifold which is *nonlinearly* embedded in the high dimensional space.

Problems of nonlinear dimensionality reduction arise in many applications with natural signals, including computer vision, speech recognition, astrophysics, climatology, and genetics. Consider the following high-dimensional spaces of visual images (represented as vectors of raw pixel brightness values):

1. Viewpoints of a scene.
2. images of hand written letters.
3. images of a hand in various positions.

For each of these examples, a low dimensional nonlinearly embedded manifold contains the points. In the first example, the dimensionality is 3 (left-right/up-down/near-far) and there may be an extra 3 dimensions due to rotations at a fixed point of view. The intrinsic dimensionality of hand-written letters and hand positions is not so easy to specify a priori, but they are still characterizable as contained in a d-dimensional manifold which is nonlinearly embedded within $\mathbb{R}^n$ where n is the number of pixels in an image and $d \ll n$. PCA will inherently perform poorly on these examples because of the nonlinearity. Assumption-free approaches will not perform well because $O(\log m/\epsilon^2)$ is often similar to n in practice which implies that little complexity is removed.

There are two general nonlinear manifold extraction algorithms which have recently been proposed, Isomap [5] and LLE [11]. Both of these algorithms behave well on datasets of the above form. It is not clear when and how these algorithms will fail. The assumption free results imply they *must* fail in some cases. In what cases will these algorithms work? Where will they fail?

An ideal result would have the form: "Natural assumption A and B are necessary and sufficient conditions for manifold extraction up to error tolerance ϵ." We have achieved half of this: sufficient conditions for extraction with the Isomap algorithm given some natural assumptions. A similar result is not known for the LLE algorithm.

Isomap works in 3 simple stages starting with a set of points in $\mathbb{R}^n$ (for n typically large):

1. Construct a graph G where every point is connected to every nearby[1] point.

[1] The exact algorithm used was either "within some ϵ" or "K nearest neighbors" in the high

2. Compute all points shortest path on the graph G.
3. Use MDS (MultiDimensional Scaling) to embed this graph into some $\mathbb{R}^d$ for d small.

The output of Isomap is an embedding of the input points in $\mathbb{R}^d$ for d small (typically 2, 3, or 4). The essential goal of this paper is analyzing when the Isomap algorithm works. Roughly, we show that if:

1. The input points have a small cover scale (defined later) for a Euclidean manifold M,
2. The radius of curvature for the embedding of the manifold M into $\mathbb{R}^d$ is not too small,
3. The branch separation (distance between different "arms" of the manifold) is large enough

then the distance between two points in $\mathbb{R}^d$ will be near to the *geodesic* distance (in M). As the scale of the covering approaches 0, the distortion goes to 0.

The paper is organized into 4 sections.

- In Section 2 we state formal definitions.
- In Section 3 we show half of a two-step convergence.
- In Section 4 we show the other half of two-step convergence and state the main theorem.
- In Section 5 we discuss the relationship of the results here with the Isomap algorithm.

2 Finite samples on submanifolds

Let $M = M^d$ be a compact d-dimensional smooth submanifold of the Euclidean space $\mathbb{R}^n$. Boundary and boundary corners are permitted, so for instance M may be a d-dimensional cube.

The natural Riemannian structure on M (induced from the Euclidean metric on $\mathbb{R}^n$) gives rise to a manifold metric d_M defined by:

$$d_M(x, y) = \inf_{\gamma}\{\text{length}(\gamma)\}$$

where γ varies over the set of (piecewise) smooth arcs connecting x to y in M. Note that $d_M(x, y)$ is generally different from the Euclidean distance $\|x - y\|$.

Let $\{x_i\} \subset M$ be a finite set, whose elements we will refer to as "data points" or "sample points". These points may be chosen randomly, or obtained in some other manner. The Isomap algorithm attempts to recover the manifold distances $d_M(x_i, x_j)$, given only the data points $\{x_i\} \subset \mathbb{R}^n$. Of course, this can only be done approximately.

dimensional Euclidean space.

The construction makes use of a graph G on the data points. Given such a graph we can define two further metrics, just on the set of data points. Let x, y belong to the set $\{x_i\}$. We define:

$$
\begin{aligned}
d_G(x,y) &= \min_P \left(\|x_0 - x_1\| + \ldots + \|x_{p-1} - x_p\|\right) \\
d_S(x,y) &= \min_P \left(d_M(x_0, x_1) + \ldots + d_M(x_{p-1}, x_p)\right)
\end{aligned}
$$

where $P = (x_0, \ldots, x_p)$ varies over all paths along the edges of G connecting x ($\equiv x_0$) to y ($\equiv x_p$).

Given the data points and graph G, one can compute d_G without knowledge of the manifold M. This is a key stage in the Isomap algorithm. Since the real goal is to estimate d_M, we must show that d_G is a good approximation to d_M, under favorable circumstances.

The metric d_S is an intermediate approximation. The proof that $d_G \approx d_M$ falls naturally into two parts: $d_M \approx d_S$ and $d_S \approx d_G$.

Proposition 1. We have the inequalities:

$$
\begin{aligned}
d_M(x,y) &\leq d_S(x,y) \\
d_G(x,y) &\leq d_S(x,y)
\end{aligned}
$$

Proof. The first expression is just the triangle inequality for the metric d_M. The second inequality holds because for the path $P = x_0, \ldots, x_p$ achieving the minimum in the definition of d_S, we have:

$$
d_S(x,y) \geq \|x_0 - x_1\| + \ldots + \|x_{p-1} - x_p\| \geq d_G(x,y)
$$

since $d_M(x_i, x_{i+1}) \geq \|x_i - x_{i+1}\|$ for the left inequality, and the right inequality follows by the minimum in d_G. $\square$

The main task will be to show that d_S is not too much bigger than d_M, and that d_G is not too much smaller than d_S.

3 The sampling condition

There are simple conditions on the data set $\{x_i\}$ and the graph G which guarantee that d_S is a good approximation to d_M.

One parameter of the proof is the covering scale of the samples, $\{x_i\}$.

$$
C(d_M, \{x_i\}) = \max_{x \in M} \min_{\{x_i\}} d_M(x_i, x)
$$

The covering scale can be thought of as an inversion of the covering number often used in learning theory in the context of VC bounds [12].

Theorem 2. *For all $\epsilon > 0$ and $\delta > 0$, with $4\delta < \epsilon$. If:*

1. *The graph G contains all edges xy for which $d_M(x,y) \leq \epsilon$.*
2. *The covering scale is small, $C(d_M, \{x_i\}) \leq \delta$*

Then for all pairs of data points x, y we have

$$d_M(x,y) \leq d_S(x,y) \leq (1+4\delta/\epsilon)d_M(x,y)$$

Proof. Let γ be any piecewise-smooth arc connecting x to y, with length $\ell = \text{length}(\gamma)$. We will find a path from x to y along edges of G whose length $d_M(x_0,x_1) + \ldots + d_M(x_{p-1},x_p)$ is less than $(1+4\delta/\epsilon)\ell$. The right-hand inequality will follow upon taking the infimum over γ.

If $\ell \leq \epsilon - 2\delta$ then x and y are connected by an edge, which we can use as our path.

If $\ell > \epsilon - 2\delta$ we can write $\ell = \ell_0 + (\ell_1 + \ell_1 + \ldots + \ell_1) + \ell_0$, where $\ell_1 = \epsilon - 2\delta$ and $\epsilon - 2\delta \geq \ell_0 \geq (\epsilon - 2\delta)/2$. (There may be no ℓ_1 terms if ℓ is small.)

Now cut up the arc γ into pieces in accordance with this decomposition of its length ℓ. This gives us a sequence of points $\gamma_0 = x, \gamma_1, \ldots, \gamma_p = y$ along γ, which divide the length of γ as indicated.

Each point γ_i (for $i = 1, \ldots, p-1$) lies within distance δ of a sample point x_i. We claim that the path $xx_1x_2\ldots x_{p-1}y$ satisfies our requirements. We estimate the length of each edge as follows:

$$\begin{aligned}
d_M(x_i, x_{i+1}) &\leq d_M(x_i,\gamma_i) + d_M(\gamma_i,\gamma_{i+1}) + d_M(\gamma_{i+1},x_{i+1}) \\
&\leq \delta + \ell_1 + \delta \\
&= \epsilon \\
&= \ell_1\epsilon/(\epsilon - 2\delta)
\end{aligned}$$

and similarly:

$$\begin{aligned}
d_M(x, x_1) &\leq \ell_0\epsilon/(\epsilon - 2\delta) \\
d_M(x_{p-1}, y) &\leq \ell_0\epsilon/(\epsilon - 2\delta)
\end{aligned}$$

Since $\ell_0\epsilon/(\epsilon - 2\delta) \leq \epsilon$, we find that each edge has length at most ϵ and so belongs to G. Adding the inequalities, we obtain:

$$\begin{aligned}
d_M(x_0,x_1) + \ldots + d_M(x_{p-1},x_p) &\leq \ell\epsilon/(\epsilon - 2\delta) \\
&< \ell(1+4\delta/\epsilon)
\end{aligned}$$

The last step makes use of the general fact that $1/(1-t) < 1+2t$ when $0 < t < 1/2$. This completes the proof. □

It follows that the distance d_S can approximates d_M arbitrarily well, as the neighborhood range ϵ and the covering scale approach zero.

4 Parameters of the manifold embedding

The assertion that $d_S \approx d_G$ requires us to consider various parameters of the embedded manifold.

The *minimum radius of curvature* $r_0 = r_0(M)$ is defined by:

$$\frac{1}{r_0} = \max_{\gamma,t}\{\|\ddot{\gamma}(t)\|\}$$

where γ varies over all unit-speed geodesics in M and t is in the domain D of γ. The second derivative is computed by regarding γ as a map $D \to \mathbb{R}^n$.

Any Euclidean sphere of radius r_0 has minimum radius of curvature equal to r_0; in particular this is true of circles of radius r_0 contained in some 2-dimensional plane. Intuitively, geodesics in M curl around "less tightly" than circles of radius less than $r_0(M)$.

The *minimum branch separation* $s_0 = s_0(M)$ is defined to be the largest positive number for which $\|x-y\| < s_0$ implies $d_M(x,y) \le \pi r_0$, for $x, y \in M$.

The existence of r_0 and s_0 is guaranteed by the compactness of M. We now make use of these quantities.

Lemma 3. *If γ is a geodesic in M connecting points x and y, and if $\ell = \mathrm{length}(\gamma) \le \pi r_0$ then:*

$$2r_0 \sin(\ell/2r_0) \le \|x-y\| \le \ell$$

The right-hand inequality is just the fact that the line segment between x and y is the shortest arc connecting x to y. The proof of the left-hand inequality is omitted, but see [3] for the proof. Equality on the left is achieved in the case where γ is an arc of a circle of radius r_0. Thus Lemma 3 crystallizes the notion that geodesics in M curl around less tightly than circles of radius r_0.

Using the fact that $\sin(t) \ge t - t^3/6$ for $t \ge 0$, we can write down a weakened form of Lemma 3:

$$(1 - \ell^2/24r_0^2)\ell \le \|x-y\| \le \ell$$

We make this change purely for cosmetic purposes; it makes it easy to see that when ℓ is small, all three terms are approximately equal.

There is also the "first-order" weakening

$$(2/\pi)\ell \le \|x-y\| \le \ell,$$

which is valid in the range $\ell \le \pi r_0$.

We now have the main ingredients required to show $d_G \approx d_S$.

Corollary 4 *For all $\lambda > 0$, if points x, y in M satisfy:*

$$\begin{aligned} \|x-y\| &< s_0 \\ \|x-y\| &\le (2/\pi) r_0 \sqrt{24\lambda} \end{aligned}$$

and x, y is connected by a geodesic of length $d_M(x,y)$, then

$$(1-\lambda) d_M(x,y) \le \|x-y\| \le d_M(x,y)$$

Proof. The first assumption implies that $\ell = d_M(x,y) \le \pi r_0$. Since we assume that this distance is represented by a geodesic arc, we can apply the weak form of Lemma 3. First of all we have

$$\ell \le (\pi/2)\|x_i - x_{i+1}\| \le r_0\sqrt{24\lambda},$$

using the first-order approximation. It follows that $1 - \lambda \leq 1 - \ell^2/24r_0^2$ and so Lemma 3 implies the desired result. $\square$

Corollary 4 relates to the crucial inequality $(1-\lambda)d_S \leq d_G \leq d_S$, since d_S and d_G are defined in terms of lengths of geodesics and lengths of straight line segments, respectively. However, one must be careful to ensure that all relevant graph edges have a corresponding geodesic arc.

This last assumption is not always valid. If M is a convex domain in $\mathbb{R}^n$ with a hole punched out of its interior, then points on opposite sides of the hole are not connected by a shortest geodesic.

It is useful to have a criterion to identify when this assumption is valid. We say that M is *geodesically convex* if any two points x, y in M are connected by a geodesic of length $d_M(x,y)$.

Examples. Convex domains in $\mathbb{R}^n$ are geodesically convex, as are compact Riemannian manifolds without boundary. In general, if M is a compact Riemannian manifold, then M is geodesically convex if and only if its boundary is convex, in an appropriate sense.

This notion can be put to use in theorems of the following kind.

Main Theorem A. *Let M be a compact submanifold of $\mathbb{R}^n$ with minimum radius of curvature, r_0 and branch separation s_0. Let $\{x_i\}$ be a finite set of data points in M. Let G be a graph with vertices $\{x_i\}$, and $0 < \lambda_1, \lambda_2 < 1$. Suppose*

1. *The graph G contains all edges xy s.t. $\|x-y\| \leq 4C(d_M, \{x_i\})/\lambda_2$;*
2. *All edges of G have length $\|x-y\| \leq (2/\pi)r_0\sqrt{24\lambda_1} < s_0$;*
3. *The submanifold M is geodesically convex.*

then for all x, y in $\{x_i\}$:

$$(1-\lambda_1)d_M(x,y) \leq d_G(x,y) \leq (1+\lambda_2)d_M(x,y)$$

The theorem gives us conditions for ensuring that $d_G \approx d_M$ to any given degree of accuracy, by setting λ_1 and λ_2 as small as required. Note that if conditions (1) and (2) can not be simultaneously satisfied, the theorem says nothing. This implies that our covering scale must be smaller than the scales implied by the radius of curvature and branch separation.

Proof.

Let $\epsilon_{min} = 4C(d_M, \{x_i\})/\lambda_2$.

Since $\|x-y\| \leq d_M(x,y)$, condition 1 implies that G contains all edges xy for which $d_M(x,y) \leq \epsilon_{min}$. We can therefore apply Theorem 2, which with Proposition 1 gives

$$d_G(x,y) \leq d_S(x,y) \leq (1+4C(d_M,\{x_i\})/\epsilon_{min})d_M(x,y) \leq (1+\lambda_2)d_M(x,y),$$

establishing the right-hand inequality.

To prove the left-hand inequality, choose a path $x_0x_1\ldots x_p$ connecting x $(= x_0)$ to y $(= x_p)$ along graph edges which minimizes the total graph length. The conditions on G and ϵ_{max} imply that each pair x_i, x_{i+1} satisfies the hypotheses of Corollary 4 with $\lambda = \lambda_1$. Then:

$$\begin{aligned}
d_M(x,y) &\leq d_M(x_0,x_1) + \ldots + d_M(x_{p-1},x_p) \\
&\leq (1-\lambda_1)^{-1}\|x_0 - x_1\| + \ldots + (1-\lambda_1)^{-1}\|x_{p-1} - x_p\| \\
&= (1-\lambda_1)^{-1} d_G(x,y)
\end{aligned}$$

This establishes the left-hand inequality. □

Aside. It may seem that geodesic convexity is a much stronger condition than we really need for our proof. We could say that M is ϵ-geodesically convex if whenever $\|x - y\| \leq \epsilon$ there is a geodesic of length $d_M(x,y)$ connecting x to y. Taking $\epsilon = \epsilon_{max}$, this is all that is required in the proof. Curiously enough, it turns out that ϵ-geodesic convexity implies full geodesic convexity for a metrically complete manifold, such as a compact manifold, so nothing is really gained by doing this.

5 Relationship to the Isomap Algorithm

One piece of theoretical work is necessary in order to show that the Isomap algorithm works. In particular, it is necessary to show that the graph constructed by the Isomap algorithm satisfies the properties required in Main Theorem A. This is relatively simple when the notion of neighborhood is ϵ nearest neighbor and somewhat more complicated when the K nearest neighbors. For reasons of space, this is not included, but see [3] for details.

There are two main claims in [5]: that d_G converges asymptotically to d_M, and therefore that Isomap asymptotically recovers the true structure for data manifolds that are isometric to a convex domain of Euclidean space. We consider each of these pieces in turn.

The claim that d_G converges asymptotically to d_M is an application of Theorem A for an appropriate ϵ nearest neighbor rule.

The assertion that Isomap asymptotically recovers the true structure of intrinsically Euclidean data sets is really a statement about how the d-dimensional Euclidean embedding Y produced by Isomap relates to a hypothetical Euclidean configuration that can be thought of as the first stage of a generative model of the data. We imagine that the data were generated by first sampling points from V, a convex region of d-dimensional Euclidean space, and then mapping those points isometrically (and in general, nonlinearly) into a high-dimensional observation space X. The observed points in X (or more properly, their distances d_X) are then the inputs to Isomap, and the goal of the algorithm is to invert the nonlinear isometric mapping $V \to X$. Let $\mathbf{v}_i$ denote the coordinates of point i in the original Euclidean space V, which we would like to recover, and let $\mathbf{y}_i$ denote the coordinates of i in the embedding produced by Isomap. Our claim is that in the limit of infinitely many data points, the embedding coordinates $\{\mathbf{y}_i\}$ produced by applying MDS to any fixed finite set of points converge asymptotically to their original coordinates $\{\mathbf{v}_i\}$, up to Euclidean isomorphism (translation, rotation, scaling).

This claim rests on several properties of the algorithm for classical multidimensional scaling (MDS), which serves as the final step of Isomap. MDS takes as input the pairwise distances for a set of points and constructs a d-dimensional Euclidean embedding that captures those distances as closely as possible. When MDS is given as input actual Euclidean distances for a configuration of points in a d-dimensional space, it will construct a d-dimensional embedding that captures those distances perfectly and thereby recovers exactly the original configuration of the points, up to isomorphism. In general, any lower dimensional embedding will not capture the distances perfectly, so it is easy to detect that d is the true dimensionality of the data. Moreover, MDS is continuous: small bounded variations in the input distances lead to small bounded variations in the output embedding. These properties are described in detail by Cox & Cox [8].

The second step of Isomap produces distance estimates d_G that converge to the true geodesic distances d_M. If the original coordinates $\{\mathbf{v}_i\}$ were sampled from a convex Euclidean domain V, then the geodesic distances d_M will be equal to the Euclidean distances in V. These two facts guarantee that MDS (and hence Isomap) will asymptotically recover the original Euclidean configuration of the data. We can obtain quantitative bounds on how quickly the embedding coordinates $\{\mathbf{y}_i\}$ converge to the original coordinates $\{\mathbf{v}_i\}$ by combining the bounds in our Main Theorem with the perturbation analysis of MDS on page 38 of [8].

6 Conclusion

We have shown that it is possible to recover a low-dimensional representation of a compact Euclidean manifold embedded in a high dimensional space given that the manifold is embedded "nicely" (with a reasonable branch separation and radius of curvature) and that the points cover the manifold on a small enough scale. However, many question remain open.

If we do not insist on step 3 of the Isomap algorithm (embed the graph in $\mathbb{R}^d$ using MDS), then this approach is much more general: it works for any geodesically convex submanifold M rather than just submanifolds which are isometric to a convex subset of $\mathbb{R}^d$.

This paper is about *sufficient* conditions allowing Isomap (or some other algorithm) to extract a manifold from a high dimensional space. It is interesting to consider *necessary* conditions as well. What are necessary conditions for manifold extraction up to error ϵ?

We have not analyzed observation noise. Suppose the data points are perturbed by some low-amplitude isotropic noise process. We expect that the asymptotic result of applying Isomap to this noisy data will be a "thickened" embedding, with the top d embedding dimensions accounting for most of the variance in graph distances and the bottom $n - d$ dimensions accounting for some small residual variance whose amplitude can be bounded in terms of the magnitude of the noise. Manifold reconstruction in the presence of embedding noise should be possible, but it is not yet analyzed (see [6] for some thoughts).

The theory described here was developed in tandem with the current version of the Isomap algorithm (see [4] for an earlier version), and led to design changes with greatly improved performance: significantly lower empirical error obtained in an order of magnitude less time, using an order of magnitude less training data.

Acknowledgments

We thank S. Dasgupta, R. Lehrer, S. Mahajan, and H. Moore for helpful comments on this work.

References

[1] Nina Amenta, Marshall Bern, David Eppstein, "The crust and the beta-skeleton: combinatorial curve reconstruction", Graphical models and image processing, 60/2:2, pages 125-135 (1998)

[2] Nina Amenta, Sunghee Choi, Ravi Kolluri. "The Power Crust", Proceedings of 6th ACM Symposium on Solid Modeling, 2001, pages 249-260.

[3] Mira Bernstein, Vin de Silva, John Langford and Josh Tenenbaum Graph approximations to geodesics on embedded manifolds Technical report 2000, http://isomap.stanford.edu/BdSLT.ps

[4] J. B. Tenenbaum (1998). Mapping a manifold of perceptual observations. Advances in Neural Information Processing Systems 10.

[5] J. B. Tenenbaum, V. de Silva, J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** (5500), 22 Dec 2000.

[6] J. B. Tenenbaum, V. de Silva, J. C. Langford (2000). The Isomap Algorithm and Topological Stability. *Science* **295** (5552), 4 Jan 2002.

[7] H. Chernoff (1952). A measure of asymptotic efficiency for test of a hypothesis based upon the sum of observations. *Annals of Mathematical Statistics* **23**, 493-507.

[8] T. F. Cox and M. A. A. Cox (1994). *Multidimensional Scaling*. London: Chapman and Hall.

[9] W. Johnson and J. Lindenstrauss "Extensions of Lipschitz maps into a Hilbert Space." Contemp. Math. 26, 1984, pp. 189-206.

[10] Piotyr Indyk, "Algorithmic applications of low-distortion geometric embeddings" FOCS2001 tutorial and survey paper.

[11] Sam Roweis and Lawrence Saul, "Nonlinear dimensionality reduction by locally linear embedding." Science, v.290 no.5500, Dec.22, 2000. pp.2323–2326.

[12] Vladimir N. Vapnik, "Statistical Learning Theory", Wiley, December 1999.