Seminar Report

Applying Semi-Supervised Locally Linear Embedding

Department of Statistics
Ludwig-Maximilians-Universität München

By Lisa Wimmer
Under the supervision of Jann Goschenhofer, Ph.D.

Munich, month day$^{\text{th}}$, 2021

Abstract

Storyline

- Goal: present SS-LLE as a local, graph-based manifold learning method incorporating prior knowledge
- Step 0: define basic mathematical concepts required to understand argumentation (plus notation)
- Step 1: introduce idea of **isometry** (most basic: MDS)
- Step 2: introduce idea of **graph-based** models
    - Achieve non-linearity
    - Common structure: build graph → derive matrix as quadratic form over graph function → derive embedding from eigenvalue problem
    - Most basic: ISOMAP (global, dense, convex)
- Step 3: introduce idea of **locality**
    - Relax global to local isometry
    - Find sparse rather than dense matrices
    - **Laplacian eigenmaps** as concept in which the others can be generalized
        - Define weighting scheme for neighborhood
        - Use Laplacian to derive matrix
        - Solve sparse eigenvalue problem
- Step 4: introduce **local linearity**
    - **LLE**
        - Obtain weights via linear reconstructions
        - Can be shown to approximate graph Laplacian (Belkin & Niyogi (2006))
    - **Hessian LLE**
        - Replace Laplacian by Hessian
- Step 5: introduce **prior knowledge**
    - **SS-LLE**
    - Improve results by pre-specifying some manifold coordinates

Contents

List of Figures

List of Tables

1 Introduction

Machine learning problems increasingly employ data of high dimensionality. While a large amount of samples is beneficial to learning, high-dimensional feature spaces, such as in speech recognition or gene processing, pose serious obstacles to the performance and convergence of most algorithms (Cayton, 2005).
Three aspects strike as particularly problematic: computational operations, interpretation of results, and geometrical idiosyncrasies. Computational cost must be considered but is becoming less of an issue with the evolution of technology (Leist et al., 2009). By contrast, the demand for explainable results (for reasons of, say, safety or ethics) is rather intensified by the advance of complex methods. Alas, interpretation in more than a few dimensions is virtually inaccessible to humans (Doshi-Velez and Kim, 2017). The geometric aspect is often addressed as *curse of dimensionality*, a term subsuming various phenomena of high-dimensional spaces. It is generally not straightforward to infer properties of objects in complex spaces as geometric intuition developed in two or three dimensions can be misleading. Crucially, the exponential increase of spatial volume induces sparsity. Consequences of this behavior are, among others, a sharp incline in the number of points required to sample the feature space and a loss in meaningfulness of distances. Many learners, however, rely on these concepts[1] and see their functionality deteriorate (Verleysen and Francois, 2005).
These challenges make the case for *dimensionality reduction*, that is, the endeavor of compressing problem dimensionality to a manageable size. Far from undue simplification, dimensionality reduction relies on the idea that the latent data-generating process is indeed of much lower dimension than is observed. Consider, for example, textual data that are often represented in high-dimensional document-term matrices counting the frequency of single terms across documents. It is reasonable to suppose the texts are in fact characterized by much lower-dimensional semantic features, or topics (Roweis and Saul, 2000). More formally, the data are assumed to lie on a d-dimensional *manifold* embedded in the D-dimensional observation space, with $d \ll D$. The goal is thus to uncover the structure of this manifold in an unsupervised manner (Cayton, 2005).

Various approaches have been proposed to learn points' manifold coordinates so they can be mapped to the corresponding d-dimensional Euclidean space[2]. A taxonomy can for example be found in van der Maaten et al. (2009). Many methods rely on spectral techniques, trying to find a matrix representation of the data whose principal eigenvectors are used to span a d-dimensional subspace. Among these spectral methods some are confined to learning linear embeddings (such as *principal component analysis (PCA)* or *multi-dimensional scaling (MDS)*). Since linearity is a strong assumption that will not hold for general manifolds, non-linear techniques are more widely applicable. They can be further divided along the scope of the structure they attempt to preserve: full spectral methods (for instance, *ISOMAP*) retain a global notion of distance, whereas sparse approaches focus on local properties. Locality al-

[1] For instance, consider support vector machines and k-nearest neighbors, both of which rely on distances, or tuning, which requires extensive sampling of the hyperparameter space.

[2] The most intuitive example of this is probably the representation of the Earth, which is a two-dimensional manifold enclosed in three-dimensional space, on two-dimensional maps.

lows sparse methods to better capture non-convex structures, where global isometry is not appropriate (van der Maaten et al., 2009).

One such technique is *locally linear embedding (LLE)*, proposed by Roweis and Saul (2000). It is based on the idea that points on the manifold lie within locally linear neighborhoods reflecting intrinsic geometric properties. Consequently, weights of linear reconstruction from neighboring points in the D-dimensional original space should be the same as for the d-dimensional manifold coordinates. LLE thus maps vicinity structures, characterized by neighborhood graphs, to the d-dimensional subspace and finds the coordinates that preserve them best. This requires solving the least-squares problem of minimizing reconstruction error and then the sparse eigenvalue problem of minimizing embedding cost. Convexity of the latter guarantees globality of any local optimum.
The original LLE algorithm uses no prior information. As Yang et al. (2006) argue, however, prior knowledge can improve performance by anchoring the unsupervised task to some known coordinates. The results presented in their work indicate considerable success of *semi-supervised locally linear embedding (SS-LLE)*.

It is the aim of this report to (1) reproduce these results, thereby creating an open-source implementation of SS-LLE, and (2) to apply SS-LLE to further manifold learning tasks for a more thorough assessment of its performance. The rest of the report is organized as follows: chapter 2 provides a mathematical framework where fundamental concepts are briefly introduced; chapter 3 explains the idea of local graph-based manifold learning; chapter 4 presents SS-LLE in detail; chapter 5 discusses the results of the conducted experiments; and chapter 6 draws final conclusions.

2 Mathematical Framework

2.1 Basic Geometric Concepts

This chapter introduces the main geometric concepts considered necessary to provide a solid understanding of SS-LLE[3]. It must be noted that everything discussed here is presented through the lens of machine learning, deliberately forsaking the generality inherent to topology. Therefore, assuming features can be represented by coordinates in D-dimensional Euclidean space, all concepts are examined with regard to their meaning in $\mathbb{R}^D$. Dimensionality reduction techniques take the data observed in $\mathbb{R}^D$ to actually lie in a d-dimensional topological space that is not necessarily Euclidean but exhibits some specific properties.

Topological spaces. A *topological space* is constituted by a set X equipped with a *topology* $\mathcal{T}$. A topology is a general way of describing relations between elements in X. Consider a function $\mathcal{T} : X \to 2^X, x \mapsto \mathcal{T}(x)$, which assigns to $x \in X$ a set of subsets of X called a *neighborhood*. For $\mathcal{T}$ to be a topology[4] on X, the following properties must hold (Brown, 2006):

[3] Obviously, the list of concepts discussed is by no means extensive. Theory is presented much more in detail (and mathematical rigor) in, for example, good book.

[4] Alternative definitions employ open subsets of X, see for example Waldmann (2014).

1. If $\mathcal{T}$ is a neighborhood of x, then $x \in \mathcal{T}$.
2. If $\mathcal{T}$ is a subset of X containing a neighborhood of x, then $\mathcal{T}$ is a neighborhood of x.
3. The intersection of two neighborhoods of x is again a neighborhood to x.
4. Any neighborhood $\mathcal{T}$ of x contains a neighborhood $\mathcal{T}'$ of x such that $\mathcal{T}$ is a neighborhood of each element in $\mathcal{T}'$.

Note that, in this general definition, neighborhoods are based on an abstract notion of "nearness". Learning the structure of a topological space effectively boils down to learning neighborhood relations. In Euclidean topological space these are directly based on distance: neighborhoods are constructed by ϵ-balls containing all elements within a Euclidean distance of ϵ from x. The resulting topology is also called the *metric topology* (McCleary, 2006).
Topological spaces in general are not accessible via distances. The ultimate goal is again the interpretation of the data in a Euclidean space, albeit one with lower dimensionality, where such concepts are meaningful. The next step is thus to study how a (potentially highly non-linear and complicated) topological space might relate to $\mathbb{R}^d$.

Homeomorphisms. Consider two topological spaces $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ (denoted by the respective shorthands X, Y from here) and a mapping function $f : X \to Y$. If f is bijective and continuous and $f^{-1} : Y \to X$ is also continuous, f is called a *homeomorphism*. Intuitively, this is equivalent to $f(\mathcal{T})$ being a neighborhood of $f(x)$ if $\mathcal{T}$ is a neighborhood of x (Brown, 2006). Topological spaces for which such a mapping exists are *homeomorphic* to each other. Any properties of X that Y shares when it is homeomorphic to X are referred to as topological properties. Two homeomorphic spaces are thus topologically equivalent (McCleary, 2006).
If there exists a non-negative integer d such that for every x in a topological space X a local neighborhood is homeomorphic to an open subset of $\mathbb{R}^d$, X is *locally Euclidean*[5]. In local neighborhoods X then behaves like $\mathbb{R}^d$, which is conceivably a desirable property in this context (Ma and Fu, 2011).

Manifolds. *Manifolds* are now precisely such locally Euclidean topological spaces, with some additional properties. For $\mathcal{M}$[6] to be a d-dimensional manifold (also: d-manifold) it must meet the following conditions (Waldmann, 2014):

1. $\mathcal{M}$ is Hausdorff.
2. $\mathcal{M}$ is second-countable.
3. $\mathcal{M}$ is locally homeomorphic to $\mathbb{R}^d$.

The Hausdorff condition is a separation property and ensures that for any two distinct points from $\mathcal{M}$ disjoint neighborhoods can be found (Brown, 2006). Second-countability restricts the manifold's size via the number of open sets it may possess (Waldmann, 2014).

[5] For locally Euclidean topological spaces it is thus meaningful to speak of elements as points.

[6] This is again a shorthand, omitting the explicit notation of the corresponding topology to enhance readability.

Manifolds can now be *embedded* in Euclidean space. Consider $\mathbb{R}^k \supset \mathbb{R}^d$[7]. $\mathbb{R}^d$ is endowed with the so-called *subspace topology* that results from intersecting open subsets of $\mathbb{R}^k$ with $\mathbb{R}^d$ (for $\mathbb{R}^2$, these are ϵ-circles obtained by intersecting $\mathbb{R}^3$-ϵ-balls with the coordinate planes). For a manifold $\mathcal{M}$ to be embedded in $\mathbb{R}^k$ means that $\mathcal{M}$ is enclosed by $\mathbb{R}^k$ but locally homeomorphic to $\mathbb{R}^d$, thereby inheriting the metric subspace topology from $\mathbb{R}^d$ (Waldmann, 2014). It can be shown that $k = 2d + 1$ is sufficient to create an embedding, but k may be smaller (Ma and Fu, 2011).
This now has important consequences for manifold learning: data lying on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^k$ are observed as k-dimensional points but may locally be treated like points from $\mathbb{R}^d$.

Figure 1 shows the well-known *S-curve* manifold embedded in $\mathbb{R}^3$. Clearly, the S-curve as a whole is far from linear, but local patches on its surface behave like flat surfaces from $\mathbb{R}^2$. So the S-curve is two-dimensional and feature dimensionality can in effect be compressed from $\mathbb{R}^3$ to $\mathbb{R}^2$. The challenge is now to unravel the manifold in a way that preserves its structure to maximum extent. Obviously, a simple projection to $\mathbb{R}^2$ (onto any coordinate plane) will not accomplish this task. Instead, manifold learning must capture the intrinsic neighborhood structures and map these to $\mathbb{R}^2$, which, in this case, can be imagined as a "flattening-out" of the S-curve.

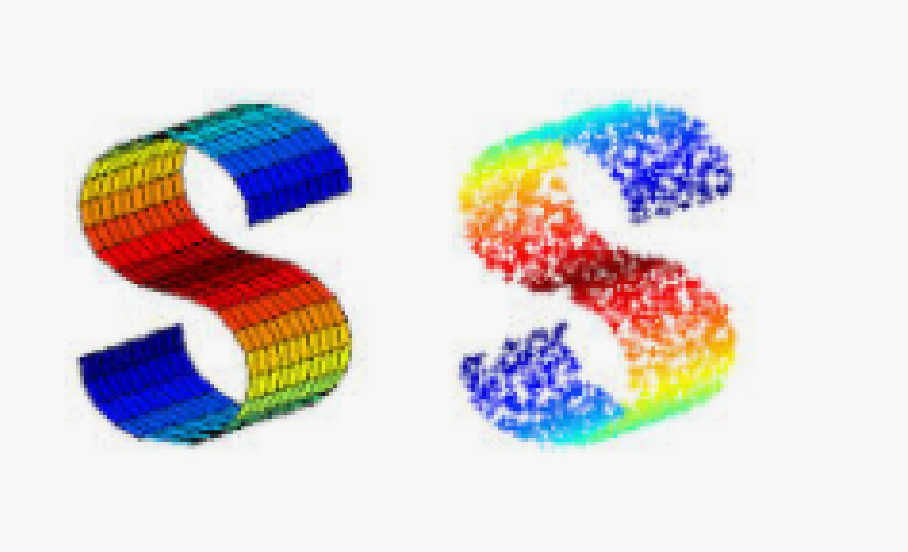

Figure 1: XY points sampled from the S-curve manifold, obtained by bla.

Geodesic. One last aspect remains open, namely how to handle distances on manifolds. Figure 1 illustrates that standard Euclidean distances are not meaningful here: rather than measuring "shortcuts" between points across $\mathbb{R}^3$ (where, for instance, points in the upper blue area would be considered quite close to points in the red area), it makes intuitive sense to constrain distances to the manifold surface. In order to enable the construction of such a metric, manifolds must fulfill two additional properties: being *Riemannian* and being *connected* (Ma and Fu, 2011).

[7] Here, k is used to denote a general higher-dimensional space (a manifold embedded in $\mathbb{R}^k$ is also embedded in $\mathbb{R}^{k+1}$, and so on, as homeomorphisms are transitive (Waldmann, 2014)). This is deliberate to distinguish it from the more specific notation D indicating the number of observed features.

Riemannian manifolds are differentiable

Connectedness means that no separation $\{U, V\}$ of a manifold $\mathcal{M}$ exists with open, non-empty and disjoint $U, V \subset \mathcal{M}$, $\mathcal{M} = U \cup V$. For manifolds, connectedness is immediately equivalent with path-connectedness, which is perhaps more intuitive: informally stated, any two points on a connected manifold can be linked by a path (McCleary, 2006).

For connected Riemannian manifolds it is now possible to define a distance metric, or *geodesic distance*. Geodesic distance is the length of the shortest curve (*geodesic*) between two points $\mathbf{p}, \mathbf{q} \in \mathcal{M}$ (such a curve must exist due to connectedness).

A curve c in $\mathcal{M}$ is a smooth mapping from an open interval $\Lambda \subset \mathbb{R}$ into $\mathcal{M}$. c is parametrized by a point $\lambda \in \Lambda$, such that $c(\lambda) = (c_1(\lambda), ..., c_d(\lambda))^T$ (all $c_j, j = 1, ..., d$, having a sufficient number of continuous derivatives) is a curve in $\mathbb{R}^d$. Component-wise differentiation with respect to λ yields the *velocity* of c in λ, $c'(\lambda) = (c'_1(\lambda), ..., c'_d(\lambda))^T$. The *speed* of c is given by $\|c'(\lambda)\|_2^2$, where $\| \cdot \|_2^2$ denotes the square norm. Then, distance along this curve is measured by the arc-length $L(c) = \int_{\mathbf{p}}^{\mathbf{q}} \|c'(\lambda)\|_2^2 d\lambda$.

Finally, geodesic distance can be derived as the length of the shortest such curve, out of the set of differentiable curves in $\mathcal{M}$ that connect $\mathbf{p}$ and $\mathbf{q}$, $\mathcal{C}(\mathbf{p}, \mathbf{q})$: $d^{\mathcal{M}}(\mathbf{p}, \mathbf{q}) = \inf_{c \in \mathcal{C}(\mathbf{p}, \mathbf{q})} L(c)$ (Ma and Fu, 2011).

Intuitively, geodesic distance can be identified with Euclidean distance in Euclidean spaces where shortest curves are but straight lines.

2.2 Spectral Decomposition

- Eigenvalues/eigenvectors
- Spectral decomposition

3 Local Graph-Based Manifold Learning

3.1 Concept of Isometry

- Notion of distance
- Preserving distances in manifold learning
- MDS (very brief)

3.2 Graph-Based Models

3.2.1 Neighborhoods

- k-/ϵ-neighborhoods and neighborhood graphs
- Linear reconstruction and reconstruction error

3.2.2 Basics of Spectral Graph Theory

- Degree and adjacency matrices
- Laplacian operators

3.2.3 General Structure of Graph-Based Models

- ○ Neighborhood graph
- ○ Weight matrix
- ○ Eigenwert problem

3.2.4 ISOMAP

- ○ (One of the) earliest, simplest variant(s)
- ○ MDS with geodesics

3.3 Laplacian Eigenmaps

- ○ Notion of locality
- ○ Laplacian eigenmaps

3.4 Locally Linear Embedding (LLE)

- ○ Notion of local linearity
- ○ Approximation of graph Laplacian

3.5 Hessian Locally Linear Embedding (HLLE)

- ○ Hessian instead of Laplacian (eigenmaps)
- ○ Hessian instead of LS fit (LLE)

4 Semi-Supervised Locally Linear Embedding (SS-LLE)

4.1 Employment of Prior Information

- ○ Why use labels in the first place?
- ○ How will that help?
- ○ How do we even find prior points?
- ○ Exact vs inexact knowledge

4.2 SS-LLE Algorithm

- ○ What is different wrt standard LLE?

4.3 Strengths and Drawbacks of SS-LLE

Potential shortcoming: what if manifold is not well-sampled? Not a problem with synthetic data, but IRL. But probably problematic with all manifold approaches Also: generalization to new points (w/o recomputing everything) neighborhood-preserving propositions

5 Experiment Results

5.1 Data

5.2 Experimental Design

- Implementation details
- Hyperparameters
- Evaluation criteria

5.3 Results and Discussion

6 Conclusion

Lorem ipsum

A Appendix

Lorem ipsum

B Electronic Appendix

Data, code and figures are provided in electronic form.

References

Brown, R. (2006). *Topology and Groupoids. A Geometric Account of General Topology, Homotopy Types and the Fundamental Groupoid*, 2 edn, Createspace.

Cayton, L. (2005). Algorithms for manifold learning, *Technical Report CS2008-0923*, University of California, San Diego (UCSD).

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning, *arXiv: Machine Learning* .

Leist, A., Playne, D. P. and Hawick, K. A. (2009). Exploiting graphical processing units for data-parallel scientific applications, *Concurrency and Computation. Practice and Experience* **21**(18): 2400–2437.

Ma, Y. and Fu, Y. (2011). *Manifold Learning. Theory and Applications*, CRC Press.

McCleary, J. (2006). *A First Course in Topology. Continuity and Dimension*, American Mathematical Society.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**(5500): 2323–2326.

van der Maaten, L., Postma, E. and van den Herik, J. (2009). Dimensionality reduction: A comparative review, *Technical Report TiCC TR 2009-005*, Tilburg University.

Verleysen, M. and Francois, D. (2005). The curse of dimensionality in data mining and time series prediction, *in* J. Cabestany, A. Prieto and F. Sandoval (eds), *Computational Intelligence and Bioinspired Systems*, Springer.

Waldmann, S. (2014). *Topology. An Introduction*, Springer.

Yang, X., Fu, H., Zha, H. and Barlow, J. (2006). Semi-supervised nonlinear dimensionality reduction, *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA.

Declaration of Authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.