Lecture 1. Linear Dimensionality Reduction

Juno Kim

Department of Mathematics & Statistics
Seoul National University

Manifold Learning, Spring 2022

Table of Contents

1 Introduction

2 Principal Component Analysis

3 Multidimensional Scaling

What is Manifold Learning?

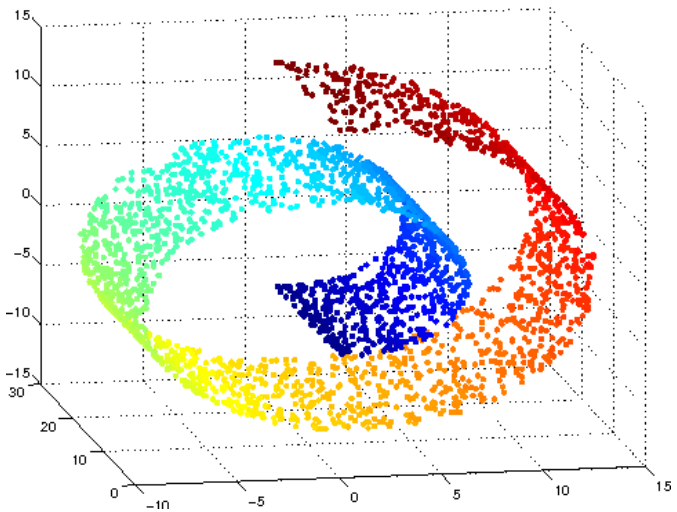

Figure: The Swiss roll dataset.

- A class of algorithms for recovering a low-dimensional manifold embedded in a high-dimensional ambient space
- Usually considered as a topic in unsupervised learning

What is Manifold Learning?

- Finitely many data points $\{y_i\}$, $i = 1 \cdots n$ are drawn from a t-dimensional Riemannian manifold $(\mathcal{M}^t, d^{\mathcal{M}})$, possibly with boundary
- This data is embedded into some high-dimensional Euclidean input space $\mathbb{R}^r$, $r \gg t$ via a smooth embedding $\psi : \mathcal{M} \to \mathbb{R}^r$
- Given the data $x_i = \psi(y_i)$ or the proximity matrix of distances between them, our goal is to recover $\mathcal{M}$, ψ, and y_i up to isometry
- The algorithm returns t'-dimensional estimates $\{\hat{y}_i\}$; we are successful if $t = t'$. We also wish to visualize the data, identify geometric features, make predictions, etc

Prerequisites

2nd year – Multivariate Calculus, Linear Algebra, Analysis

3rd year (Math) – Topology, Riemannian Geometry

3rd year (Stat) – Mathematical Statistics, Linear Regression

We will be drawing upon various statistical methods and ideas from numerical optimization and discrete analysis. Involved concepts will be explained during the lecture.

I also aim to implement differential topological considerations in original research. Any collaboration is welcome.

"Motivation is the only true prerequisite" – Me

Table of Contents

1 Introduction

2 Principal Component Analysis

3 Multidimensional Scaling

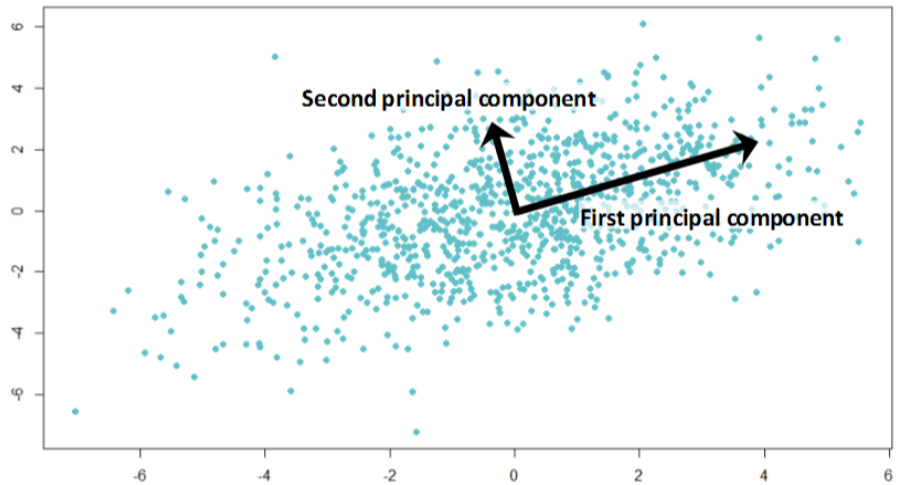

Figure: Principal components of data cloud.

PCA is a technique for deriving a reduced set of orthogonal projections of correlated variables, which maximize the original information. In short, it performs linear dimensionality reduction and is one of the most employed multivariate statistical methods.

Setup

- The input variables are the components of a random r-vector $X = (x_1, \cdots, x_r)^T$ with mean $\mu_X = \mathbb{E}X$ and covariance matrix $\Sigma_X = \mathbb{E}\left[(X - \mu_X)(X - \mu_X)^T\right]$.
- We replace the variables x_i, $i \leq r$ by the principal components $z_j := v_j^T (X - \mu_X)$, $j \leq t$, for some $t \leq r$, where the unit r-vectors v_j are to be decided.
- The z_j are to be orthogonal and ordered by decreasing variance (score).
- Let $Z = (z_1, \cdots, z_t)$ and $\mathbf{V} = (v_1, \cdots, v_t)$, so $Z = \mathbf{V}^T (X - \mu_X)$.

Derivation

Method 1. Since we want Z to retain as much information of the original data as possible, least-squares regress X on Z:

$$\underset{\nu,\mathbf{A},\mathbf{V}}{\operatorname{argmin}} \mathbb{E} \left\| X - \nu - \mathbf{A}\mathbf{V}^T (X - \mu_X) \right\|^2$$

where $\mathbf{A}\mathbf{V}^T$ functions as a single $(r \times r)$-matrix of reduced-rank t.

Method 2. Sequentially find the direction w maximizing variance:

$$\operatorname{var}\left(w^T X\right) = w^T \Sigma_X w$$

update X by $X - \Pi_w X$, $\Pi_w = ww^T$, and repeat.

Exercise. Derive PCA from both approaches.

- We have the spectral decomposition $\Sigma_X = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$
- $\mathbf{\Lambda} = \text{diag}\,(\lambda_1, \cdots, \lambda_r)$ consists of the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$
- The corresponding eigenvectors u_j form the columns of $\mathbf{U}$.
- Both methods yield $v_j = u_j$, $j \leq t$. Thus, PCA returns the eigenvectors of Σ_X with the t largest eigenvalues. The score of each component is $\|z_j\| = \sqrt{\lambda_j}$.
- High scores indicate high spread, while low scores detect multicollinearity. Hence, t can be chosen after PCA is performed, by discarding z_j with low information.
- $\hat{\mathcal{M}}$ is the subspace spanned by $u_1, \cdots, u_t$ translated by μ_X, and the estimated data $\hat{X}_i$ are the projections of X_i to $\hat{\mathcal{M}}$.

Remark

In practice, we do not know μ_X and Σ_X. Thus, we use the usual unbiased estimators: the sample mean $\hat{\mu}_X = \bar{X}$ and the sample covariance matrix $\hat{\Sigma}_X = \frac{1}{n}\sum_i \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T$. However, the eigenstructure of $\hat{\Sigma}_X$ tends to be distorted outwards when r/n is not too small.

Such problems are studied in *random matrix theory*. Topics include calculating distributions of extremal eigenvalues or proposing more stable, optimal estimators.

Table of Contents

1 Introduction

2 Principal Component Analysis

3 Multidimensional Scaling

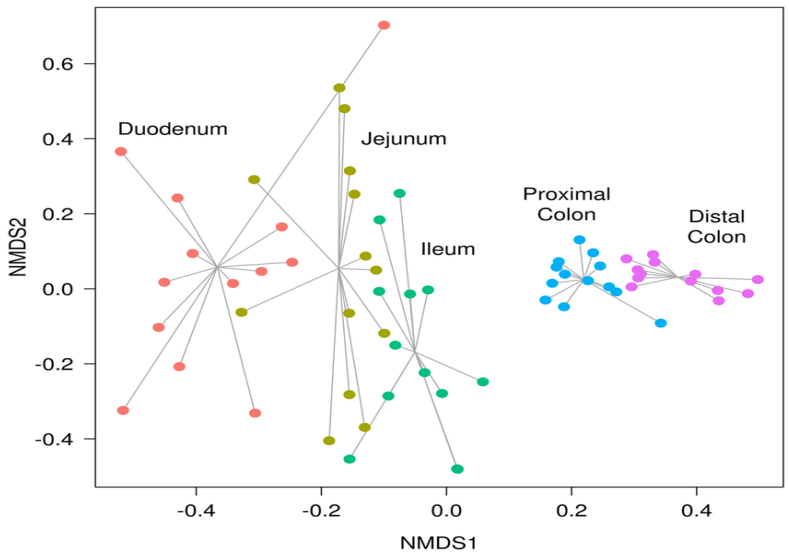

Figure: Identifying clusters via a low-dimensional visualization.

MDS translates information about pairwise 'distances' between a set of objects into a geometric configuration of points in lower dimensional Euclidean space, while preserving the distances as best as possible.

Setup

- For n objects, we are given the proximity δ_{ij} between each pair of objects, stored in an $(n \times n)$-matrix $\Delta = (\delta_{ij})$.
- Proximity may not be an actual distance, but any measure of dissimilarity, possibly subjective and qualitative.
- For sake of exposition, we assume that δ satisfies the conditions to be a metric, and in particular is computed from $X_1, \cdots, X_n \in \mathbb{R}^r$ via $\delta_{ij} = \|X_i - X_j\|$. (The same method will work for general δ.)
- We wish to find $Y_1, \cdots, Y_n \in \mathbb{R}^t$, $t < r$, with pairwise distances 'close' to δ_{ij}.

Isometric Embedding

Discrete metric spaces with ≥ 4 points generally do not isometrically embed into *any* Euclidean space. In particular, a finite metric space $P = \{p_0, \cdots, p_n\}$ has an isometric embedding into $\mathbb{R}^k$ iff the $(n \times n)$-matrix $\mathbf{B}$ given by

$$\mathbf{B}_{ij} = \frac{1}{2}\left(d\left(p_0, p_i\right)^2 + d\left(p_0, p_j\right)^2 - d\left(p_i, p_j\right)^2\right)$$

is positive semi-definite and of rank $\leq k$.

Exercise. Prove the theorem.

Derivation

- Assume that the data is centered: $\sum X_i = 0$.
- Define $a_{ij} := -\frac{1}{2}\delta_{ij}^2$ and $b_{ij} := a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot} = X_i^T X_j$ (show this).
- This 'double centering' can be expressed by the matrices $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b_{ij})$ and $\mathbf{H} = \mathbf{I}_n - n^{-1}11^T$ as: $\mathbf{B} = \mathbf{HAH}$.
- We wish to find a rank $\leq t$ matrix $\mathbf{B}^* = (b_{ij}^*)$ minimizing:

$$\mathrm{tr}\{(\mathbf{B} - \mathbf{B}^*)^2\} = \sum_{i,j}(b_{ij} - b_{ij}^*)^2$$

- Here, $\mathbf{B}$ is the Gram matrix $\mathbf{X}^T\mathbf{X}$ where $\mathbf{X} = (X_1, \cdots, X_n)$, and $\mathbf{B}^* = \mathbf{Y}^T\mathbf{Y}$. We are performing a least-squares fit w.r.t. the inner products, which we can calculate from Δ.

- Let $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be its spectral decomposition. Taking only the largest t eigenvalues and their associated eigenvectors solves MDS:

$$\mathbf{B}^* = \mathbf{U}^*\mathbf{\Lambda}^*\mathbf{U}^{*T} \quad \text{i.e.} \quad \mathbf{Y} = \mathbf{\Lambda}^{*1/2}\mathbf{U}^{*T}$$

  where $\mathbf{\Lambda}^* = \text{diag}\,(\lambda_1, \cdots, \lambda_t)$ and $\mathbf{U}^* = (u_1, \cdots, u_t)$ are the truncated versions. The columns $Y_1, \cdots, Y_n$ are our desired estimates.
- The loss equals $\sum_{i=1}^{t} \min(\lambda_i, 0)^2 + \sum_{i=t+1}^{n} \lambda_i^2$ for general δ
- Simultaneously applying any rigid transformation to $Y_1, \cdots, Y_n$ (currently centered) also yields a solution.

Exercise. Derive the general-case SSE loss.

Choosing t

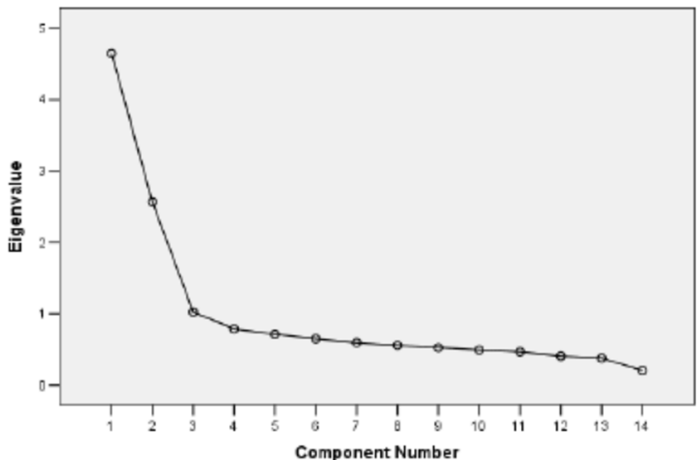

Figure: A scree plot with elbow at $t = 2$.

Setting $t := \operatorname{rank} \mathbf{B}$ gives an exact fit; see the Remark. To assess or reduce dimensionality of noisy data, in particular when Δ does not come from a metric, plot the eigenvalues and choose t such that the eigenvalues after λ_t stabilize (scree test).

Remark

- Classical MDS with Δ derived from Euclidean data as above is equivalent to empirical PCA. To see this, note that for centered data $\mathbf{X} = (X_1, \cdots, X_n)$, $\hat{\Sigma}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ and:

$$\frac{1}{n}\mathbf{X}\mathbf{X}^T v = \lambda v \quad \Leftrightarrow \quad \mathbf{X}^T\mathbf{X} w = n\lambda w, \quad w = \mathbf{X}^T v$$

  (excluding $\lambda = 0$, which we discard in MDS)
- While our main goal is to learn *nonlinear* manifolds, many nonlinear algorithms rely on applying MDS or related spectral methods to locally constructed proximity data to find representations of $\mathcal{M}$ as a linear subset of Euclidean space.