Notes on Statistical and Machine Learning

Multidimensional Scaling

Chapter: *28* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 14, Unsupervised Learning* in Hastie, Tibshirani, and Friedman (2009), and
- *Chapter 13, Multidimensional Scaling and Distance Geometry* in Izenman (2009).

I. Introduction

1. **Overview:** Given only a two-way table of proximities of data points, the problem of multidimensional scaling (MDS) attempts to find a lower-dimensional representation of data that preserves the pairwise distances as well as possible.

2. **Setup:** We are given

   (a) the distances $d_{i,j}$ between the i-th and the j-th observations, or

   (b) the similarity measurements $s_{i,j}$ between the i-th and the j-th observations,

   for all $i, j = 1, 2, \cdots, n$. In particular, we do *not* have the values of the original observations.

3. **Categories of Multidimensional Scaling:** There are two broad categories of approaches to multidimensional scaling:

   (a) *Metric Scaling:* Utilizes the actual similarity or dissimilarity measurements are used;

   Examples. Least squares scaling, Sammon scaling, and classical scaling.

   (b) *Non-metric Scaling:* Only utilizes the ranks of dissimilarity measurements.

   Example. Shephard-Kruskal non-metric scaling.

II. Metric Scaling

1. **Least Squares Scaling:**

   (a) *Main Idea:* The main idea here is that to find a lower-dimensional representation of the data that preserves the pairwise distances as well as possible.

(b) *Formulation:* The *least squares scaling* seeks values $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n \in \mathbb{R}^k$ to minimize the following objective function

$$S_{\text{ls}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n) := \sum_{i \neq j} (d_{i,j} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2. \tag{1}$$

The function S_{ls} is known as the *stress function*.

Remark. This approach to multidimensional scaling is also called Kruskal-Shephard scaling.

2. **Sammon Mapping:** A variation of the least squares scaling is the *Sammon mapping* which minimizes

$$S_{\text{Sammon}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n) := \sum_{i \neq j} \frac{(d_{i,j} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2}{d_{i,j}}, \tag{2}$$

where more emphasis is put on preserving smaller pairwise distances.

3. **Classical Scaling:**

(a) *Formulation:* Suppose we are given the *similarity measurements* as the inner product of centered data, i.e.,

$$s_{i,j} := \langle \mathbf{x}_i - \bar{\mathbf{x}}, \mathbf{x}_j - \bar{\mathbf{x}} \rangle, \qquad \text{for all } i, j = 1, 2, \cdots, n,$$

where $\bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. The *classical scaling* problem attempts to minimize

$$S_{\text{cs}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n) := \sum_{i,j=1}^n (s_{i,j} - \langle \mathbf{z}_i - \bar{\mathbf{z}}, \mathbf{z}_j - \bar{\mathbf{z}} \rangle)^2, \tag{3}$$

where $\mathbf{z}_i \in \mathbb{R}^k$ for all $i = 1, 2, \cdots, n$.

(b) *Alternative Formulation:* Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ with the (i, j)-th entry being $\langle \mathbf{z}_i, \mathbf{z}_j \rangle$, where we assume each $\mathbf{z}_i$ has already been centered so that $\sum_{i=1}^n \mathbf{z}_i = \mathbf{0}_k$. We can then write $\mathbf{M}$ as

$$\mathbf{M} = \begin{pmatrix} \mathbf{z}_1^\top \\ \mathbf{z}_2^\top \\ \vdots \\ \mathbf{z}_n^\top \end{pmatrix} \begin{pmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_n \end{pmatrix}.$$

We can write the criterion (3) as

$$S_{\text{cs}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n) = \text{trace}((\mathbf{S} - \mathbf{M})^\top (\mathbf{S} - \mathbf{M})) = \|\mathbf{S} - \mathbf{M}\|_F^2.$$

Since $\mathbf{z}_i \in \mathbb{R}^k$ for all $i = 1, 2, \cdots, n$, the classical scaling problem reduces to the best rank-k approximation problem for $\mathbf{S}$.

(c) *Derivation of Solution:* Using Eckart-Young theorem, the solution is given by the eigen-decomposition of $\mathbf{S}$. Let $\mathbf{S} = \mathbf{E}\mathbf{D}^2\mathbf{E}^\top$, where $\mathbf{D}^2 := \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with eigenvalues of $\mathbf{S}$ on the diagonal, columns of $\mathbf{E}$ are the eigenvectors of $\mathbf{S}$. Let $\mathbf{e}_i$ be the eigenvector associated with the i-th largest eigenvalue of $\mathbf{S}$. The minimizer to S_{cs} is

$$\begin{aligned} \widehat{\mathbf{M}} &:= \arg\min S_{\mathrm{cs}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n) \\ &= \sum_{\ell=1}^{k} \lambda_\ell \mathbf{e}_\ell \mathbf{e}_\ell^\top \\ &= (\mathbf{E}_k\mathbf{D}_k)(\mathbf{E}_k\mathbf{D}_k)^\top, \end{aligned}$$

where $\mathbf{D}_k := \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_k})$ and $\mathbf{E}_k := (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k) \in \mathbb{R}^{n\times k}$.
In particular, if we let $(\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \cdots, \hat{\mathbf{z}}_n) := \arg\min S_{\mathrm{cs}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)$, then $\hat{\mathbf{z}}_i^\top$ is given by the i-th row of $\mathbf{E}_k\mathbf{D}_k$.

4. **Connection with PCA:** If the similarities are in fact centered inner-products, classical scaling is exactly equivalent to principal components, an inherently linear dimension-reduction technique.

III. Non-Metric Scaling

1. **Non-metric Scaling:** *Shephard-Kruskal non-metric scaling* uses only ranks and seeks to minimize the stress function

$$S_{\mathrm{NM}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n, \theta) := \frac{\sum_{i\neq j}(\|\mathbf{z}_i - \mathbf{z}_j\|_2 - \theta(d_{i,j}))^2}{\sum_{i\neq j}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}$$

over $\mathbf{z}_i$'s and an arbitrary increasing function θ.

Minimizing $S_{\mathrm{NM}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n, \theta)$ involves the following two steps:

(1) With the function θ fixed, we minimize over $\mathbf{z}_i$ by gradient descent;

(2) With $\mathbf{z}_i$'s fixed, we use the method of isotonic regression to find the best monotonic approximation $\theta(d_{i,j})$ to $\|\mathbf{z}_i - \mathbf{z}_j\|_2$.

These two steps are iterated until the solutions stabilize.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.