A Comprehensive Mathematical Note: Kruskal's Steepest Descent

Koji Kosugi

1 preface

Assuming that m objects are placed in a P-dimensional space, let x_{it} be the configure of object i in the tth dimension, and define the distance separation d_{ij} as follows.

$$d_{ij} = \sqrt{\sum_{t=1}^{P}(x_{it} - x_{jt})^2}$$

Although it is possible to use Minkowski's general distance as the distance, in this section, for the sake of simplicity, we restrict ourselves to the Euclidean distance.

For the dissimilarity s_{ij} of the data calculated from this coordinate distribution, we would like to have a monotonic relationship such that $d_{ij} > d_{jk}$ when $s_{ij} > s_{ik}$. However, it is difficult to match directly, so we consider the disparity $\hat{d}_{ij}$ as an intermediate variable and take the procedure of finding X_{ij} so that $\hat{d}_{ij}$ has a completely monotonic relationship with S_{ij} while minimizing the error between $\hat{d}_{ij}$ and d_{ij} so that the error between $\hat{d}_{ij}$ and d_{ij} is minimized.

This error is specifically called Stress, and Stressη can be calculated in the following two ways:

$$\eta_1 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(d_{ij} - \hat{d}_{ij})^2/d_{ij}}$$

$$\eta_2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(d_{ij} - \hat{d}_{ij})^2/\sum_{i=1}^{m}\sum_{j=1}^{m}(d_{ij} - d)^2}$$

where

$$d_{ij} = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{\substack{j=1 \\ (j \neq i)}}^{m} d_{ij}.$$

In the following, η_2 will be discussed simply as η.

The problem now boils down to finding the gradient to update to the optimal value by sequential computation, given an initial value of some coordinate x_{it}. This note is a follow-up note on the calculation of the partial derivative to compute the gradient, following up on the intermediate calculations of Kruskal[1].

2 Method of steepest descent

Successively update the value of $\boldsymbol{x}$ as $\boldsymbol{x}^{q+1} = \boldsymbol{x}^q + \alpha^q \boldsymbol{d}^q$. where $\boldsymbol{d}$ is the gradient vector and α is the step size or learning rate. To obtain this gradient vector, differentiate stress η by $\boldsymbol{X}$.

In computing $\dfrac{\partial \eta}{\partial \boldsymbol{X}}$, let A, B be as follows.

$$A = \sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2$$

$$B = \sum_i \sum_j (d_{ij} - d)^2$$

This allows η to be expressed as follows.

$$\eta = (A/B)^{1/2}$$

Differentiating η by x_{it}. In the following equation expansion, the following formulas for the derivative of the composite function and the derivative of the quotient are used, and should be checked.

■Differentiation of composite functions :

$$\{f(g(x))\}' = f'(g(x)) \cdot g'(x)$$

■Differentiation of quotient :

$$\left\{ \frac{f(x)}{g(x)} \right\}' = \left\{ \frac{f'(x)g(x) - g'(x)f(x)}{\{g(x)\}^2} \right\}$$

Let's begin.

$$\begin{aligned}
\frac{\partial \eta}{\partial x_{it}} &= \frac{\partial (A/B)^{1/2}}{\partial x_{it}} \\
&= \frac{1}{2}(A/B)^{-1/2} \cdot \frac{\partial (A/B)}{\partial x_{it}} \\
&= \frac{1}{2}\frac{1}{\eta}\frac{\partial (A/B)}{\partial x_{it}} \\
&= \frac{1}{2}\frac{1}{\eta}\left\{ \frac{\frac{\partial A}{\partial x_{it}} B - \frac{\partial B}{\partial x_{it}} A}{B^2} \right\} \\
&= \frac{1}{2}\frac{1}{\eta}\left\{ \frac{\frac{\partial A}{\partial x_{it}} B}{B^2} - \frac{\frac{\partial B}{\partial x_{it}} A}{B^2} \right\} \\
&= \frac{1}{2}\frac{1}{\eta}\left\{ \frac{\partial A}{\partial x_{it}}\frac{1}{B} - \frac{\partial B}{\partial x_{it}}\frac{A}{B^2} \right\}
\end{aligned}$$

We now check that η can be transformed as follows.

$$\eta = \sqrt{\frac{A}{B}},\ \eta^2 = \frac{A}{B},\ \frac{1}{B} = \frac{\eta^2}{A},\ \frac{\eta^2}{B} = \frac{A}{B^2}$$

The following development follows from this:

$$= \frac{1}{2}\frac{1}{\eta}\left\{\frac{\partial A}{\partial x_{it}}\frac{1}{B} - \frac{\partial B}{\partial x_{it}}\frac{A}{B^2}\right\}$$

$$= \frac{1}{2}\frac{1}{\eta}\left\{\frac{\partial A}{\partial x_{it}}\frac{\eta^2}{A} - \frac{\partial B}{\partial x_{it}}\frac{\eta^2}{B}\right\}$$

$$= \frac{1}{2}\left(\frac{\eta}{A}\frac{\partial A}{\partial x_{it}} - \frac{\eta}{B}\frac{\partial B}{\partial x_{it}}\right)$$

We now turn our attention to $\frac{\partial A}{\partial x_{it}}$ and $\frac{\partial B}{\partial x_{it}}$.The $\frac{\partial A}{\partial x_{it}}$ is as follows.

$$\frac{\partial A}{\partial x_{it}} = \frac{\partial}{\partial x_{it}}\sum_i\sum_j(d_{ij} - \hat{d}_{ij})^2$$

Since A is a function of d and d is a function of x, we transform as follows:

$$\frac{\partial A}{\partial x_{it}} = \frac{\partial A}{\partial d_{ij}}\frac{\partial d_{ij}}{\partial x_{it}}.$$

For $\frac{\partial A}{\partial d_{ij}} = \frac{\partial}{\partial d_{ij}}\sum\sum(d_{ij} - \hat{d}_{ij})^2$, this is also the derivative of the composite function.

We now consider the derivative of the composite function, $f(g(x))'$, as $f(x) = x^2, g(d_{ij}) = (d_{ij} - \hat{d}_{ij})^2$. Note that the disparity $\hat{d}_{ij}$ in $g(d_{ij})$ is a distance d_{ij} is a quantity that does not depend on the distance d_{ij}. So the derivative here is 1, and the calculation is as follows:

$$\frac{\partial A}{\partial d_{ij}} = \sum\sum 2(d_{ij} - \hat{d}_{ij}) \cdot 1 \cdot \frac{\partial}{\partial d_{ij}}(d_{ij} - \hat{d}_{ij}).$$

If we treat ∂d_{ij} as if it were a symbol for a small quantity and expand the equation, we can organize it as follows.

$$\begin{aligned}\frac{\partial A}{\partial x_{it}} &= \frac{\partial A}{\partial d_{ij}}\frac{\partial d_{ij}}{\partial x_{it}} \\ &= \sum\sum 2(d_{ij} - \hat{d}_{ij}) \cdot \frac{\partial}{\partial d_{ij}}(d_{ij} - \hat{d}_{ij}) \cdot \frac{\partial d_{ij}}{\partial x_{it}} \\ &= \sum\sum 2(d_{ij} - \hat{d}_{ij}) \cdot \frac{\partial(d_{ij} - \hat{d}_{ij})}{\partial x_{it}}\end{aligned}$$

Similarly, $\frac{\partial B}{\partial x_{it}}$ can be expanded as follows.

$$\begin{aligned}\frac{\partial B}{\partial x_{it}} &= \frac{\partial b}{\partial d_{ij}}\frac{\partial d_{ij}}{\partial x_{it}} \\ &= \sum\sum 2(d_{ij} - d) \cdot \frac{\partial}{\partial d_{ij}}(d_{ij} - d) \cdot \frac{\partial d_{ij}}{\partial x_{it}} \\ &= \sum\sum 2(d_{ij} - d) \cdot \frac{\partial(d_{ij} - d)}{\partial x_{it}}\end{aligned}$$

where $\frac{\partial \hat{d}_{ij}}{\partial x_{it}}$ is the change in disparity for small changes in x_{it}. However, since disparity is a quantity independent of x_{it}, $\frac{\partial \hat{d}_{ij}}{\partial x_{it}} = 0$. Additionally, since $\frac{\partial d_{ij}}{\partial x_{it}}$ is represented as $d_{ij} = d_{ji}$, the changes cancel each other out when considering all i, j combinations, leading to $\frac{\partial d_{ij}}{\partial x_{it}} = 0$. *1

So far, we have been able to transform the following.

$$\frac{\partial A}{\partial x_{it}} = \sum \sum 2(d_{ij} - \hat{d}_{ij}) \frac{\partial d_{ij}}{\partial x_{it}}$$

$$\frac{\partial B}{\partial x_{it}} = \sum \sum 2(d_{ij} - d) \frac{\partial d_{ij}}{\partial x_{it}}$$

The remainder is the derivative of the distance, $\frac{\partial d_{ij}}{x_{ij}}$. This can be expanded as follows.

$$\frac{\partial d_{ij}}{x_{ij}} = \frac{1}{\partial x_{it}} \sqrt{\sum_{t=1}^{P} (x_{it} - x_{jt})^2}$$

Differentiation of composite functions

$$= \frac{1}{2} (\sum (x_{it} - x_{jt})^2)^{-1/2} \frac{\partial}{\partial x_{it}} (\sum (x_{it} - x_{jt})^2)$$

$$= \frac{1}{2\sqrt{\sum (x_{it} - x_{jt})^2}} \frac{\partial}{\partial x_{it}} (\sum (x_{it} - x_{jt})^2)$$

Since $\sum_{t=1}^{P}$ is not relevant except for the t-th dimension

$$= \frac{1}{2\sqrt{\sum (x_{it} - x_{jt})^2}} 2(x_{it} - x_{jt})$$

$$= \frac{x_{it} - x_{jt}}{d_{ij}}$$

Oh dear. We now have all the elements that differentiate the stress value η. Putting them together, we can draw the following conclusions.

$$\begin{aligned} \frac{\partial \eta}{\partial x_{it}} &= \frac{1}{2} \left(\frac{\eta}{A} \frac{\partial A}{\partial x_{it}} - \frac{\eta}{B} \frac{\partial B}{\partial x_{it}} \right) \\ &= \frac{1}{2} \frac{\eta}{A} 2 \sum_i \sum_j (d_{ij} - \hat{d}_{ij}) \cdot \frac{x_{it} - x_{jt}}{d_{ij}} - \frac{\eta}{B} 2 \sum_i \sum_j (d_{ij} - d) \frac{x_{it} - x_{jt}}{d_{ij}} \\ &= \eta \sum_i \sum_j \frac{x_{it} - x_{jt}}{d_{ij}} \left(\frac{d_{ij} - \hat{d}_{ij}}{A} - \frac{d_{ij} - d}{B} \right) \end{aligned}$$

*1 Takahashi[2] notes that this follows from the monotone regression principle. I believe that the monotone regression principle is a rule defining the correspondence between distance and disparity, and that these differential values being zero are not directly relevant to that principle.

This time we limited ourselves to the Euclidean distance, but when considering Minkowski's general distance, it is necessary to consider the sign of the direction of movement, among other factors. Additionally, note that Kruskal[1] considers a method to fine-tune the step width based on experience[*2].

参考文献

[1] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2), 115-129.

[2] 高橋和子. (1986). 多次元尺度法 –Kruskal の方法を中心に. 茨城大学政経学会雑誌, 51, 97-117.

[*2] Takahashi[2] precedes this with 2η, which is probably a typo.