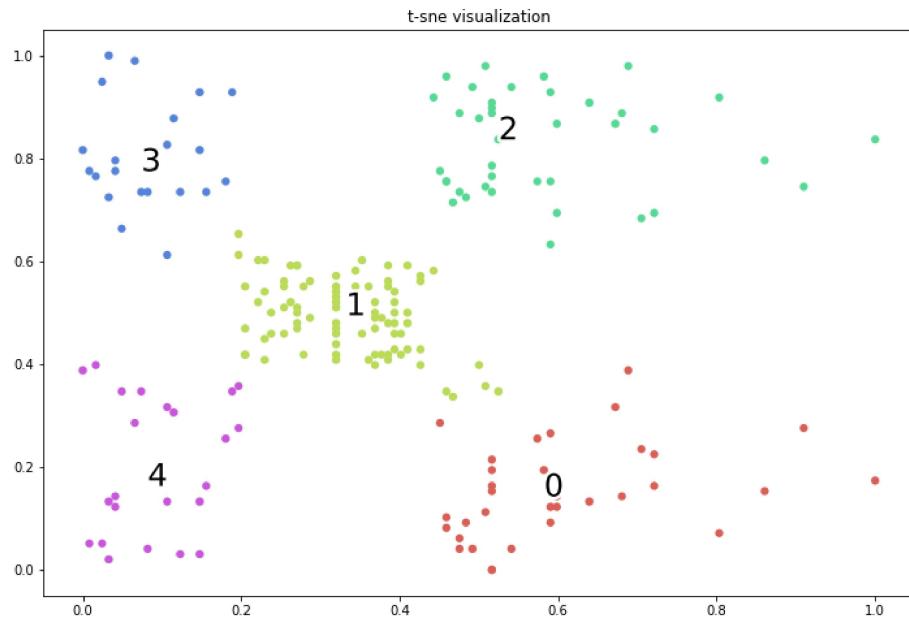


```

means:
[[[0.60502531 0.15433196]
[0.33368985 0.49394756]
[0.58393969 0.82673863]
[0.0829305 0.80743088]
[0.09861098 0.21597752]]
covariances:
[[[ [ 0.01818446 0.00433814]
[ 0.00433814 0.00873064]
[[ 0.00613567 -0.00231927]
[-0.00231927 0.0051635 ]]
[[ 0.01808598 -0.00031096]
[-0.00031096 0.0091568 ]]
[[ 0.00337483 -0.0001437 ]
[-0.0001437 0.01026088]]
[[ 0.00453005 0.00255303]
[ 0.00255303 0.01918353]]]
```

Hình ảnh phân bố của các cụm ( $K = 5$ ):



### 5.5 Tổng kết

GMM là một mô hình xác suất. Mô hình này thể hiện sự cải tiến so với K-Means, đó là các điểm dữ liệu được sinh ra từ một phân phối hỗn hợp của hữu hạn các phân phối Gaussian đo chiều. Tham số của những phân phối này được giả định là chưa biết. Để tìm ra tham số huấn luyện cho mô hình thì chúng ta tối đa hóa hàm auxiliary thông qua thuật toán EM, thuật toán này sẽ cập nhật nghiệm sau mỗi vòng lặp để di đến điểm cực trị. Chúng ta có thể coi rằng GMM như là một dạng khái quát của thuật toán K-Means nhằm kết hợp với thông tin về hiệp phương sai của dữ liệu cũng như là tâm của các phân phối Gaussian tiềm ẩn.

### 6 Chia tỉ lệ nhiều chiều

Giả sử có  $n$  quan sát trong không gian  $p$  chiều. Giữa mỗi cặp quan sát bất kì  $(r, s)$  lại có một phép đo  $\delta_{rs}$  biểu thị mức độ khác nhau giữa hai vật  $r$  và  $s$ . Giờ ta mong muốn biểu diễn các vật trên trong không gian để xem xét mối liên hệ của chúng. Đơn giản thì ta có thể coi mỗi quan sát thứ  $r$  là một điểm với các tọa độ  $(n_{r1}, n_{r2}, \dots, n_{rp})$ , nhưng làm thế thì chỉ có thể biểu diễn các điểm với điều kiện  $p \leq 3$ . Ngoài ra nếu  $p \leq 3$ , khi đó các điểm biểu diễn theo cách này lại không thể hiện được độ sai khác giữa mỗi cặp phần tử là  $\delta$ .

Chia tỉ lệ nhiều chiều (*Multidimensional Scaling* hay viết tắt là MDS) là tập các phương pháp để tìm một cấu hình trong không gian ít chiều hơn  $p$ , mỗi điểm trong không gian tương ứng cho một vật và khoảng cách giữa các vật là phù hợp với sự khác nhau  $\delta$ .

Với  $n$  vật thì ta có  $n(n - 1)/2$  giá trị của  $\delta$ . Nếu ta chỉ quan tâm đến thứ tự của các giá trị  $\delta$  để tìm cấu hình biểu diễn các vật, gọi là lớp phương pháp *Non-metric multidimensional scaling*. Nếu ta dùng đến các giá trị  $\delta$  để tìm cấu hình, đó là lớp phương pháp *metric multidimensional scaling*.

Ngoài ra, thay cho độ đo sự khác nhau  $\delta$  còn có phép đo sự giống nhau  $s$ . Tuy nhiên ta có thể biến đổi linh hoạt giữa hai dạng phép đo này tùy vào mục đích sử dụng cũng như yêu cầu cần giải quyết.

#### 6.1 Dùng khoảng cách trong chia tỉ lệ nhiều chiều và cách làm cổ điển

Giả sử có  $n$  vật với phép đo khác nhau  $\{\delta_{rs}\}$ . Metric MDS sẽ cố gắng tìm một tập hợp các điểm trong không gian, ở đó mỗi điểm tương ứng với một vật và khoảng cách giữa các điểm là  $\{d_{rs}\}$  sao cho

$$d_{rs} \approx f(\delta_{rs}) \quad (16)$$

Với  $f$  là hàm liên tục đơn điệu.

Gọi  $\mathbb{O}$  là tập chứa các vật với độ khác nhau  $\delta$  được định nghĩa trên tập  $\mathbb{O} \times \mathbb{O}$ . Với  $\phi$  là một ánh xạ từ tập  $\mathbb{O}$  vào tập  $\mathbb{E}$ , nơi chứa các điểm tương ứng cho các vật. Đặt  $\phi(r) = x_r$ , ( $r \in \mathbb{O}, x_r \in \mathbb{E}$ ), và  $\mathbb{X} = \{x_r : r \in \mathbb{O}\}$  là tập ảnh. Khoảng cách giữa hai điểm  $x_r, x_s$  là  $d_{rs}$ . Mục đích của Metric MDS là tìm ánh xạ  $\phi$  sao cho  $d_{rs}$  xấp xỉ bằng với  $f(\delta_{rs})$  với mọi  $r, s \in \mathbb{O}$ .

Chia tỉ lệ cổ điển có nguồn gốc từ những năm 1930 khi Young và Householder (1938) chỉ với ma trận khoảng cách giữa các điểm trong không gian Euclidean, tọa độ của các điểm có thể được tìm thấy sao cho khoảng cách được bảo toàn. Torgerson (1952) đã đưa chủ đề này trở nên phổ biến bằng cách sử dụng vào kỹ thuật chia tỷ lệ.

##### Khôi phục lại tọa độ

Cho  $n$  điểm trong không gian Euclidean  $p$  chiều. Tọa độ của điểm thứ  $r$  là  $\mathbf{x}_r = (x_{r1}, x_{r2}, \dots, x_{rp})^T$ , ( $r = 1, 2, \dots, n$ ). Do đó khoảng cách giữa hai điểm  $r$  và  $s$  được định nghĩa như sau:

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T (\mathbf{x}_r - \mathbf{x}_s) \quad (17)$$

Và ma trận tích trong  $\mathbf{B}$  với

$$[\mathbf{B}]_{rs} = b_{rs} = \mathbf{x}_r^T \mathbf{x}_s. \quad (18)$$

Từ khoảng cách đã biết  $\{d_{rs}\}$ , ta tìm được ma trận  $\mathbf{B}$ , rồi tìm lại được tọa độ các điểm.

##### Phương pháp tìm ma trận $\mathbf{B}$

Hiển nhiên khi có một cấu hình thỏa mãn trong không gian nào, bằng việc xoay và tịnh tiến ta lại

thu được các câu hình khác cũng thỏa mãn. Do vậy để tìm được một câu hình duy nhất, ta đặt ra điều kiện:

$$\sum_{r=1}^n x_{ri} = 0, (i = 1, 2, \dots, p). \quad (19)$$

Như vậy trọng tâm của câu hình là điểm  $(0, 0, \dots, 0)$ . Để tìm  $\mathbf{B}$ , trước tiên ta khai triển đẳng thức (17):

$$d_{rs}^2 = \mathbf{x}_r^T \mathbf{x}_r + \mathbf{x}_s^T \mathbf{x}_s - 2\mathbf{x}_r^T \mathbf{x}_s. \quad (20)$$

Kết hợp với (19) ta thu được các đẳng thức:

$$\frac{1}{n} \sum_{r=1}^n d_{rs}^2 = \frac{1}{n} \sum_{r=1}^n \mathbf{x}_r^T \mathbf{x}_r + \mathbf{x}_s^T \mathbf{x}_s \quad (21)$$

$$\frac{1}{n} \sum_{s=1}^n d_{rs}^2 = \mathbf{x}_r^T \mathbf{x}_r + \frac{1}{n} \sum_{s=1}^n \mathbf{x}_s^T \mathbf{x}_s \quad (22)$$

$$\frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n d_{rs}^2 = \frac{2}{n} \sum_{r=1}^n \mathbf{x}_r^T \mathbf{x}_r \quad (23)$$

Thay (21), (22), (23) vào (20)

$$\begin{aligned} d_{rs}^2 &= \mathbf{x}_r^T \mathbf{x}_r + \mathbf{x}_s^T \mathbf{x}_s - 2\mathbf{x}_r^T \mathbf{x}_s \\ &= \frac{1}{n} \sum_{r=1}^n d_{rs}^2 + \frac{1}{n} \sum_{s=1}^n d_{rs}^2 - \frac{2}{n} \sum_{r=1}^n \mathbf{x}_r^T \mathbf{x}_r - 2\mathbf{x}_r^T \mathbf{x}_s \\ &= \frac{1}{n} \sum_{r=1}^n d_{rs}^2 + \frac{1}{n} \sum_{s=1}^n d_{rs}^2 - \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n d_{rs}^2 - 2\mathbf{x}_r^T \mathbf{x}_s \end{aligned}$$

Như vậy

$$\begin{aligned} b_{rs} &= \mathbf{x}_r^T \mathbf{x}_s \\ &= -\frac{1}{2} \left( d_{rs}^2 - \frac{1}{n} \sum_{r=1}^n d_{rs}^2 - \frac{1}{n} \sum_{s=1}^n d_{rs}^2 + \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n d_{rs}^2 \right) \\ &= a_{rs} - a_{r.} - a_{.s} + a_{..} \end{aligned}$$

Trong đó:

$$\begin{aligned} a_{rs} &= -\frac{1}{2} d_{rs}^2 \\ a_{r.} &= n^{-1} \sum_s a_{rs} \\ a_{.s} &= n^{-1} \sum_r a_{rs} \\ a_{..} &= n^{-2} \sum_r \sum_s a_{rs} \end{aligned}$$

Đặt ma trận  $\mathbf{A}$  với  $[\mathbf{A}]_{rs} = a_{rs}$ . Ta thu được ma trận  $\mathbf{B}$  ở dưới dạng:

$$\mathbf{B} = \mathbf{H} \mathbf{A} \mathbf{H} \quad (24)$$

## 6 Chia tỉ lệ nhiều chiều

---

Với  $\mathbf{H}$  là ma trận nửa xác định dương:

$$\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$$

Trong đó  $\mathbf{I}$  là ma trận đơn vị và  $\mathbf{1} = (1, 1, \dots, 1)^T$  là vector có  $n$  số 1.

Đẳng thức (24) có thể chứng minh như sau:

Trước tiên ta khai triển:

$$\begin{aligned}\mathbf{H}\mathbf{A}\mathbf{H} &= (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbf{A}(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T) \\ &= \mathbf{I}\mathbf{A}\mathbf{I} - \mathbf{I}\mathbf{A}n^{-1}\mathbf{1}\mathbf{1}^T - n^{-1}\mathbf{1}\mathbf{1}^T\mathbf{A}\mathbf{I} + n^{-1}\mathbf{1}\mathbf{1}^T\mathbf{A}n^{-1}\mathbf{1}\mathbf{1}^T \\ &= \mathbf{A} - n^{-1}\mathbf{A}\mathbf{1}\mathbf{1}^T - n^{-1}\mathbf{1}\mathbf{1}^T\mathbf{A} + n^{-2}\mathbf{1}\mathbf{1}^T\mathbf{A}\mathbf{1}\mathbf{1}^T \\ &= \mathbf{C}\end{aligned}$$

Ta tính giá trị  $c_{rs}$  thông qua 4 số hạng ở vế phải. Tại vị trí hàng  $r$  cột  $s$  của:

$$\begin{aligned}\mathbf{A}_{rs} &= -\frac{1}{2}d_{rs}^2 \\ (n^{-1}\mathbf{A}\mathbf{1}\mathbf{1}^T)_{rs} &= -\frac{1}{2n} \sum_{s=1}^n d_{rs}^2 \\ (n^{-1}\mathbf{1}\mathbf{1}^T\mathbf{A})_{rs} &= -\frac{1}{2n} \sum_{r=1}^n d_{rs}^2 \\ (n^{-2}\mathbf{1}\mathbf{1}^T\mathbf{A}\mathbf{1}\mathbf{1}^T)_{rs} &= \frac{1}{2n^2} \sum_{r=1}^n \sum_{s=1}^n d_{rs}^2\end{aligned}$$

Từ đó suy ra:

$$\begin{aligned}c_{rs} &= -\frac{1}{2}d_{rs}^2 + \frac{1}{2n} \sum_{r=1}^n d_{rs}^2 + \frac{1}{2n} \sum_{s=1}^n d_{rs}^2 - \frac{1}{2n^2} \sum_{r=1}^n \sum_{s=1}^n d_{rs}^2 \\ &= -\frac{1}{2} \left( d_{rs}^2 - \frac{1}{n} \sum_{r=1}^n d_{rs}^2 - \frac{1}{n} \sum_{s=1}^n d_{rs}^2 + \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n d_{rs}^2 \right) \\ &= b_{rs}\end{aligned}$$

Vậy đẳng thức (24) đã được chứng minh.

### Khôi phục lại tọa độ từ ma trận $\mathbf{B}$

Ma trận tích trong  $\mathbf{B}$  có thể được viết lại dưới dạng:

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T$$

Trong đó  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$  là một ma trận cỡ  $n \times p$ . Ma trận  $\mathbf{B}$  là ma trận đối xứng nên viết được dưới dạng:

$$\mathbf{B} = \mathbf{V}\Lambda\mathbf{V}^T$$

Trong đó  $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$  là ma trận đường chéo chứa các giá trị riêng  $\lambda$  của  $\mathbf{B}$  và  $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]$  là ma trận chứa các vector riêng trực chuẩn.

Từ đẳng thức  $\mathbf{B} = \mathbf{XX}^T$ , kèm theo điều kiện  $\mathbf{B}$  là ma trận nửa xác định dương thì:

$$\mathbf{X} = \mathbf{V}\Lambda^{\frac{1}{2}} \quad (25)$$

Nếu ma trận  $\mathbf{B}$  không là nửa xác định dương thì ta có thể loại bỏ các giá trị riêng âm:

$$\mathbf{X} = \mathbf{V}_1\Lambda_1^{\frac{1}{2}} \quad (26)$$

Với  $\Lambda_1$  là ma trận chéo chỉ chứa các giá trị riêng không âm và  $\mathbf{V}_1$  là ma trận chứa các vector riêng trực chuẩn tương ứng. Tuy nhiên việc loại bỏ các giá trị riêng âm sẽ khiến cấu hình mà ta thu được có thể trở nên thiếu chính xác hơn.

Số chiều của cấu hình thu được là số hàng của ma trận giá trị riêng  $\Lambda$  hay  $\Lambda_1$ . Nếu ma trận giá trị riêng này có chứa giá trị riêng 0, thì tương ứng ma trận  $\mathbf{X}$  sẽ có cột toàn là 0. Hay nói cách khác số chiều của cấu hình là số các vector riêng khác 0 trong ma trận giá trị riêng.

Như vậy để có một cấu hình với số chiều bất  $q$ , ta tạo một ma trận  $\Lambda_q$  chứa  $q$  giá trị riêng không âm và ma trận  $\mathbf{V}_q$  chứa  $q$  vector riêng tương ứng. Tọa độ các điểm trong cấu hình thu được qua công thức:

$$\mathbf{X} = \mathbf{V}_q\Lambda_q^{\frac{1}{2}}$$

Do ma trận  $\mathbf{H}$  có giá trị riêng 0 với vector riêng  $\mathbf{1}$  nên ma trận  $\mathbf{B}$  cũng có giá trị riêng 0. Vậy cấu hình thu được có tối đa  $n - 1$  chiều.

### Thêm hằng số vào $d_{rs}$ để ma trận $\mathbf{B}$ nửa xác định dương

Như đã thấy ở trên, nếu  $\mathbf{B}$  có giá trị riêng âm, các giá trị trong ma trận  $\mathbf{X}$  theo công thức (25) sẽ chứa các số phức. Nói cách khác là cấu hình mà ta đang tìm không thuộc trong không gian Euclidean. Có nhiều phương pháp thêm hằng số để giúp ma trận  $\mathbf{B}$  trong công thức (24) là nửa xác định dương. Trong tài liệu này tác giả trình bày phương pháp khá đơn giản của Francis Cailliez, bạn đọc có thể tìm hiểu thêm tại [6].

Phương pháp của Cailliez là tìm một hằng số  $c^*$  nhỏ nhất sao cho mọi phép đo  $d^{(c)}$  được định nghĩa:

$$d_{rs}^{(c)} = \begin{cases} d_{rs} + c & (r \neq s) \\ 0 & (r = s) \end{cases} \quad (27a)$$

$$(27b)$$

Đều có một cấu hình đại diện trong không gian Euclidean với mọi  $c \geq c^*$ . Để thuận tiện cho việc biến đổi cũng như chứng minh, ta đặt:

$$\mathbf{B}_d = \mathbf{H}\mathbf{A}_d\mathbf{H}$$

Với giả thiết  $\mathbf{B}_d$  không là nửa xác định dương.

Trong đó  $\mathbf{A}_d$  chứa các phần tử dưới dạng  $a_{rs} = -\frac{1}{2}d_{rs}^2$ . Như vậy ta có:

$$\begin{aligned} \mathbf{B}_{d^{(c)}} &= \mathbf{H}\mathbf{A}_{d^{(c)}}\mathbf{H} \\ &= -\frac{1}{2}\mathbf{H} \begin{bmatrix} 0 & (d_{12} + c)^2 & \dots & (d_{1n} + c)^2 \\ (d_{21} + c)^2 & 0 & \dots & (d_{2n} + c)^2 \\ \dots & \dots & \dots & \dots \\ (d_{n1} + c)^2 & (d_{n2} + c)^2 & \dots & 0 \end{bmatrix} \mathbf{H} \\ &= \mathbf{B}_d + 2c\mathbf{B}_{d^{\frac{1}{2}}} - \frac{1}{2}\mathbf{H} \begin{bmatrix} 0 & c^2 & \dots & c^2 \\ c^2 & 0 & \dots & c^2 \\ \dots & \dots & \dots & \dots \\ c^2 & c^2 & \dots & 0 \end{bmatrix} \mathbf{H} \end{aligned}$$

## 6 Chia tỉ lệ nhiều chiều

---

Ta biến đổi số hạng cuối cùng trong về phải của đẳng thức trên như sau:

$$\begin{aligned} -\frac{1}{2}\mathbf{H} \begin{bmatrix} 0 & c^2 & \dots & c^2 \\ c^2 & 0 & \dots & c^2 \\ \dots & \dots & \dots & \dots \\ c^2 & c^2 & \dots & 0 \end{bmatrix} \mathbf{H} &= -\frac{c^2}{2} (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T) (\mathbf{1}\mathbf{1}^T - \mathbf{I}) \mathbf{H} \\ &= \frac{c^2}{2} (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T) \mathbf{H} \\ &= \frac{c^2}{2} \mathbf{H}^2 = \frac{c^2}{2} \mathbf{H} \end{aligned}$$

Như vậy ta thu được:

$$\mathbf{B}_{d^{(c)}} = \mathbf{B}_d + 2c\mathbf{B}_{d^{\frac{1}{2}}} + \frac{c^2}{2}\mathbf{H} \quad (28)$$

Đẳng thức (28) giúp ta dễ dàng chứng minh định lý sau:

**Định lý 1.** *Tồn tại một hằng số  $c^*$  sao cho mọi phép đo  $d^{(c)}$  được định nghĩa bởi (27a) (27b) đều có một đại diện (cấu hình) trong không gian Euclide với mọi  $c \geq c^*$ . Ngoài ra cấu hình với phép đo  $d^{(c^*)}$  có nhiều nhất ( $n - 2$ ) chiều ( $n$  là số quan sát).*

Trước tiên ta chứng minh tồn tại hằng số  $c^*$ . Ta muốn phép đo  $d^{(c)}$  cho một cấu hình trong không gian Euclide hay với mọi vector cột  $\mathbf{x}$  trong không gian đều thoả mãn:

$$\begin{aligned} \mathbf{x}^T \mathbf{B}_{d^{(c)}} \mathbf{x} &\geq 0 \\ \Leftrightarrow \mathbf{x}^T \mathbf{B}_d \mathbf{x} + \mathbf{x}^T 2c \mathbf{B}_{d^{\frac{1}{2}}} \mathbf{x} + \mathbf{x}^T \frac{c^2}{2} \mathbf{H} \mathbf{x} &\geq 0 \end{aligned}$$

Gọi  $\lambda_n, \mu_n$  lần lượt là giá trị riêng nhỏ nhất của ma trận  $\mathbf{B}_d$  và  $\mathbf{B}_{d^{\frac{1}{2}}}$ . Do  $\mathbf{B}_d$  không là nửa xác định dương nên  $\lambda_n < 0$ . Gọi  $\Lambda_d$  là ma trận giá trị riêng của  $\mathbf{B}_d$  và  $\mathbf{V}_d$  là ma trận vector riêng tương ứng. Với mọi  $\mathbf{x}$  trong không gian Euclide:

$$\begin{aligned} \mathbf{x}^T \mathbf{B}_d \mathbf{x} &= \mathbf{x}^T \mathbf{V}_d \Lambda_d \mathbf{V}_d^T \mathbf{x} \\ &= (\mathbf{V}_d^T \mathbf{x})^T \Lambda_d (\mathbf{V}_d^T \mathbf{x}) \\ &= \mathbf{u}^T \Lambda_d \mathbf{u} \geq \lambda_n \mathbf{u}^T \mathbf{I} \mathbf{u} \end{aligned}$$

Mà  $\lambda_n \mathbf{u}^T \mathbf{I} \mathbf{u} = \lambda_n \mathbf{x}^T \mathbf{V}_d \mathbf{I} \mathbf{V}_d^T \mathbf{x} = \lambda_n \mathbf{x}^T \mathbf{x}$ . Do cách định nghĩa ma trận  $\mathbf{H}$ , ta lại có:

$$\mathbf{x}^T \mathbf{x} \geq \mathbf{x}^T \mathbf{H} \mathbf{x}$$

Tóm lại ta có bất đẳng thức sau:

$$\mathbf{x}^T \mathbf{B}_d \mathbf{x} \geq \lambda_n \mathbf{x}^T \mathbf{H} \mathbf{x}$$

Tương tự, ta cũng chứng minh được:

$$\mathbf{x}^T \mathbf{B}_{d^{\frac{1}{2}}} \mathbf{x} \geq \mu_n \mathbf{x}^T \mathbf{H} \mathbf{x}$$

Từ đó ta suy ra:

$$\mathbf{x}^T \mathbf{B}_{d(c)} \mathbf{x} \geq \left( \lambda_n + 2c\mu_n + \frac{c^2}{2} \right) \mathbf{x}^T \mathbf{H} \mathbf{x}$$

Vậy để  $\mathbf{B}_{d(c)}$  là nửa xác định dương, do  $\mathbf{H}$  là nửa xác định dương nên ta chỉ cần:

$$\left( \lambda_n + 2c\mu_n + \frac{c^2}{2} \right) \geq 0$$

Do  $c > 0$  nên:

$$c \geq -2\mu_n + (4\mu_n^2 - 2\lambda_n)^{\frac{1}{2}}$$

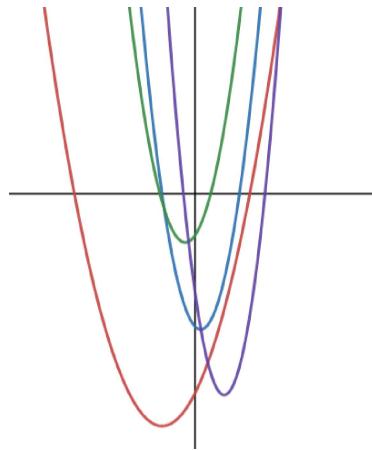
Hay nói cách khác,  $c^*$  tồn tại và bị chặn bởi một số dương:

$$c^* \leq -2\mu_n + (4\mu_n^2 - 2\lambda_n)^{\frac{1}{2}} \quad (29)$$

Như vậy, ta có thể thêm hằng số  $c$  lớn hơn  $c^*$  trong công thức (29). Tuy nhiên ta mới chỉ khẳng định  $c^*$  bị chặn và chưa tìm được giá trị nhỏ nhất của  $c^*$ .

Xét đẳng thức (28), ta thấy  $\mathbf{x}^T \mathbf{B}_{d(c)} \mathbf{x}$  là một hàm của  $c$  có đồ thị dạng parabol lồi. Với mỗi  $\mathbf{x}$  ta lại được một parabol. Với mỗi parabol,  $c$  chỉ cần lớn hơn  $\alpha(x)$  (giao điểm bên phải của parabol với trục hoành) thì giá trị tại  $c$  sẽ lớn hơn hoặc bằng 0. Ngoài ra parabol giao với trục tung tại điểm có tung độ là  $\mathbf{x}^T \mathbf{B}_d \mathbf{x}$ .

Xét tập các vector  $\mathbf{x}$  làm cho  $\mathbf{x}^T \mathbf{B}_d \mathbf{x} < 0$ . Khi đó lớp các parabol  $\mathbf{x}^T \mathbf{B}_{d(c)} \mathbf{x}$  có dạng:



Như vậy  $\alpha(x)$  trong các lớp parabol trên là dương. Kết hợp với (29) ta có thể chọn  $c^*$ :

$$c^* = \sup_{\mathbf{x}' \mathbf{B}_d \mathbf{x} < 0} \alpha(\mathbf{x}) = \alpha(\mathbf{x}^*) \quad (30)$$

Từ đó, ta thu được:

$$\begin{aligned} \mathbf{x}^T \mathbf{B}_{d(c)} \mathbf{x} &\geq 0, c \geq c^* \\ (\mathbf{x}^*)^T \mathbf{B}_{d(c^*)} \mathbf{x}^* &= 0 \end{aligned}$$

## 6 Chia tì lệ nhiều chiều

Với mọi  $c \geq c^*$  thì phép đo  $d^{(c)}$  có một cấu hình đại diện trong không gian Euclidean. Phép đo  $d^{(c^*)}$  có cấu hình tối đa  $(n - 2)$  chiều do  $\mathbf{B}_d^{(c^*)}$  có hai vector riêng ứng với giá trị riêng 0 là  $\mathbf{x}^*$  và  $\mathbf{1}$ . Vậy định lý 1 được chứng minh.

Giờ ta đi tìm  $c^*$  trong công thức (30). Chú ý rằng  $\mathbf{1}\mathbf{1}^T\mathbf{x} = k\mathbf{1}$  với số  $k$  nào đó:

$$\left( \mathbf{B}_d + 2c^*\mathbf{B}_{d^{\frac{1}{2}}} + \frac{(c^*)^2}{2}\mathbf{H} \right) \mathbf{Hx}^* = 0$$

Đặt  $2\mathbf{B}_d\mathbf{Hx}^* = c^*\mathbf{y}$ , ta viết lại đẳng thức trên dưới dạng:

$$\mathbf{y} + 4\mathbf{B}_{d^{\frac{1}{2}}}\mathbf{Hx}^* + c^*\mathbf{Hx}^* = 0$$

Từ những đẳng thức trên, rất thú vị khi ta có thể viết được hệ phương trình sau:

$$\begin{pmatrix} 0 & 2\mathbf{B}_d \\ -\mathbf{I} & -4\mathbf{B}_{d^{\frac{1}{2}}} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{Hx}^* \end{pmatrix} = c^* \begin{pmatrix} \mathbf{y} \\ \mathbf{Hx}^* \end{pmatrix}$$

Như vậy  $c^*$  là một giá trị riêng của ma trận:

$$\mathbf{W} = \begin{pmatrix} 0 & 2\mathbf{B}_d \\ -\mathbf{I} & -4\mathbf{B}_{d^{\frac{1}{2}}} \end{pmatrix}$$

Xét  $a$  là một giá trị riêng của  $\mathbf{W}$  với vector riêng  $[\mathbf{z} \quad \mathbf{t}]^T$ , ta có  $\mathbf{t}^T\mathbf{B}_{d(a)}\mathbf{t} = 0$ . Do  $c^* = \sup_{\mathbf{x}^T\mathbf{B}_d\mathbf{x} < 0} \alpha(x)$  nên  $c^* \geq a$ , hay nói cách khác  $c^*$  là giá trị riêng thực lớn nhất của ma trận  $\mathbf{W}$ .

### Lựa chọn cấu hình thích hợp

Xét ma trận  $\mathbf{B}$  là nửa xác định dương. Từ đẳng thức (25) và  $\mathbf{B}$  luôn có giá trị riêng 0, cấu hình ta thu được có nhiều nhất  $n - 1$  chiều. Nếu số chiều của cấu hình lớn thì việc giảm số chiều dữ liệu sẽ ít hiệu quả do ta khó quan sát được các đặc điểm của dữ liệu. Do đó ta thường tìm một hình chiểu của cấu hình ban đầu lên không gian  $q$  chiều ( $q$  thường là 2 hoặc 3). Tọa độ trong không gian  $q$  chiều có dạng:

$$\mathbf{X} = \mathbf{V}_q \Lambda_q^{\frac{1}{2}}$$

Với  $\Lambda_q$  là ma trận đường chéo cỡ  $q \times q$  chứa  $q$  giá trị riêng của ma trận  $\mathbf{B}$ , tương ứng  $\mathbf{V}_q$  là ma trận cỡ  $n \times q$  chứa các vector riêng. Gọi khoảng cách giữa các điểm trong không gian  $q$  chiều là  $d^*$ . Ta chọn  $q$  giá trị riêng và  $q$  vector riêng tương ứng sao cho biểu thức dưới đây đạt giá trị nhỏ nhất:

$$\sum_{r=1}^n \sum_{s=1}^n (d_{rs}^2 - (d_{rs}^*)^2) \tag{31}$$

Đặt  $\mathbf{B}_q = \mathbf{V}_q \Lambda_q \mathbf{V}'_q$ . Chú ý đẳng thức (23):

$$\sum_{r=1}^n \sum_{s=1}^n (d_{rs}^2 - (d_{rs}^*)^2) = n (tr\mathbf{B} - tr\mathbf{B}_q)$$

Vậy (31) đạt giá trị nhỏ nhất nếu  $tr\mathbf{B}_q$  lớn nhất, hay  $\Lambda_q$  chứa  $q$  giá trị riêng lớn nhất của  $\mathbf{B}$ .

Để đánh giá độ tốt của cấu hình, bạn đọc có thể tham khảo hàm Stress tại 6.2.

### Các bước thực hiện của thuật toán chia tỉ lệ cổ điển

Có rất nhiều công đoạn mà tác giả đã giới thiệu ở trên, thuật toán có thể tóm gọn lại thành một số bước sau:

---

#### Algorithm 6 Thuật toán chia tỉ lệ cổ điển

---

- 1: Tính các  $\delta_{rs}$ .
  - 2: Tìm ma trận  $\mathbf{A} = \left[ -\frac{1}{2}\delta_{rs}^2 \right]$ .
  - 3: Tìm ma trận  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ .
  - 4: Tìm giá trị riêng của  $\mathbf{B}$  và các vector riêng tương ứng. Nếu  $\mathbf{B}$  không là nửa xác định dương có thể dùng kỹ thuật thêm hằng số và quay lại bước 2.
  - 5: Chọn không gian tốt nhất với số chiều  $p$  mong muốn.
  - 6: Tính toán tọa độ các điểm trong không gian  $p$  chiều và vẽ biểu đồ.
- 

Với cách làm này, khoảng cách giữa điểm  $r$  và  $s$  trong cấu hình thu được từ ma trận  $\mathbf{B}$  là  $\delta_{rs}$ .

## 6.2 Dùng thứ hạng trong chia tỉ lệ nhiều chiều và cách tiếp cận của Kruskal

Cho tập các vật thuộc tập  $\mathbb{O}$  và phép đo độ khác nhau giữa hai vật  $r$  và  $s$  thuộc  $\mathbb{O}$  là  $\delta_{rs}$ . Gọi  $\phi$  là ánh xạ từ tập  $\mathbb{O}$  vào tập  $\mathbf{X}$  với  $\mathbf{X}$  là tập con của không gian được sử dụng để biểu diễn các vật. Vật  $r, s$  tương đương với hai điểm  $x_r$  và  $x_s$  trong  $\mathbf{X}$  với khoảng cách giữa hai điểm là  $d_{x_r x_s}$ . Ta còn định nghĩa một hàm đo độ lệch là  $\hat{d}$  trên tập  $\mathbb{O} \times \mathbb{O}$  dùng để đo sự khớp của  $d_{x_r x_s}$  với độ khác nhau  $\delta_{rs}$ . Mục tiêu của lớp phương pháp *Nonmetric Multidimensional Scaling* là tìm ánh xạ  $\phi$  sao cho  $d_{x_r x_s}$  xấp xỉ bằng với  $\hat{d}_{rs}$ , và thường được tìm bởi một hàm mất mát nào đó. Các điểm thuộc  $\mathbf{X}$  cùng với các khoảng cách giữa các điểm tạo nên một cấu hình.

Tập  $\mathbf{X}$  có thể là tập con của  $\mathbb{R}^2$  với khoảng cách Euclidean, hoặc là tập con của  $\mathbb{R}^3$  với khoảng cách Minkowski,... Với phép đo độ khác nhau  $\delta$  xác định và các phương pháp để tính độ lệch  $\hat{d}$ , vấn đề Nonmetric MDS trở thành vấn đề tìm thuật toán làm cực tiểu hàm mất mát.

Ngoài ra, ta chỉ sử dụng thứ hạng của các phép đo độ khác nhau  $\delta_{rs}$  cho phương pháp này. Đó cũng là lý giải của từ "Nonmetric".

Thông thường, khoảng cách giữa các điểm trong  $\mathbf{X}$  là khoảng cách Minkowski. Với hai điểm  $r$  và  $s$  trong  $\mathbf{X}$  có  $p$  chiều thì khoảng cách giữa chúng tính bởi:

$$d_{rs} = \left[ \sum_{i=1}^p |x_{ri} - x_{si}|^\lambda \right]^{\frac{1}{\lambda}} \quad (\lambda > 0)$$

Với điểm  $r$  có tọa độ là  $\mathbf{x}_r = (x_{r1}, \dots, x_{rp})^T$ .

Tập các độ lệch  $\{\hat{d}_{rs}\}$  được xem như là hàm của  $\{d_{rs}\}$ :

$$\hat{d}_{rs} = f(d_{rs})$$

Với  $f$  là hàm đơn điệu sao cho:

$$\delta_{rs} < \delta_{tu} \Rightarrow \hat{d}_{rs} \leq \hat{d}_{tu} \quad (32)$$

Trong phần này, tác giả trình bày phương pháp của Joseph Bernard Kruskal. Độc giả có thể đọc tìm hai bài báo [7] và [8] để tìm hiểu chi tiết hơn.

## 6 Chia tỉ lệ nhiều chiều

### Định nghĩa hàm măt măt

Sử dụng các ký hiệu mà tác giả đã nêu trên, hàm măt măt  $S$  (viết tắt của Stress) được Kruskal định nghĩa như sau:

$$S = \sqrt{\frac{S^*}{T^*}} \quad (33)$$

Trong đó  $S^* = \sum_{r < s} (d_{rs} - \hat{d}_{rs})^2$  và  $T^* = \sum_{r < s} d_{rs}^2$ . Với cách định nghĩa này, tập  $\{\delta_{rs}\}$  "nhảy vào"  $S$  thông qua điều kiện (32).

Nhờ có  $S$ , ta có thể đánh giá độ tốt của cấu hình như sau:

Bảng đánh giá độ phù hợp:

Stress (%)	Goodness of fit
20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

### Cách xác định $\hat{d}_{rs}$

Hiển nhiên cho một tập khoảng cách trong cấu hình  $\{d_{rs}\}$ , cách xác định tập  $\{\hat{d}_{rs}\}$  sẽ cho ra các giá trị của  $S^*$  khác nhau. Điều này sẽ làm ảnh hưởng đến quá trình tìm cực tiểu của  $S$  và dẫn đến nhiều cấu hình "tốt" theo các cách định nghĩa tập  $\{\hat{d}_{rs}\}$ . Để tránh khỏi vấn đề này cũng như tìm được cấu hình ứng ý nhất, ta chọn tập  $\{\hat{d}_{rs}\}$  sao cho  $S^*$  đạt cực tiểu với tập  $\{d_{rs}\}$  cho trước.

Để thuận tiện, ta gán nhãn lại tập  $\{\delta_{rs}\}$  thành tập  $\{\delta_i : i = 1, \dots, N\}$  được sắp theo thứ tự tăng dần. Tương tự ta có tập các khoảng cách  $\{d_i : i = 1, \dots, N\}$  với  $d_i$  tương đương với  $\delta_i$ .

### Ví Dụ

Ta có bốn vật với độ khác nhau giữa các vật như sau:

$$\delta_{12} = 2.1, \delta_{13} = 3.0, \delta_{14} = 2.4, \delta_{23} = 1.7, \delta_{24} = 3.9, \delta_{34} = 3.2$$

Và bốn điểm trong cấu hình đại diện với khoảng cách:

$$d_{12} = 3.3, d_{13} = 4.5, d_{14} = 5.7, d_{23} = 3.3, d_{24} = 4.3, d_{34} = 1.3$$

Ta gán nhãn lại  $\{d_{rs}\}$  và  $\{\delta_{rs}\}$ :

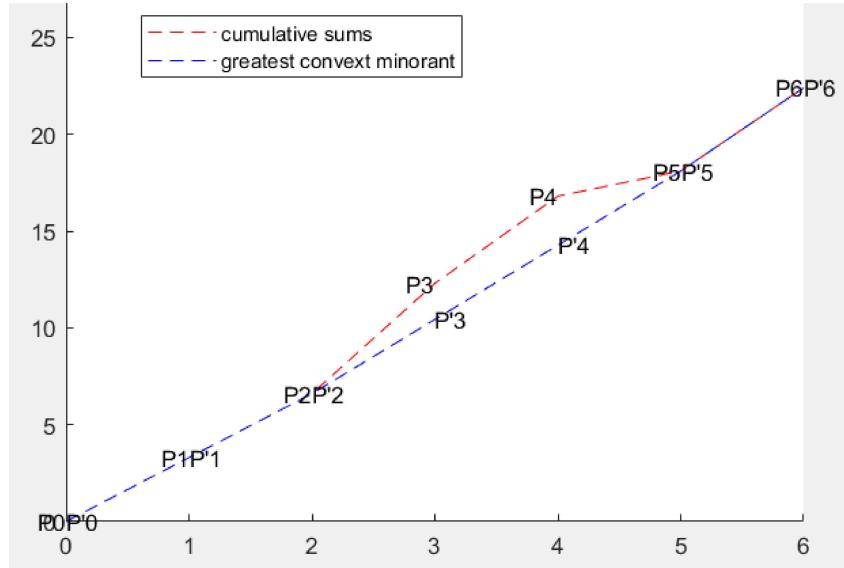
$$\delta_1 = 1.7, \delta_2 = 2.1, \delta_3 = 2.4, \delta_4 = 3.0, \delta_5 = 3.2, \delta_6 = 3.9$$

$$d_1 = 3.3, d_2 = 3.3, d_3 = 5.7, d_4 = 4.5, d_5 = 1.3, d_6 = 4.3$$

Cực tiểu hóa  $S$  tương đương với việc cực tiểu hóa  $S^* = \sum_i (d_i - \hat{d}_i)^2$ . Gọi  $D_i$  là tổng tích lũy của  $d_i$ :

$$D_i = \sum_{j=1}^i d_j \quad (i = 1, \dots, N)$$

Ta vẽ các điểm có tọa độ  $(0, 0)$ ,  $(i, D_i)$  và nối các điểm liên tiếp trên đồ thị và gán nhãn các điểm là  $P_0, \dots, P_N$ . Độ dốc của đoạn thẳng nối  $P_{i-1}$  với  $P_i$  là  $d_i$ . Hàm lồi tốt nhất (The greatest convex minorant) của các tổng tích lũy là cận trên bé nhất của tất cả các hàm lồi có đồ thị nằm dưới đồ thị của các tổng tích lũy. Hình sau minh họa cho ví dụ trên:



**Hình 26:** Đồ thị của các tổng tích lũy và hàm lồi tốt nhất

Ta có các điểm  $P'_0, \dots, P'_N$  là các điểm nằm trên hàm lồi tốt nhất với tọa độ là  $(0, 0)$  và  $(i, D'_i)$ . Ta xác định  $\{\hat{d}_{rs}\}$  như sau:

$$\begin{aligned}\hat{d}_1 &= D'_1 \\ \hat{d}_i &= D'_i - D'_{i-1}\end{aligned}$$

Ngoài ra, do hàm các tổng tích lũy là đồng biến và tính chất của hàm lồi tốt nhất, nếu  $D'_i < D_i$  thì  $\hat{d}_i = \hat{d}_{i+1}$ .

Giờ ta đi chứng minh cách xác định tập  $\{\hat{d}_{rs}\}$  như bên trên sẽ làm cực tiểu hóa  $S^*$  với tập  $\{d_{rs}\}$  cố định. Giả sử ta có tập đo độ lệch khác là  $\{d_{rs}^*\}$  cũng thỏa mãn điều kiện (32). Ta phải chỉ ra rằng:

$$\sum_{i=1}^N (d_i - d_i^*)^2 \geq \sum_{i=1}^N (d_i - \hat{d}_i)^2 \quad (34)$$

Đặt

$$\begin{aligned}D_i^* &= \sum_{j=1}^i d_j^* \\ D'_i &= \sum_{j=1}^i \hat{d}_j\end{aligned}$$

## 6 Chia tỉ lệ nhiều chiều

Công thức Abel:

$$\sum_{i=1}^N a_i b_i = \sum_{i=1}^{N-1} A_i (b_i - b_{i+1}) + A_N b_N$$

Trong đó  $A_i = \sum_{j=1}^i a_j$ .

Xét:

$$\begin{aligned} \sum_{i=1}^N (d_i - d_i^*)^2 &= \sum_{i=1}^N \left[ (d_i - \hat{d}_i) + (\hat{d}_i - d_i^*) \right]^2 \\ &= \sum_{i=1}^N (d_i - \hat{d}_i)^2 + \sum_{i=1}^N (\hat{d}_i - d_i^*)^2 + 2 \sum_{i=1}^N (d_i - \hat{d}_i)(\hat{d}_i - d_i^*) \end{aligned}$$

Áp dụng công thức Abel cho số hạng cuối ở vế phải:

$$\begin{aligned} &\sum_{i=1}^N (d_i - \hat{d}_i)(\hat{d}_i - d_i^*) \\ &= \sum_{i=1}^{N-1} (D_i - D'_i)(\hat{d}_i - \hat{d}_{i+1}) - \sum_{i=1}^{N-1} (D_i - D'_i)(d_i^* - d_{i+1}^*) + (D_N - D'_N)(\hat{d}_N - d_N^*) \end{aligned}$$

Có  $D_N - D'_N = 0$  do điểm cuối của hàm lồi tốt nhất và  $P_N$  là trùng nhau. Giờ ta xét  $(D_i - D'_i)(\hat{d}_i - \hat{d}_{i+1})$ . Nếu  $D_i = D'_i$  thì ta có số hạng thứ  $i$  bằng 0. Trường hợp còn lại là  $D_i > D'_i$ , lúc này  $\hat{d}_i = \hat{d}_{i+1}$  nên số hạng thứ  $i$  cũng bằng 0. Hay nói cách khác là  $\sum_{i=1}^{N-1} (D_i - D'_i)(\hat{d}_i - \hat{d}_{i+1}) = 0$ . Ta có  $d_i^* \leq d_{i+1}^*$  nên  $-\sum_{i=1}^{N-1} (D_i - D'_i)(d_i^* - d_{i+1}^*)$  là một số dương. Tóm lại:

$$\begin{aligned} \sum_{i=1}^N (d_i - d_i^*)^2 &\geq \sum_{i=1}^N (d_i - \hat{d}_i)^2 + \sum_{i=1}^N (\hat{d}_i - d_i^*)^2 \\ \Leftrightarrow \sum_{i=1}^N (d_i - d_i^*)^2 &\geq \sum_{i=1}^N (d_i - \hat{d}_i)^2 \end{aligned}$$

Vậy bất đẳng thức (34) được chứng minh. Quay lại ví dụ trên, ta xác định:

$$\hat{d}_1 = \hat{d}_2 = 3.3, \hat{d}_3 = \hat{d}_4 = \hat{d}_5 = \frac{23}{6}, \hat{d}_6 = 4.3$$

Lúc này  $S = 0.33$ .

### Cấu hình làm cực tiểu Stress

Giờ ta muốn tìm một cấu hình trong không gian  $p$  chiều với khoảng cách Minkowski  $\lambda > 0$ . Ta có định giá trị của  $p$  và  $\lambda$ .

Tất cả các điểm trong cấu hình đều có thể miêu tả như một vector (một điểm) trong không gian  $np$  chiều (còn gọi là không gian cấu hình) với tọa độ  $x_{il}$  trong đó  $i = 1$  tới  $n$  và  $l = 1$  đến  $p$  là tất cả các tọa độ của các điểm trong cấu hình.

$$\mathbf{x} = (x_{11}, \dots, x_{1p}, x_{21}, \dots, x_{n1}, \dots, x_{np})$$

Với mỗi điểm trong không gian cầu hình, hay với mỗi cầu hình, lại cho một giá trị của  $S$ . Nói cách khác  $S$  là một hàm:

$$S = S(x_{11}, \dots, x_{1p}, x_{21}, \dots, x_{n1}, \dots, x_{np})$$

Vấn đề bây giờ là tìm một điểm trong không gian cầu hình làm cực tiểu hóa hàm  $S$ , hay nói cách khác là tìm cực tiểu của hàm số nhiều biến. Ta sẽ dùng phương pháp "method of steepest descent" hay "method of gradients". Ta bắt đầu từ một điểm, di chuyển nó một chút theo hướng ngược với gradient tại đó (hướng giảm nhanh nhất). Gradient tại một điểm là một vector:

$$\left( \frac{\partial S}{\partial x_{11}}, \dots, \frac{\partial S}{\partial x_{1p}}, \dots, \frac{\partial S}{\partial x_{np}} \right)$$

Vector trên luôn dương, đi theo vector đối sẽ tới điểm cực tiểu. Lặp lại quá trình trên, sẽ đến một điểm mà vector gradient tại đó bằng vector 0 (hoặc gần bằng), đó là điểm làm cực tiểu  $S$  hay là cầu hình cần tìm.

Cách làm trên sẽ cho ta một điểm cực tiểu địa phương trong không gian cầu hình, chưa chắc là cực tiểu toàn cục. Để khắc phục điều này, ta có thể bắt đầu từ nhiều điểm xuất phát và chọn điểm cực tiểu nhỏ nhất rồi hy vọng nó là cực tiểu toàn cục. Tất nhiên Stress đủ nhỏ thì ta vẫn có một cầu hình đủ tốt mà không cần bận tâm liệu  $S$  đã nhỏ nhất chưa.

Điểm bắt đầu có rất nhiều cách lựa chọn, như tọa độ của điểm tuân theo phân phôi đều liên tục trong đoạn  $[-1; 1]$  hay tuân theo phân phôi Poisson,...

Gọi  $g$  là gradient tại điểm  $x$  trong không gian cầu hình,  $\alpha$  là hệ số nhảy. Cầu hình tiếp theo được tìm thông qua công thức:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \frac{\alpha}{\text{mag}(g)} g \quad (35)$$

Trong đó:

$$\text{mag}(g) = \sqrt{\frac{\sum_{r,s} g_{rs}^2}{\sum_{r,s} x_{rs}^2}}$$

Giá trị bắt đầu của  $\alpha = 0.2$  và thay đổi sau mỗi lần lặp. Cụ thể:

$$\alpha_{\text{present}} = \alpha_{\text{previous}} \cdot (\text{angle factor}) \cdot (\text{relaxation factor}) \cdot (\text{good luck factor})$$

$$\theta = \frac{\sum_{r,s} g_{rs} g''_{rs}}{\sqrt{\sum_{r,s} g_{rs}^2} \sqrt{\sum_{r,s} g''_{rs}}} \quad (g'' \text{ là gradient trước đó, } g \text{ là gradient hiện tại})$$

$$\text{angle factor} = 4.0^{\cos^3 \theta}$$

$$\text{relaxation factor} = \frac{1.3}{1 + (5\text{step ratio})^5}$$

$$5 \text{ step ratio} = \min \left[ 1, \left( \frac{\text{present stress}}{\text{stress 5 iterations ago}} \right) \right]$$

$$\text{good luck factor} = \min \left[ 1, \frac{\text{present stress}}{\text{previous stress}} \right]$$

Cách chọn  $\alpha$  như trên dựa trên kinh nghiệm của Kruskal và không có bằng chứng rằng cách chọn như trên là tối ưu.

## 6 Chia tỉ lệ nhiều chiều

Giờ ta tính đạo hàm riêng của  $S$  theo  $x_{ui}$  bất kỳ:

$$\begin{aligned}\frac{\partial S}{\partial x_{ui}} &= \frac{1}{2} \sqrt{\frac{T^*}{S^*}} \frac{\left(T^* \frac{\partial S^*}{\partial x_{ui}} - S^* \frac{\partial T^*}{\partial x_{ui}}\right)}{(T^*)^2} \\ &= \frac{1}{2} S \left( \frac{1}{S^*} \frac{\partial S^*}{\partial x_{ui}} - \frac{1}{T^*} \frac{\partial T^*}{\partial x_{ui}} \right) \\ \frac{\partial S^*}{\partial x_{ui}} &= 2 \sum_{r < s} (d_{rs} - \hat{d}_{rs}) \frac{\partial d_{rs}}{\partial x_{ui}} \\ \frac{\partial T^*}{\partial x_{ui}} &= 2 \sum_{r < s} d_{rs} \frac{\partial d_{rs}}{\partial x_{ui}}\end{aligned}$$

Với khoảng cách Minkowski:

$$\frac{\partial d_{rs}}{\partial x_{ui}} = d_{rs}^{1-\lambda} (x_{ri} - x_{si})^{\lambda-1} (\beta^{ru} - \beta^{su}) \operatorname{sig}(x_{ri} - x_{si})$$

Với:

$$\beta^{rs} = \begin{cases} 0 & (r \neq s) \\ 1 & (r = s) \end{cases} \quad (36a)$$

(36b)

Tóm lại:

$$\frac{\partial S}{\partial x_{ui}} = S \sum_{r < s} (\delta^{ru} - \delta^{su}) \left[ \frac{d_{rs} - \hat{d}_{rs}}{S^*} - \frac{d_{rs}}{T^*} \right] \frac{|x_{ri} - x_{si}|^{\lambda-1}}{d_{rs}^{\lambda-1}} \operatorname{sig}(x_{ri} - x_{si}) \quad (37)$$

### Kỹ thuật lắp của Kruskal

Ta gói gọn kỹ thuật qua các bước của thuật toán sau:

---

#### Algorithm 7 Kỹ thuật lắp của Kruskal

---

- 1: Chọn cấu hình ban đầu.
  - 2: Chuẩn hóa cấu hình sao cho trọng tâm tại điểm gốc và bình phương khoảng cách trung bình tới điểm gốc là 1. Ta làm vậy do giá trị  $S$  không đổi bởi phép tịnh tiến và co dãn. Việc lắp có thể dẫn đến cấu hình bị nở ra quá lớn.
  - 3: Tìm tập khoảng cách ( $d_{rs}$ ).
  - 4: Tìm tập độ lệch  $\{\hat{d}_{rs}\}$ .
  - 5: Tìm gradient  $g$  tại điểm hiện tại. Dừng lắp nếu  $g$  đủ nhỏ hoặc bằng 0
  - 6: Tính hệ số nhảy  $\alpha$ .
  - 7: Tính cấu hình mới theo công thức (35).
  - 8: Quay lại bước 2.
- 

### 6.3 Ví dụ

Ta có dữ liệu về khoảng cách theo đường chim bay giữa các địa điểm sau:

	Hà Nội	Hà Đông	Hòa Bình	Mai Châu	Mộc Châu	Sơn La	Tuần Giáo	Điện Biên Phủ	Mường Lay	Lai Châu	Sa Pa	Lào Cai	Yên Bái	Vĩnh Yên	Việt Trì
Hà Nội	0														
Hà Đông	12	0													
Hòa Bình	82	64	0												
Mai Châu	147	129	65	0											
Mộc Châu	206	189	124	60	0										
Sơn La	328	311	246	191	122	0									
Tuần Giáo	406	389	324	269	200	78	0								
Điện Biên Phủ	478	461	396	341	272	150	72	0							
Mường Lay	492	475	410	355	286	164	86	93	0						
Lai Châu	406	413	356	398	329	207	182	189	96	0					
Sa Pa	361	360	361	403	334	212	188	256	163	67	0				
Lào Cai	329	328	393	435	366	244	220	288	195	99	32	0			
Yên Bái	171	170	155	220	181	215	283	365	371	275	208	176	0		
Vĩnh Yên	55	54	111	176	224	258	336	408	422	373	306	274	116	0	
Việt Trì	80	79	86	151	199	233	311	383	397	348	281	249	91	25	0

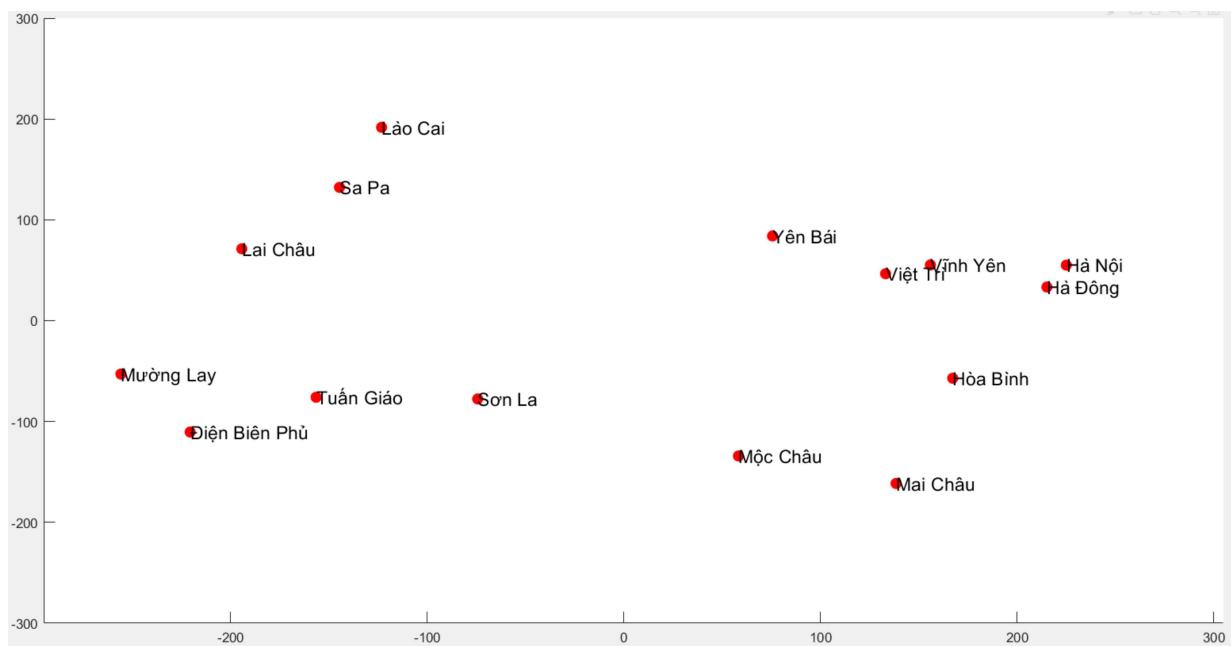
Hình 27: Khoảng cách giữa các tỉnh thành phố

Các giá trị riêng của ma trận B:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4.1313e+05	1.4984e+05	-5.3889e+04	2.1210e+04	-1.8679e+04	9.5233e+03	7.8541e+03	-6.0389e+03	-5.5666e+03	4.4243e+03	-2.2704e+03	-264.6756	-5.1401e-12	172.1581	315.9152

Hình 28: Giá trị riêng của ma trận B

Ta chọn ra hai giá trị riêng lớn nhất để xây dựng cấu hình:



Hình 29: Vị trí tương đối các tỉnh thành phố với cách làm cổ điển

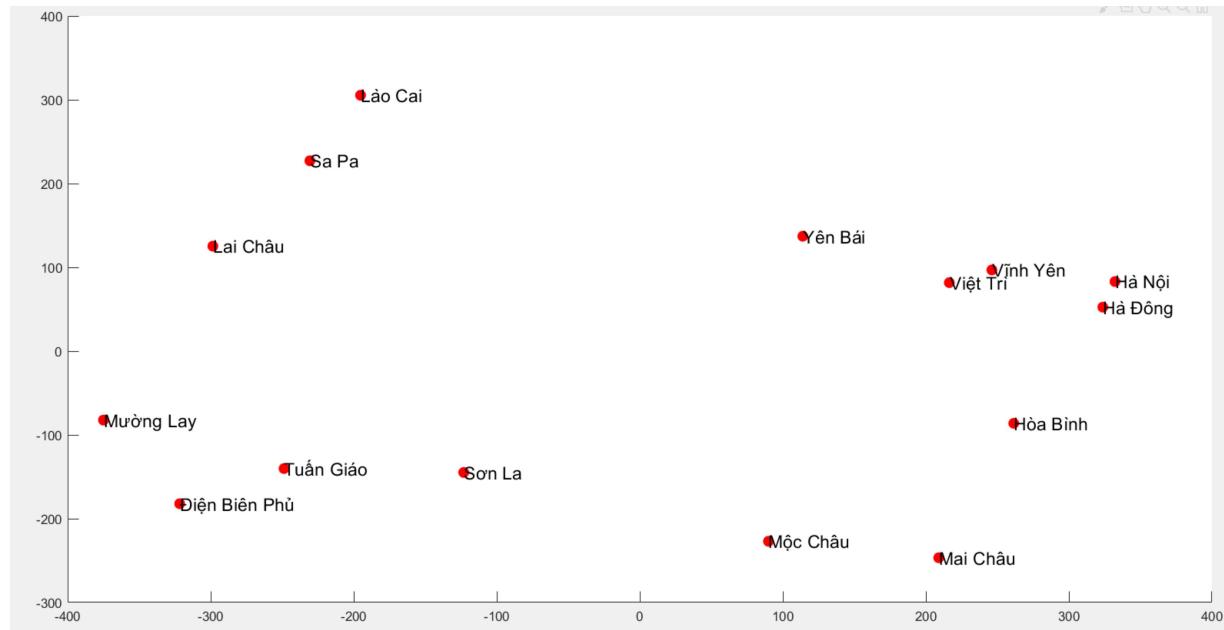
Để ý thấy ma trận B không phải nửa xác định dương, ta thêm hằng số  $c^* = 287.2911$ . Các giá trị riêng mới như sau:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
9.5724e+05	4.0540e+05	1.2244e+05	1.0899e+05	8.0592e+04	2.6456e+04	7.1001e+04	6.2016e+04	5.7752e+04	4.2351e+04	4.5564e+04	4.7445e+04	4.7027e+04	4.1003e-10	1.6457e-11

Hình 30: Giá trị riêng mới của B sau khi thêm hằng số

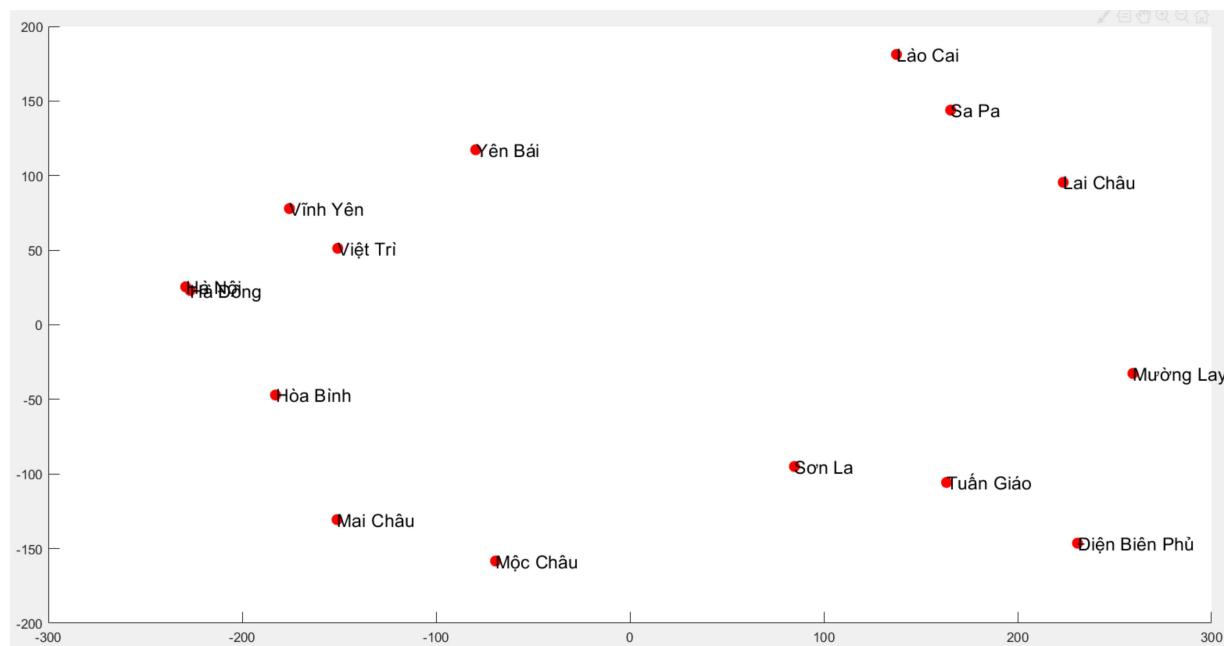
Lại chọn ra hai giá trị riêng lớn nhất và các vector riêng tương ứng, ta được cấu hình:

## 6 Chia tì lệ nhiều chiều



Hình 31: Cấu hình mới với B nửa xác định dương

Nếu dùng thuật toán của Kruskal để tìm cấu hình trong không gian hai chiều, ta thu được cấu hình như sau:



Hình 32: Cấu hình với thuật toán Kruskal

Cả ba cấu hình trên nhìn chung đều tốt với *Stress* của mỗi cấu hình đều nhỏ hơn 5%. Các cấu hình có thể sai lệch với thực tế, tuy nhiên vẫn cho ta thấy được các địa điểm gần nhau như thế nào, cũng như vị trí tương đối giữa các địa điểm.

## 6.4 Tổng kết

Như vậy qua hai phần kiến thức vừa nêu, tác giả đã trình bày hai phương pháp đơn giản nhất thuộc hai lớp phương pháp lớn. Nhìn chung mỗi phương pháp đều có ưu và nhược điểm riêng. Với cách làm cổ điển, cấu hình tìm được khá dễ dàng thông qua các phép nhân ma trận và tìm giá trị cũng như vector riêng. Tuy nhiên chỉ thu được một cấu hình và có thể cấu hình đó không tốt. Với cách làm của Kruskal, việc lập trình thuật toán cũng như tìm kiếm cấu hình sẽ phức tạp hơn do các công thức phải tính toán nhiều, khi có nhiều quan sát thì quá trình tìm cực tiểu mất nhiều thời gian. Tuy nhiên ta lại thu được nhiều cấu hình hơn nên khả năng tìm được cấu hình ưng ý cũng cao hơn. Cuối cùng, các kiến thức mà tác giả giới thiệu đều dừng lại ở mức cơ bản và lấy từ các bài báo đã lâu, độc giả có thể đọc các bài báo mà tác giả đã nêu trên để hiểu rõ hơn cũng như các bài báo mới hơn để có thêm những cải tiến của các phương pháp. Ngoài ra, chia tỉ lệ nhiều chiều còn rất nhiều lớp phương pháp khác, bạn đọc có thể tìm hiểu tại [9].

## Tài liệu

- [1] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*. Pearson Education, 2007, vol. 6.
- [2] D.Wishart, “An algorithm for hierarchical classifications,” *Biometrics*, vol. 25, no. 1, 1969.
- [3] G.N.Lance and W.T.Williams, “A general theory of classificatory sorting strategies: 2. clustering systems,” *The Computer Journal*, 1967.
- [4] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: an efficient data clustering method for very large databases,” in *SIGMOD '96*, 1996.
- [5] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” *Society for Industrial and Applied Mathematics, United States*, 2007.
- [6] F. Cailliez, “The analytical solution of the additive constant problem,” *Psychometrika*, vol. 48, no. 2, 1983.
- [7] J. B. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis,” *Psychometrika*, vol. 29, no. 1, 1964.
- [8] ———, “Non-metric multidimensional scaling: A numerical method.” *Psychometrika*, vol. 29, no. 1, 1964.
- [9] T. F.Cox and M. A.A.Cox, *Multidimensional Scaling*. Chapman & Hall/CRC, 2001.