Algorithms for Grand Tour and Parallel Coordinates

Huh, Moon Yul
Department of Statistics
Sungkyunkwan University
Seoul, Korea

We will consider two visualization tools for data mining in this talk: Grand Tour and Parallel Coordinates. For grand tour, many statistical packages (XLISP-STAT, SPlus, Statistica, and etc.) and visual data mining tools like XGobi (http://www.research.att.com/areas/stat/xgobi) and CViz (http://www.alphaworks.ibm.com/tech/cviz) have installed the feature after the inception of this idea by Asimov (1985). In this talk, I will give the algorithms for geodesic grand tour as implemented in XLISP-STAT and a modification of the grand tour known as Tracking Grand Tour suggested by myself (2001). Parallel coordinates has attracted many visual data miners since this is another form of multivariate scatterplots but in simpler form. However, the viewer of the plot may have wrong interpretation of the data depending on the arrangement of the variables. Usual approach to set the order of the variable is the arrangement of the variables as given in the data base. In this talk, I give two arrangement algorithms: permutation method and component method. The visual effects of these algorithms will be discussed.

key words: Grand Tour, Tracking Grand Tour, parallel coordinates, permutation method, component method.

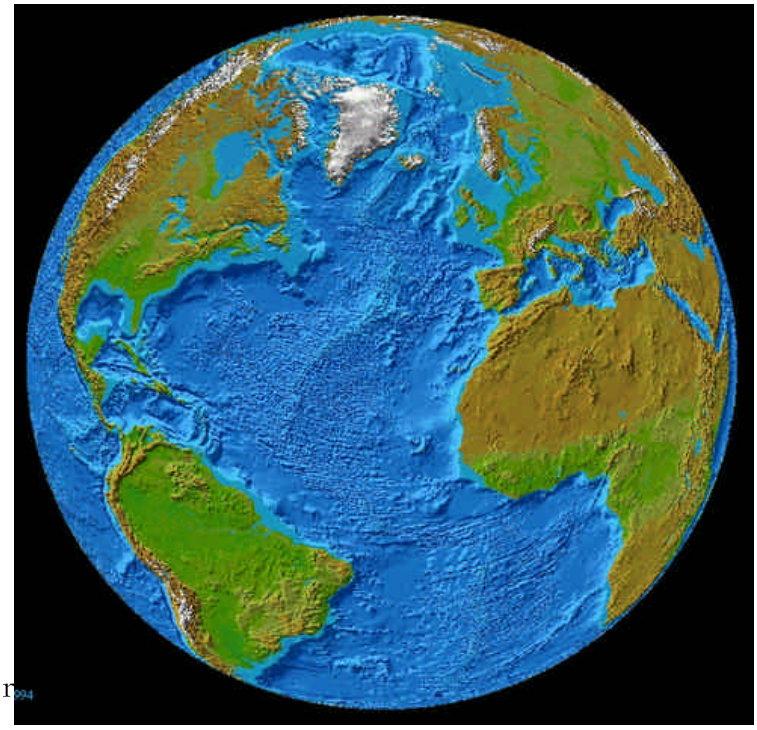


Figure 1 How can we tour all parts of the earth in the shortest total length?

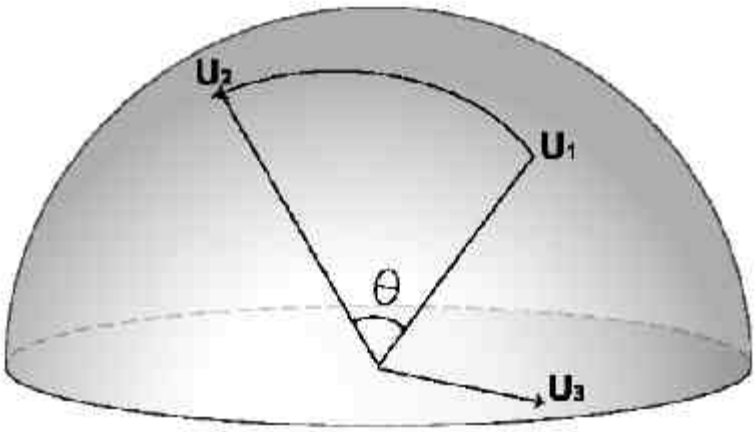


Figure 2 geodesic rotation

1. Algorithms of the Grand Tour

Steps of geodesic grand tour for p-dimensional data is as follows.

1) Choose 2 *p*-dimensional unit normal random vectors

2) construct an orthonormal basis $\mathbf{B}_{p\times p}$ that defines a rotation plane from the first vector to the second one while keeping the orthogonal complement of the plane fixed.

3) construct a rotation matrix $\mathbf{R}_{p\times p}$ on the rotation plane

4) apply $\mathbf{BRB'}$ to the data $\mathbf{X}_{n\times p}$ *a number of times*

5) draw the first 2 coordinates of the transformed data

6) repeat the above process indefinitely

Details of each step are as follows.

1) generate $\mathbf{w_1},\ \mathbf{w_2} \sim \mathbf{N(0,\ I_p)}$

2) construct orthonormal basis $\mathbf{B}_{p\times p}$ using the Gram-Schmidt process,

a) find the orthonormal basis $\{\mathbf{u_1},\ \mathbf{u_2}\}$ from $\{\mathbf{w}_1,\ \mathbf{w}_2\}$

$$\mathbf{u}_1 \Leftarrow \frac{\mathbf{w}_1}{||\mathbf{w}_1||}$$
$$\mathbf{w}_2 \Leftarrow \mathbf{w}_2-(\mathbf{w}_2,\mathbf{u}_1)\ \mathbf{u}_1$$
$$\mathbf{u}_2 \Leftarrow \frac{\mathbf{w}_2}{||\mathbf{w}_2||}$$

b) find the next *p-2* dimensional orthonormal basis

$$\{\mathbf{u_3},\dots,\mathbf{u_p}\} \perp \{\mathbf{u_1},\mathbf{u_2}\}$$
$$\mathbf{w} \Leftarrow \mathbf{w}_{k+1}-\sum_{i=1}^{k}(\mathbf{w}_{k+1},\mathbf{u}_i)\mathbf{u}_i$$
$$\mathbf{u}_{k+1} \Leftarrow \frac{\mathbf{w}}{||\mathbf{w}||}$$

for $k=2,\dots,p-1$, and $\mathbf{w_k}$ is p-dimensional vector with elements 0 except 1 on the k-th element.

c) $\mathbf{B}_{p\times p}= (\mathbf{u_1},\dots,\mathbf{u_p})$

3) construct $\mathbf{R}_{p\times p}$ for rotation angle θ as follows

$$\mathbf{R}_{p\times p}=\begin{pmatrix} \cos\theta & -\sin\theta & 0 \dots 0 \\ \sin\theta & \cos\theta & 0 \dots 0 \\ 0 & 0 & 0 \dots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \dots 0 \end{pmatrix}$$

4) transformed data: **XBRB′**

the number of rotations is chosen uniformly $[0, \frac{\pi}{2\theta}]$,

where θ is a small angle like 0.1

2. Tracking GT

With GT, the transformed data drawn on the screen is erased when the new transformed data is drawn. For TGT, do not erase the transformed data until a new 2 vectors for the new tour is selected.

3. Parallel Coordinates

3.1 Algorithms

A. permutation method

consider all pairwise permutation of the axes so that every possible adjacency is present.

Example: 4 variables named as 1,2,3, and 4.
2 permutations will do: pm1 = {1,2,4,3}, pm2 = {2,3,1,4}.
=> variable 1: (1,2) in pm1, (1,3) and (1,4) in pm2
variable 2: (2,1) in pm1, (2,3) in pm2, and (2,4) in pm1
variable 3: (3,1) and (3,2) in pm2, (3,4) in pm1
variable 4: (4,1) in pm2, (4,2) and (4,3) in pm1.

Wegman (1990) gives a simple formulation for this arrangement.

Let $v_i^{(j)}$ be the j-th permutation for the i-th variable.

The formula for the j-th permutation for p number of variables is:

$$v_i^{(j+1)} = (v_i^{(j)} + 1) \mod p,$$

$$\text{where } j = 1, \ldots, [\frac{p+1}{2}], \quad v_i^{(1)} = v_i, \quad v_1 = 1.$$

$$v_{i+1} = [v_i + (-1)^{i+1} i] \mod p, \quad i = 1, 2, \ldots, p-1,$$

$$0 \mod p = p \mod p = p, \quad x \mod p = (p+x) \mod p, \text{ if } x < 0.$$

and $[\cdot]$ s the greatest integer function.

This gives [(p+1)/2] permutations.

Now the problem is to choose the best permutation among these.

Criterion:

$$\text{minize} \quad \sum_{i=1}^{p} \sum_{j=1}^{p} n_{ij} \cdot D_{ij} = 2 \sum_{i=1}^{p-1} \sum_{j>i}^{p} n_{ij} \cdot D_{ij}$$

where D_{ij} is the disimilarity between variable i and j, and

$$n_{ij} = \begin{cases} 1 & \text{variables i and j are neighbours} \\ 0 & \text{otherwise} \end{cases}$$

a disimilarity is the Euclidean distance

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (b_{ki} - b_{kj})^2}$$

where

$$b_{ki} = \frac{x_{kj} - Min_i(x_{ij})}{Max_i(x_{ij}) - Min_i(x_{ij})}, \quad j = 1, \ldots, n, \quad k = 1, \ldots, p$$

and x_{ij} is the i-th observation on the j-th variable

B. Component method (Huh)

basic idea: rearrange the variable in the order of the magnitude of principal component

Let $X_{n\times p}$ be the data set with n observations and p variables. Then our procedure to rearrange the variables goes as follows.

Step 1. Obtain the principal component **u** corresponding to the largest eigenvalue of the similarity matrix obtained from the data matrix X.

Step 2. Obtain the index k corresponding to the largest absolute value of the vector **u**. We arrange the variable corresponding to this index in the first position.

Step 3. Reduce the data matrix X by eliminating the k-th column.

Step 4. Apply the above 3 steps until we have no more variables to select.

3.2 properties of each method

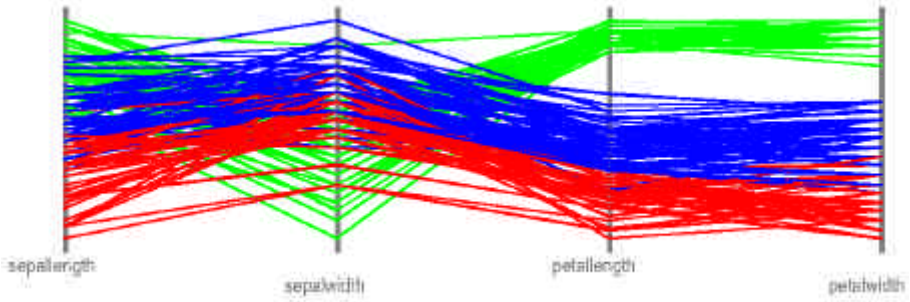
PM method: minimizing the total distance among all the possible combinations

=> find the permutation where similar variables are arranged closer each other

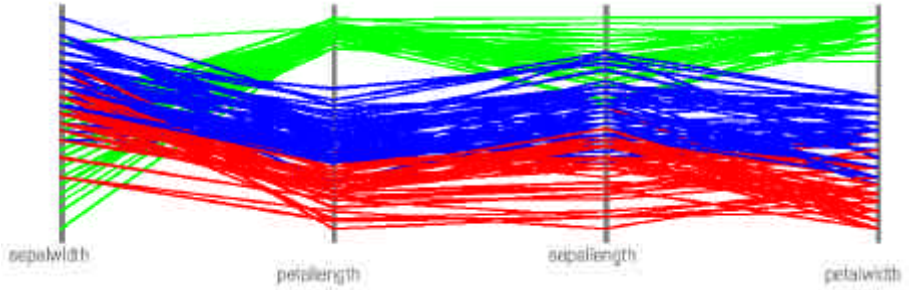
CM method: arranges the variables in the order of their capability of explanation of the data.

3.3 experimentation

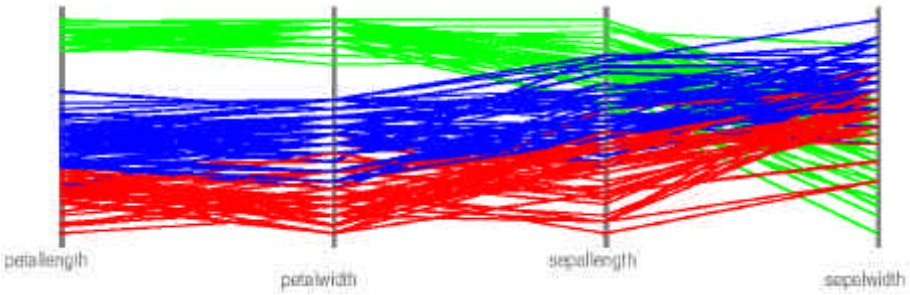
consider 2 data sets: iris data and imports car data (given in UCI data base). For each data, we applied 3-group k-means method, and then variables are arranged by PM and CM methods.



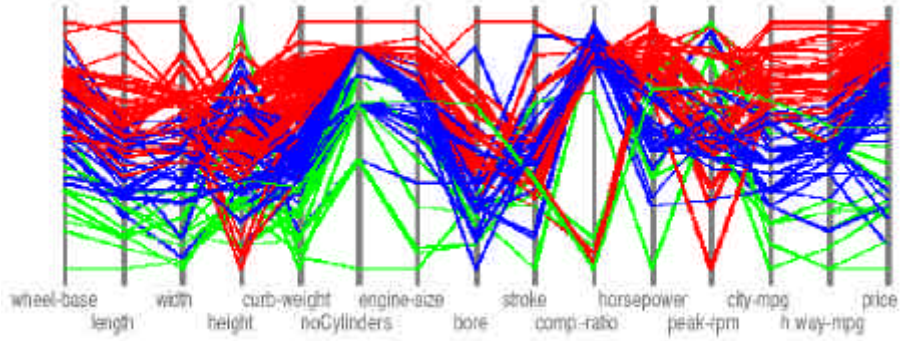
(a) iris data: variable arrangement as given in UCI data base



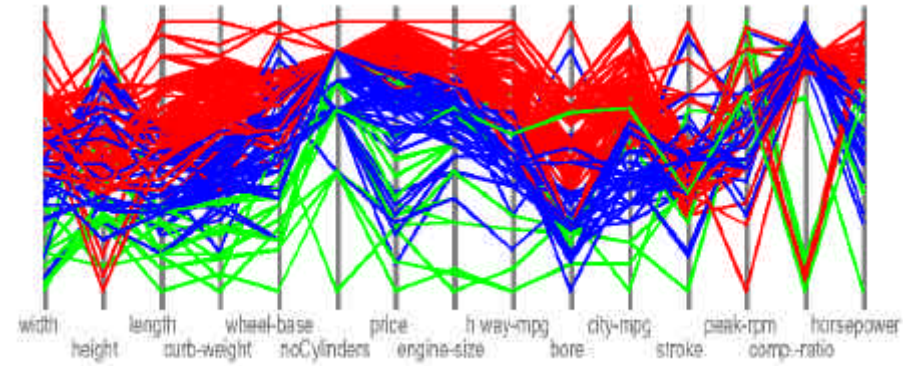
(b) iris data: arranged by PM



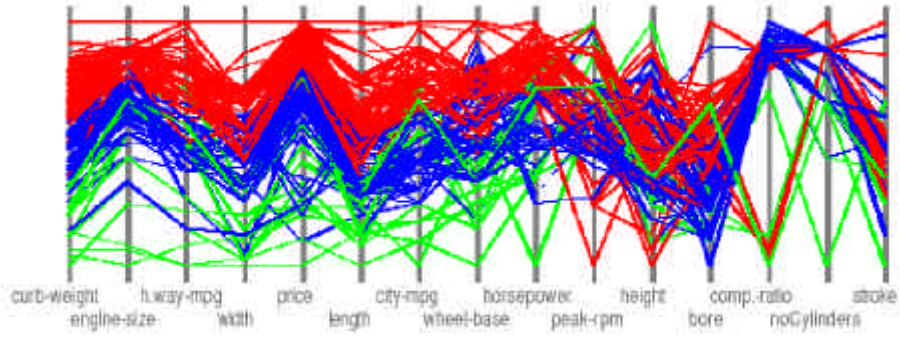
(c) iris data: arranged by CM



(a) car data: variable arrangement as given in UCI data base



(b) car data: variable arrangement by PM



(c) car data: variable arrangement by CM