Visualization of Multidimensional Data Using Modifications of the Grand Tour*

Moon Yul Huh and Kiyeol Kim†

Abstract

Current implementations of Asimov's Grand Tour (for example in XLISP-STAT by Tierney, 1990, or in XGobi by Buja et al., 1996) do not remember the path of projections and show only the current state during the touring process. We propose a modification of the Grand Tour named Tracking Grand Tour (TGT) that shows the trace of the touring process as small "comet trails" of the projected points. The usefulness of the TGT is demonstrated with a simulated and a real data set.

Key Words: Dynamic Graphics, Interactive Data Visualization.

1 INTRODUCTION AND METHOD

There exists a large literature on visualizing multivariate data. Scott (1992) gives a good survey of graphical display methods of multivariate data. The most frequently used method is the multivariate scatterdiagram (Tukey and Tukey, 1981). Several elegant ideas for representing high-dimensional data with the scatterdiagram and the combinations of scatterdiagrams with other graphical methods have been suggested (see, for example, Becker et al., 1996). Some of them are implemented in commercial statistical packages.

One breakthrough in graphical methods was the Grand Tour, suggested by Asimov (1985). Several studies have been done that extend the concept of Grand Tour and multivariate scattergrams, for example, "Projection Views" by Furnas and Buja (1994), and "Projection Pursuit Guided Tour" by Cook et al. (1995), and XGobi by Buja et al. (1996).

The Grand Tour basically examines a series of scatterplots obtained by a smoothly changing sequence of projection methods. With the Grand Tour one can dynamically explore the structure of multi-dimensional data projected onto a two dimensional window. The Grand Tour as implemented for example in XGobi, however, shows us only a series of projections of data points on a plane

*This research is partially supported by KOSEF 971-0105-028-1.

†Department of Statistics, Sung Kyun Kwan University, Seoul, Korea. The research was done while the first author was visiting CMIS, CSIRO, Canberra, Australia. The authors are grateful to Dr. Andreas Buja at AT&T Labarotories for his helpful comments during the preparation of this paper.

that interpolates a sequence of randomly generated planes. Hence with the Grand Tour one can observe only the projections of data points on the current interpolating plane.

The Tracking Grand Tour (TGT) mitigates these deficiencies of the generic Grand Tour by enabling viewers to watch traces of the changing data points in real time during the touring process. TGT is just keeping the tracking of the projections of data points on the interpolating planes. It starts a new tracking only when a new random plane is generated. We found this to be particularly useful in exploring the structure of data of small size. TGT will be effective both for small and large scale data sets. With the small scale data sets, additional ink used for the TGT will produce the effect of artificially increasing the size of data sets. With large scale data sets, data sets with similar properties will appear cluttered, and there will be seperate clutterings for each data set with different properties. This cluttering effect will provide us better understanding of the data structure.

A related idea to the TGT can be found in Buja et al. (1996, section 8), where velocity vectors attached to the points are proposed as indicators of the motion of each point.

2 EXAMPLES

To demonstrate the TGT, we use one simulated and one real data set.

The simulated data example consists of two concentric spheres of different radii (3 and 5) with 30 data points each. As is apparent in Figures 1a and 1b, a scatterdiagram and a grand tour of the data do not suggest any specific pattern in the data. That is, we cannot tell from the plots that there are two concentric spheres with different radii. Figure 1c is a snapshot of TGT of the simulated data. It clearly reflects the fact that the data consists of two concentric spheres.

The second and real data example is from the 1995 Korea census data. The original data consists of 12,958,181 ordinary households with 31 variables. The data was aggregated and averaged over 7,148 *dong*s (administrative units consisting of around 1,500 households) or *dong*-equivalent administrative units like *myun*. Out of these 7,148 *dong*s, we randomly selected 400. We also selected 7 variables for our study. These variables can be classified into 3 categorizes as follows. First category consists of one variable which is the average age of household members (variable name: Age). Second category is related to household head and consists of two variables : percentage of education at a junior college and above (collEdu.); percentage of service-related occupations such as banking, insurance, real estate and business service (servJob). Third category is related to living quarters and consists of 4 variables: percentage of households living in apartments (apartment); percentage of households living in single-family owner houses (myHouse); percentage of households living in government supported small-sized apartments less than approximately 46 square meters(14 *pyong*)(smallHouse); percentage of households living in rental houses of size larger than approximately 230 square meters(69 *pyong*) which several

families are sharing(tenement).

Figure 2a is a scatterdiagram of the census data. It is not easy to see the structure of the data from the figure. It only informs us of the relationships between the two pairwise variables. The Grand Tour (Figure 2b) suggests the possible existence of grouping in the data. Figure 2c gives two snapshots from the TGT. The snapshots strongly suggest the existence of data grouping, and also suggest there are a few points behaving differently from the majority of the data points.

It is not surprising that the TGT will be a successful tool for finding clusters and outliers. We now wish to go one step further and use the TGT to aid in visually deciding on a meaningful number of groups k produced with k-means clustering method. To this end we step through $k = 2, 3, 4$ and color the groups at each step. The goal is to use the information from the TGT to decide whether groups are natural in the sense that they are separated in data space, or whether they are just arbitrarily obtained slices through an ill-structured data cloud.

Figure 3a and 3b show, respectively, a scatterdiagram and a TGT snapshot of the data of the 2-group-clustering. The TGT representation shows that most of the data points with different colors belong to different natural groups, except for a few possible outliers. The TGT snapshot in Figure 3b also indicates further grouping is possible among the blue-colored data points. Hence we continue by partitioning the data into three groups.

The scatterdiagram suggests that the charateristics of the first group (red-colored points) are higher level of education, higher rate of service-related jobs, and higher rate of apartment living. We now partition the data points into three groups shown in Figures 4a and 4b. Our interpretation of the TGT is the same as in the case of the 2-group-clustering. That is, the data points with three different colors do not overlap during the process of TGTing, except for a few possible outliers.

Based on the interpretation of the scatterplot matrix, we may characterize the three groups as follows: first group (red-colored group), highly educated younger generation with higher rate of apartment housing; second group (blue-colored group), older and less educated generation with higher rate of living in detached own house; third group (green-colored group), mid-class group.

Stepping up further, we have the results of the 4-group clustering in Figure 5a and 5b. The TGT suggests most of the data points with different colors appear in different natural groups. We conclude that the data likely fall into at least four natural groups. Careful inspection of the scatterdiagram of Figures 5 suggest that most of the data points belonging to the green- and cyan-colored groups in the 4-group clustering come from the green-colored group ("mid-class group") in the 3-group clustering. Our interpretation from the scatterdiagram for these two new groups is as follows: green-colored group, mid-class group with moderate rate of service related jobs and higher rate of tenement housing; cyan-colored group, mid-class group with lower rate of service related jobs and moderate rate of small-sized-my-own-detached-housing.

Most of the data points belonging to the red ("highly educated younger generation living in apartments") and blue ("older and less educated generation

living in detached owner-houses") groups in the 3-group clustering carry over into corresponding groups of the 4-group clustering.

3 conclusion

In this paper, we proposed a modification of the Grand Tour called the Tracking Grand Tour. We found by experimentation that the modification can be an efficient visualization tool for understanding the structure of multi-dimensional data. It was found by experimentation that the tool is especially successful for finding and assessing clusters and outliers. The programs for the tools are available through a software DAVIS (Data VISualization system) at

`http://stat.skku.ac.kr/~myhuh/software/DAVIS/DAVIS.html`

References

[1] Asimov, D. (1985), The Grand Tour: "A Tool for Viewing Multidimensional Data", SIAM Journal on Scientific and Statistical Computing, 6, 128-143

[2] Becker, R.A., Cleveland, W.S. and Shyu, M.(1996), "The Visual Design and Control of Trellis Display", Journal of Computational and Graphical Statistics, 5, 123-155

[3] Buja, A., Cook, D. and Swayne, D.F. (1996)m "Interactive High-Dimensional Data Visualization", Journal of Computational and Graphical Statistics, 5, 78-99

[4] Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995), "Grand Tour and Projection Pursuit", Journal of Computational and Graphical Statistics, 4, 155-172

[5] Furnas, G.W. and Buja, A. (1994), "Projection Views: Dimensional Inference Through Sections and Projections", Journal of Computational and Graphical Statistics, 3, 323-353

[6] Scott, David W. (1992), Multivariate Density Estimation, John Wiley & Sons, INC

[7] Tierney, Luke (1990), LISP-STAT, John Wiley & Sons, INC

[8] Tukey, PauL A. and Tukey, John W. (1981), "Data Driven View Selection; Agglomeration and sharpening", Vic Barnett, ed., Interpreting Multivariate Data, Chichester, England, 1981, 231-232

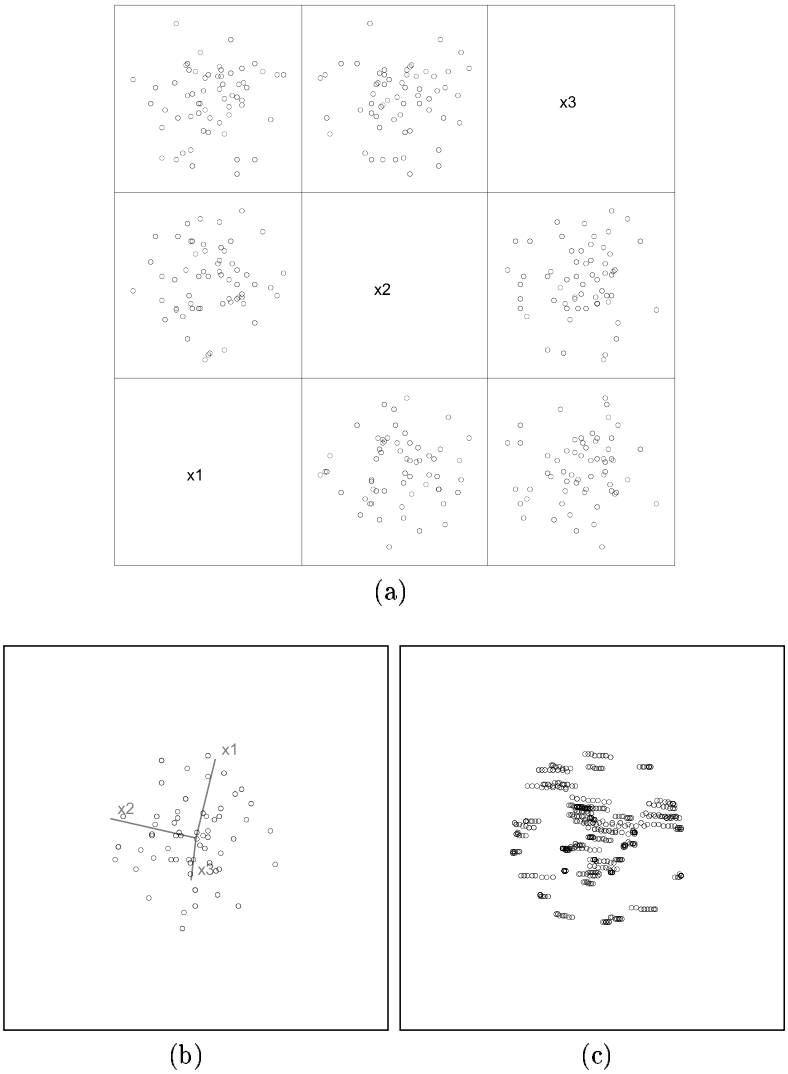
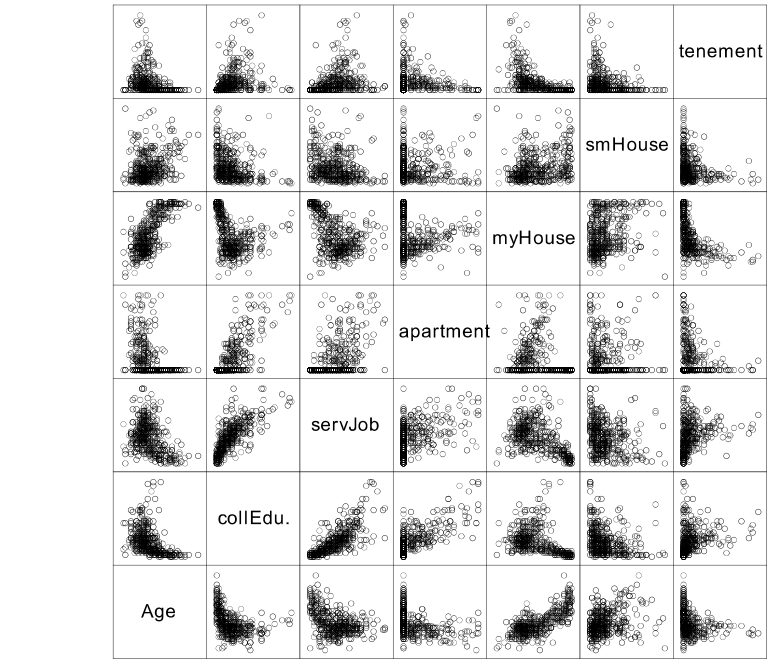
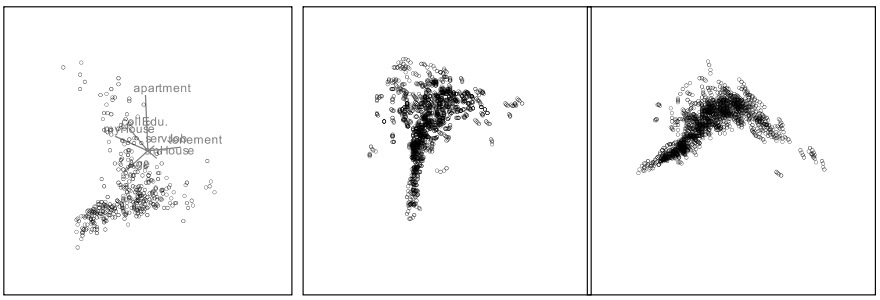


Figure 1: Scatterdiagram and the snapshots of Grand Tour and TGT of the simulated data

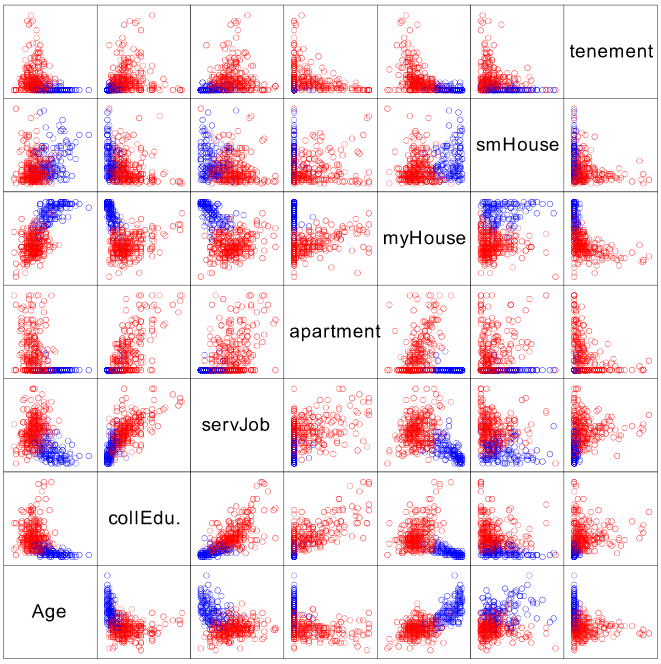


(a)

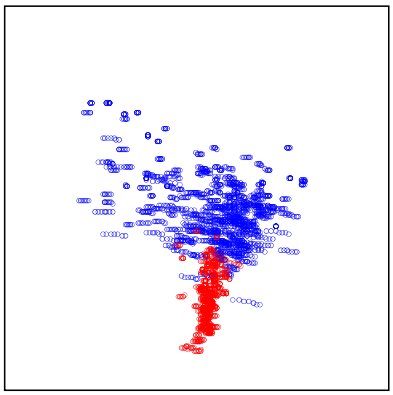


(b) (c)

Figure 2: Scatterdiagram and the snapshots of Grand Tour and TGT of the census data

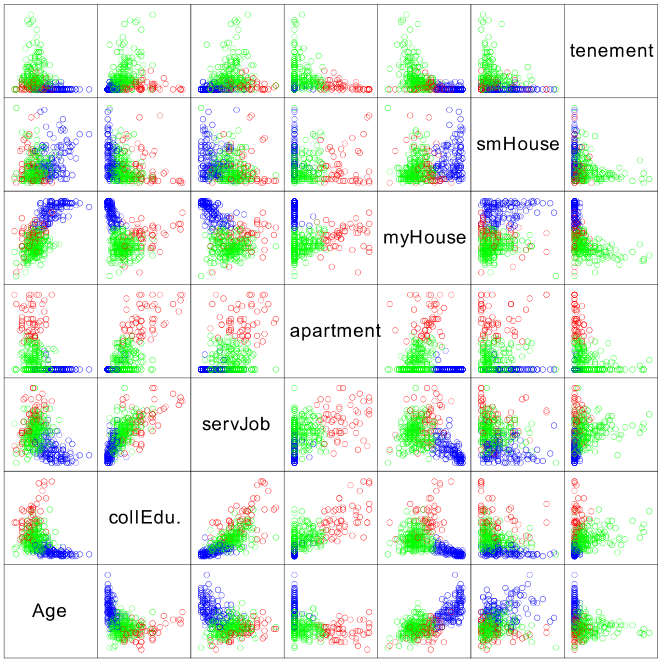


(a)

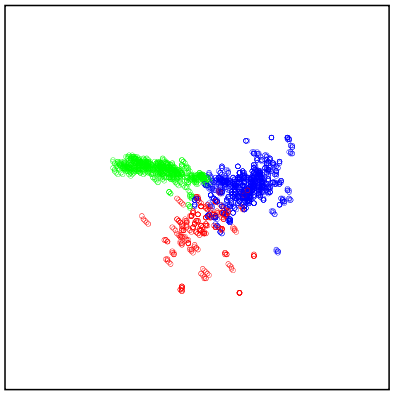


(b)

Figure 3: Scatterdiagram and the TGT snapshot of the census data after dividing the data into two groups with k-means method

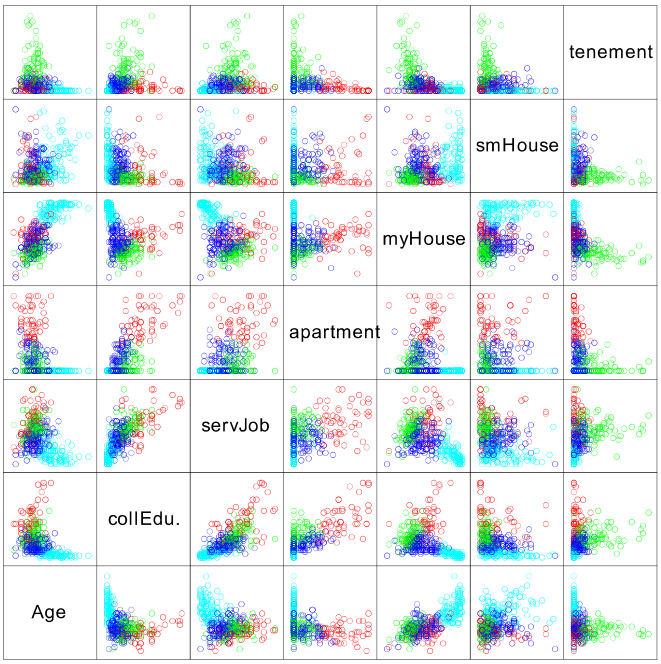


(a)

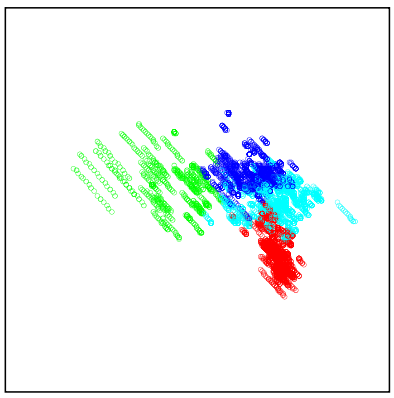


(b)

Figure 4: Scatterdiagram and the TGT snapshot of the census data after dividing the data into three groups with k-means method



(a)



(b)

Figure 5: Scatterdiagram and the TGT snapshot of the census data after dividing the data into four groups with k-means method