VISUAL EXPLORATORY DATA ANALYSIS:
PROC VISUALS, A USER-WRITTEN EXPERIMENTAL SAS PROCEDURE

Forrest W. Young, Douglas P. Kent
Thomas C. Edds, and Warren F. Kuhfeld

Psychometric Laboratory
University of North Carolina at Chapel Hill

1. Visual Exploratory Data Analysis

There have been many new developments in exploratory data analysis (EDA) in the past decade, including new EDA methods for exploring and visualizing multivariate data. During these same years there have also been tremendous improvements in graphics hardware, accompanied by significant decreases in costs.

Taken together, these developments permit a new approach to exploratory data analysis, an approach that has been recently discussed by Friedman (1985), Huber (1985), Donoho, Donoho & Gasko (1986), Tukey & Tukey (1985), Suja & Asimov (1985), ourselves (Young, et al., 1985, 1986) and others. We call this approach Visual Exploratory Data Analysis (VEDA).

The goal of VEDA is to graphically represent the structure of multivariate data so that visual exploration can provide you with fresh insights about your data's structure. To do this, the graphical representation must:

> respect your data's multidimensional geometry;

> respect your own three-dimensional perception.

Unfortunately, you must move outside of the environment provided by standard statistical packages such as SAS to take advantage of these recent developments, as, by and large, these packages have not incorporated multivariate exploratory data analysis methods, nor taken full advantage of the most recent hardware.

With this in mind, SAS Institute is supporting a project at the Psychometric Laboratory of the University of North Carolina (UNC) to develop new multivariate EDA SAS procedures. Three non-graphical multivariate EDA procedures (PRINQUAL, CONJOINT, and CORRESP) are discussed by Kuhfeld, et al. (1986). This paper presents an overview of our VEDA procedure, PROC VISUALS.

2. Overview of PROC VISUALS

PROC VISUALS was inspired by the maxim that "If a picture is worth a thousand numbers, then a moving picture must be worth at least a zillion!" The procedure is designed to help you visually discover the structure that may be hidden in the many variables of your multivariate data.

PROC VISUALS presents you with a high resolution, color picture of a three-dimensional (3D) space. Contained in the space is a cloud of objects which represent the observations in your data. The dimensions of the space are three of the variables in your data. To enhance the 3D effect, the picture is presented in perspective projection.

You are the director of the moving picture of your data. Using the cursor keys, you animate the picture to move (translate) and spin (rotate) the cloud of objects. The movements take place in real time, giving you an animated picture of your data structure. The combination of perspective projection and interactive animation helps reveal the 3D structure of your data.

Simultaneously, you can have a second 3D moving picture that is based on three additional variables, giving you a different view of your data. Its motion is also under your direction. You can interpolate between the dimensions of the two 3D pictures to get new views of your data's structure.

In addition to directing the motion picture of your data, a number of other features are under your direction. You can display labels for all observations, or for selected subsets of observations. The labels follow the objects as you animate the object cloud. You can change the color and icon (object representation) for any subset of observations. You can display grids representing the three planes formed by the 3 axes of the window space, and you can drop perpendicular lines from the objects to the planes.

In addition, you can record everything that is displayed on the screen, and playback whatever portions of the recording that you wish. The recording, along with your data, can be sent to other researchers so that they, too, can see the structure of your multivariate data.

Note that PROC VISUALS is not designed to be device independent as are SAS/GRAPH procedures. Rather, it takes full advantage of the IBM Professional Graphics Adaptor. This device has a graphics processor, 320K memory, and firmware which performs all 3D operations, (perspective projection, clipping, etc.). Thus, all graphics operations are offloaded from the host, an IBM PC/AT. Kent, et al. (1986) discusses these aspects in more detail.

840
3. Using PROC VISUALS

Since PROC VISUALS is highly interactive, it is used differently than other SAS procedures. Most of the behavior of the procedure is controlled by "commands". Each command is a single keystroke that has immediate and highly interactive effects. Ordinary SAS statements determine the VISUALS environment in which these commands operate. In section 3.1 we describe these SAS statements. In section 3.2 we explain the 3D motion picture commands; in section 3.3 we review the interpolation commands; and in section 3.4 we summarize the recorder commands.

3.1. SAS Statements use with PROC VISUALS

PROC VISUALS is invoked within the SAS system by the following statements:

```
PROC VISUALS
    DATA=dsn SNAP=dsn REC=dsn;
    SCATTER x1-var * y1-var = z1-var;
    SCATTER2 x2-var * y2-var = z2-var;
    IF expression THEN statement;
    REVEAL attributes;
    VAR varlist;
    ID varlist;
    RUN;
    QUIT;
```

PROC VISUALS is an interactive procedure, requiring the use of the RUN statement. The SCATTER statement is required before the first RUN statement. Other statements are optional, and may be used in any number and combination.

When you submit a RUN statement the previous SAS statements are used to create the VISUALS environment. You then leave the SAS display manager environment and enter the PROC VISUALS environment, and example of which is shown in Figure 1. (You can return to the SAS display manager environment by using the ESCape key).

After executing the first RUN statement, PROC VISUALS displays an initial 3D space whose dimensions are the three SCATTER variables, just as PROC G3D uses its SCATTER statement to construct a 3D space. The left-hand large square in Figure 1, which occupies 3/4ths of the screen, contains the 3D space. What you see initially is a horizontal "x" dimension and a vertical "y" dimension. The "z" dimension is aimed straight at you. The commands discussed in section 3.2 allow you to move and spin these dimensions around. The 3D space contains "objects" which represent your observations. In Figure 1 there are 12 small "3D cross" objects, and 7 long "vector" objects. The objects and their colors are determined by the _COLOR_ and _SHAPE_ variables in the DATA= dataset. The right-hand 1/4 of the screen has several information displays that are explained below.

While in the PROC VISUALS environment, the "commands" discussed in section 3.2, 3.3, and 3.4 are activated. When you ESCape back to the SAS environment the commands are deactivated, and you may enter additional SAS statements. A new RUN statement brings back the previous view of the 3D space. A QUIT statement terminates PROC VISUALS.

The ID statement lets you display object labels, and the REVEAL statement lets you display optional attributes of the objects or space. The attributes are COORDINATES (display coordinates of objects) BLINK (blink objects), PROJECTIONS (project objects onto selected planes), and PLANE (display selected plane). Both statements may be used with the IF/THEN statement to label or display attributes of selected objects. Figure 2 shows the results of

```
IF SHAPE='VECTOR' THEN ID DRUG;
```

and Figure 3 the results of

```
REVEAL PLANE='XY';
```

The IF/THEN statement is used to perform operations on subsets of observations. The logical expression can involve any variables in the DATA= dataset. The statement after the THEN clause can be an ID or REVEAL statement, or an assignment statement using DATA= variables.
3.2. 3D Motion Picture Commands

The 3D motion picture commands allow you to direct the motion of the cloud of objects. All motion picture commands are invoked by using the numeric keypad. The commands, and their assignment to the IBM/PC AT keypad, are shown in Figure 4 (a help screen from PROC VISUALS).

The motion that you direct may be "stop motion" (one movement per key stroke) or "animated" (repeated movements per key stroke). The animation feature is controlled by the SCROLL LOCK key, which acts as an ANIMATION LOCK key. That is, when SCROLL LOCK is off the motions are not animated, and when SCROLL LOCK is on the motions are animated.

3.3. Interpolation Commands

The interpolation commands are all located on the function keys, as shown in Figure 7. When you invoke PROC VISUALS with only the SCATTER statement (and not the SCATTER2 or VAR statement) you have defined only three dimensions (the x1-var, y1-var and z1-var variables). Since there are no "extra" dimensions for...
The function keys activate the interpolation features.
Caps Lock switches animation (repeated interpolation) on or off.

- **x 1** and **x 2**: Keys F1 and F2 interpolate along the "x" dimension.
- **y 1** and **y 2**: Keys F3 and F4 interpolate along the "y" dimension.
- **z 1** and **z 2**: Keys F5 and F6 interpolate along the "z" dimension.
- **less** and **more**: Keys F7 and F8 change the amount of interpolation per keypress: F7 decreases it, F8 increases it.
- **next** and **residual space**: Key F9 toggles between SCATTER, SCATTER2, & HYPERSPACE. Key F10 sets SCATTER to the interpolated HYPERSPACE, and SCATTER2 to the largest residual space.

**Figure 7**
interpolation the commands discussed in this section cannot be used.

When you use both the SCATTER and SCATTER2 statements, you have defined two spaces: The scatter space and the scatter2 space. The scatter space is displayed to you first. You can use the next space (F9) command to display the scatter2 space.

When you have used both scatter statements, you can use the interpolate commands (F1-F6) to interpolate between the two spaces, and the increment commands (F7, F8) to control the interpolation increment. When you have used the interpolate commands, you have created a third space called the interpolation space. This is the space you see on your screen.

Keys F1 and F2 cause interpolation on the "x" dimension (between x1-var and x2-var, unless the residuals command has been used), F3 and F4 interpolate on the "y" dimension, and F5 and F6 interpolate on the "z" dimension. Of these, the F1, F3, and F5 keys interpolate towards the stated dimension of the scatter space, whereas the F2, F4, and F6 keys interpolate towards the stated dimension of the scatter2 space.

The next space command (F9) toggles between three spaces: The scatter space, the current interpolation space, and the scatter2 space. If the interpolation commands are used when you are looking at a scatter or scatter2 space, the previous interpolation space is lost.

When you use the VAR statement, you define an additional space whose dimensionality equals the number of VAR variables. This space is called the "residual space", and is not displayable. However, this space may be used to modify the scatter and scatter2 spaces with the residuals command (F10). This command replaces the scatter space with the current interpolation space, and replaces the scatter2 space with the three largest (maximum variance) dimensions of the data space that are orthogonal to the interpolation space. The data space consists of all variables that you have specified on the SCATTER, SCATTER2 and VAR statements.

The interpolation commands may be used in either a "stop motion" mode (one movement per key stroke) or an "animated" mode (repeated movements per key stroke). The animation feature is controlled with the CAPS LOCK key. When CAPS LOCK is off the interpolations are not animated, and when CAPS LOCK is on they are animated.

3.4 Recoder Commands

Everything that you see in the PROC VISUALS environment may be recorded and played back using the record commands. Recorded material may be overwritten with new material, if desired. The commands are located on the ALT/function keys (i.e., when ALT is held down and a function key is pressed).

The recording commands are ALT/F1 (record) and ALT/F2 (single record). ALT/F1 is a toggle which turns the recorder on or off. When recording is on, all frames being displayed are recorded. ALT/F2 records only a single frame, the one currently being displayed. The recording is made starting at the current record head position, which is determined by the set record command. This allows erasing recorded material.

The playback commands are ALT/F3 (play), ALT/F4 (single play), and ALT/F6 (show frame). ALT/F3 is a toggle which turns the player on or off. ALT/F4 plays back a single frame, and advances the player to the next frame. ALT/F6 plays back a single frame without advancing to the next frame. Playback starts at the current playback head position, which is determined by the set play command. Note that the recorder and player may be on simultaneously.

The head positioning commands are ALT/F7 (mark frame), ALT/F8 (list marks), ALT/F9 (set play), and ALT/F10 (set record). ALT/F7, the mark frame command, marks the current frame for future reference as a frame to which the playback or record heads may be set. ALT/F8, the list marks command, lists the marked frames. ALT/F9, the set play command, positions the playback head to a specified marked frame. The next playback will start with this frame. ALT/F10, the set record command, positions the record head to a specified marked frame. The next recording will start with this frame and erase this and all of the following previously recorded and marked frames.

4. Example

The figures in this paper involve an actual set of data obtained from Fisher (1985). They are means of judgements made by 32 male and 21 female judges. About half the judges are regular users of Psychedelic Mushrooms, the other half are nonusers. The judgments are of the frequency that the judge, the judge's mother, and the judge's father use Alcohol, Tobacco, Marijuana, Cocaine, Amphetamines, Hallucinogens, and Opium. There are 12 observations (2 males x 2 user types x 3 judged people and 7 variables).

This mean frequency matrix was analyzed by PROC PRINQUAL (Kuhfeld, et al., 1986) to obtain its principal components (PC) under the assumption that the data are ordinal. The analysis, which monotonically transforms variables to maximize the variance accounted for by three PCs, accounted for 99.7% of the total variance.

The figures are 3D biplots in which the observation scores and variable coefficients on the three PCs are used to locate the objects. The observations are represented by 3D crosses, and the variables by vectors.

The most striking revelation provided by PROC VISUALS is seen in Figure 6b. Five drugs
(marijuana, tobacco, alcohol, cocaine and amphetamines) are essentially co-planar, whereas the other two (opiates and hallucinogens) are orthogonal to the plane formed by the first five. This structure corresponds nicely with the fairly widespread use of the five drugs and the infrequent use of the other two drugs.

Note that this structure is not immediately obvious from ordinary 2D biplots of pairs of PCs. However, it was easy to discover this structure with PROC VISUALS. We simply spun the 3D biplot, and watched it spin in space. This clearly revealed the easily interpreted 3D structure. We then used the non-animated move and spin commands to obtain a revealing perspective projection of the structure, the one shown in Figure 6b.

5. Conclusion

We are very pleased that PROC VISUALS seems to be a useful VEDA tool, even at its current early developmental stage.

In a few years, when hardware costs have been reduced even further, when hardware capabilities have been further enhanced, and when PROC VISUALS has matured, we should be able to smoothly animate lighted solid object models of multivariate data. This should greatly enhance the 3D effect over that obtained from the current wire-frame models that are animated at only 3-4 frames per second.

We are very optimistic that a mature PROC VISUALS will be a highly useful VEDA tool to help you explore and understand the structure of your multivariate data.

---

References


---

1 SAS is the registered trademark of SAS Institute, Inc., Cary, NC, USA.

2 PROC VISUALS is being developed at UNC, not at SAS Institute. There is no commitment by the University or by the Institute to support or distribute PROC VISUALS. Questions should be directed to the first author at the Psychometric Laboratory, Davie Hall 013-A, Chapel Hill, NC 27514 USA. Phone (919) 962-5058.

3 This paper reports on the potential design of PROC VISUALS. Portions of the description correspond to a working prototype, other portions correspond to design specifications not yet implemented. The current (Jan, 1986) prototype is a stand-alone program that does not require the SAS system, but which works with it. All commands are implemented in the current prototype, as is labeling. No promise is made that PROC VISUALS will be fully developed as a SAS procedure, and if it is, that it will be in accord with this description.