STARS AND FACES -- PROCEDURES FOR CONSTRUCTING GRAPHICAL PROFILES OF MULTIVARIATE DATA

Harold W. Ougel, General Motors Research Laboratories

ABSTRACT
Humans are exceptionally proficient at processing visual information. With the widespread availability of computer generated graphics, much interest in recent years has focused on graphical methods to supplement analytical techniques for studying complex multivariable data. This paper describes the implementation of a star for these methods as two new SAS procedures named STARS and FACES. These procedures create graphical profiles of multivariate observations in the form of star-like polygons and face-like images to collectively picture the information in the data. Color, combined with several analytical techniques, is added to current graphical methods to extend their utility at representing high-dimensional data. Ease-of-use in executing the procedures and flexibility is stressed to provide comprehensive and effective tools for exploring and summarizing multivariable data.

1. Introduction
Simple and effective means of examining multivariate data are of interest to anyone who needs to analyze or summarize sets of measurements which have been taken collectively on two or more attributes. Up until the mid 1970's nearly all multivariable informations were listed in tabular form. Since then, the proliferation of fast and inexpensive computer generated graphics has spawned numerous developments in graphical methods for representing multivariate data [12]. The effectiveness of graphics as a diagnostic tool as well as a tool for summarizing results is now well established. Because the human is so proficient at processing visual information, no other media can convey as much information so easily and effectively as good graphics.

Of the many multivariate methods developed for graphics, the most widely recognised is Chernoff Faces [1,12]. This technique maps an n-dimensional data point into a cartoon-like face. Up to 18 measurements comprising an observation can be assigned to control different characteristics of a face. This technique is of special interest since humans perceive the face as a gestalt and evidence suggests [13] that humans possess a special innate ability for visually processing faces.

Flury and Riedwyl [3] developed a new asymmetrical face technique for displaying multivariate data. The concept is similar Chernoff Faces but the resulting face diagrams are far more realistic, less likely to become degenerate, and the face parameters are nearly all independent of each other. A powerful added feature is the ability to independently control the left and right side of the face. This allows twice the number of variables, but more importantly, asymmetry of faces provides a graphical form of multivariate paired comparisons. A number of interesting applications are given in Flury and Riedwyl [3,4].

Friedman, et al. [5] describe a technique which represents an n-dimensional data point as an n-sided polygon resembling a star. This technique holds several major advantages over the Chernoff Faces; it is easier to apply and understand, data values can be read directly from it, and it is far less likely to introduce bias.

These techniques have been implemented as two new graphical profiles procedures under SAS named STARS and FACES. Extensive capabilities are provided through enhanced graphics, analytical for tests, and flexible program controls. Applications, primarily for data exploration, include:
- Data description and summarization
- Outlier detection (univariate & multivariate)
- Identification of trends in serial data
- Clustering of multi-dimensional data
- Comparisons with a standard or control group
- Assessing data compliance with specifications
- Multivariate paired comparisons
- Correlation and variance analysis

2. Discussion
Although both STARS and FACES accept the same n-dimensional data, their graphical representations are entirely different. With STARS, a star-like polygon is constructed for each observation by connecting with straight line segments the standardized variables plotted along positioned rays emanating from the center of a circle. With FACES, a human-like face is constructed for each observation using the standardized variables to control characteristics of the FACE.

Both procedures can accommodate up to 36 variables; a maximum of 36 rays spaced every 10 degrees for STARS and 18 individually addressable features on both the left and right side of a FACE. With STARS a variable is generally assigned to only one position while with FACES a variable usually controls one feature, but simultaneously on both the left and right side of the FACE. However, both procedures can be operated similarly. That is, with STARS, variables can be symmetrically assigned to left and right positions, while with FACES, left and right features can be individually controlled. The symmetry attribute in both procedures can be effectively exploited for multivariate paired comparisons.

Through device independent graphics the profiles can be drawn on virtually any device. Color can be employed to assist in differentiating between components of the profile or to highlight certain aspects of the data. For example, with color devices the grid and resulting data polygon of STARS can be colored differently. Similarly with FACES, the eyes can be given a distinct color for special effect. Data groupings can be differentiated with color and data outliers can be marked by highlighting the corresponding profile components with color. Profiles can be overlayed and distinguished with color.

In constructing profiles the variables assigned are standardized to the (0,1) unit interval. This is affected through a user selected transformation. The default transformation maps the 3-sigma limits about the mean onto the unit interval.

1 The ISSCO DISSPLA graphics system is used [8].

253
Extreme values can be "trimmed" to prevent the formation of degenerate profiles.

2.1 The STAR Profile
The STAR profile is constructed upon a grid consisting of two concentric circles placed upon a series of spokes emanating from a common center as illustrated by the sample in Fig. 1. Spokes are provided for as many positions as specified by the user, usually one spoke per variable. The concentric circles mark off the unit interval along each spoke, designating the transformation bounds for the associated variables. An optional mid-circle locates the mean for each variable. A data polygon or STAR is then created for an observation by plotting the standardized value of each variable on its assigned spoke, relative to the unit interval, and connecting the plotted points with straight line segments.

Fig. 1. A typical STAR profile.

2.2 The FACE Profile
In constructing the FACE profile the standardized data values control the position, size, and shape of components of the FACE. The 18 features listed in Table 1 can be controlled for both the left and right sides of FACEs.

Table 1: Controllable features of FACEs
<table>
<thead>
<tr>
<th>Key</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPHA</td>
<td>Upper Hair line</td>
</tr>
<tr>
<td>LOHA</td>
<td>Lower Hair line</td>
</tr>
<tr>
<td>HSSL</td>
<td>Hair Shading Slant</td>
</tr>
<tr>
<td>DAHA</td>
<td>Darkness of Hair</td>
</tr>
<tr>
<td>HPEB</td>
<td>Horizontal Position of Eyebrow</td>
</tr>
<tr>
<td>VPBR</td>
<td>Vertical Position of Eyebrow</td>
</tr>
<tr>
<td>CUBR</td>
<td>Curvature of Eyebrow</td>
</tr>
<tr>
<td>DEEB</td>
<td>Density of Eyebrow</td>
</tr>
<tr>
<td>HPEY</td>
<td>Horizontal Position of Eye</td>
</tr>
<tr>
<td>VPBE</td>
<td>Vertical Position of Eye</td>
</tr>
<tr>
<td>EYES</td>
<td>Eye Slant</td>
</tr>
<tr>
<td>EYS1</td>
<td>Eye Size</td>
</tr>
<tr>
<td>POPU</td>
<td>Position of Pupil</td>
</tr>
<tr>
<td>PUS1</td>
<td>Pupil Size</td>
</tr>
<tr>
<td>NOS1</td>
<td>Nose Size</td>
</tr>
<tr>
<td>CUMO</td>
<td>Curvature of Mouth</td>
</tr>
<tr>
<td>SMO1</td>
<td>Size of Mouth</td>
</tr>
<tr>
<td>FAL1</td>
<td>Face Line</td>
</tr>
</tbody>
</table>

Construction of FACEs consists of taking elements of the transformed data vector to control one of three construction elements:
a) Position of a component; X- or Y-coordinate or angle of rotation. For example, the vertical position of the eyebrows.
b) Parameter of a geometrically defined curve. For example, the radius of an arc of the eyes.
c) Shape and location of a parameterized curve. For example, the faceline is constructed by interpolating between base minimal and maximal 5th degree polynomials in t.

Further details are given in Flury [2].

To illustrate the FACE technique, the individual effects of the low and high settings on the three features mentioned as examples of construction elements are shown by the upper row of profiles in Fig. 2 (low settings are on the reader's left). The lower profiles show the composite effects of all features set at low, nominal, and high.

Fig. 2. FACE profiles showing individual and composite effects of extreme settings.

2.3 STARS Versus FACEs
Both techniques process n-dimensional data but their resulting graphical forms are distinctly different. This can have a marked effect on their application and interpretation. The following compares several critical areas.

Ease-of-use and Understanding STARS can be quickly understood and easily applied. FACEs requires training in application and interpretation. A set up stage is required for each application to judiciously assign variables to facial characteristics.

Effectiveness The human has a special innate ability for visually processing faces. The graphic elements of a face are instantaneously synthesized into a single image and then processed as a meaningful whole. As such, FACEs are more applicable to global applications such as clustering and identification of trends. Correlations between variables are readily noticeable with FACEs. STARS provide greater local detail. Data values can be closely approximated from the displays. Since spokes on STARS have a natural ordering, they are well suited for displaying serial data comprising an observation. Overlaid STARS provide a direct comparison of multivariate observations, overlaid FACEs are very confusing.

1 Availability of the construction algorithms is discussed in Flury and Riedwyl [3].
Bias and Weighting  Certain characteristics of FACES are far more pronounced than others. Thus, considerable bias can be introduced. This can be used to an advantage if nonuniform weighting is desired. With both STARS and FACES variables can be assigned to multiple components to achieve weighting.

Symmetry  Faces are inherently symmetric. This can be effectively exploited for paired comparison studies by assigning paired variables to corresponding left and right facial features. Paired variables can also be assigned to symmetric positions of STARS but the effect appears not nearly as forceful as with FACES.

Variable Assignments and Groupings  Natural variable groupings can be assigned to related features of FACES. Thus one can concentrate on variable subsets without having to redo the graphics. Also, certain characteristics of FACES can be linked to a physical meaning. For example, a smile/frown may represent success/failure, much hair may represent cold climates, large nose may indicate high pollution, etc.

Limitations  Certain characteristics of FACES are somewhat interdependent. For example, the pupil size cannot exceed the size of the eye. Also, the extent of certain characteristics is limited. For example, the eye can at most degenerate to a straight line. No such dependencies or restrictions apply to STARS.

Time and Cost Considerations  FACES are considerably more detailed than STARS and thus are more time consuming and costly to generate. With STARS, fewer variables lead to less complex profiles. With FACES, all detail is always drawn.

3. The STARS and FACES Procedures  The STARS and FACES procedures can be invoked with the normal SAS PROC statement or with a set of full-screen utility menus as discussed in section 3.5. The following gives the PROC statements and associated procedure information statements.

3.1 The STARS and FACES PROC Statements  The statements for invoking the procedures are:

PROC STARS/FACES options;
   ID variable;
   CLASS variable;
   COLORS variable;
   WEIGHT variable;
   BY variables;
   VARS variables;

A large number of options are provided with the procedure statements. These options provide the controls shown by the menus in Fig. 3 of section 3.5. Most menu entries are self-explanatory — the following briefly discusses several of the less obvious. (Details are given in Gogel [6]).

Marginal data limit - a value (ML), if given, initiates an inspection of all variables against low-high specification limits (specs). If ML is fractional, data values falling within ML*100 percent of the specs, on the interior, are classified as "marginal". Values falling inside these "marginal" limits are classified "good", values falling beyond the specs are classified as "bad". If ML is integer, observations with one and at most ML data values falling beyond the specs are classified as "marginal". Observations with more than ML data values falling beyond the specs are classified as "bad". Those with all values within specs are classified as "good". With color devices, the classifications can be clearly distinguished using color.

No. of obs. to aggregate - when given, successive sets of observations are aggregated to form "mean" profiles. For quality control applications, this specifies the sample size (uniform).

No. of obs. for base statistics - indicates a beginning subset of observations on which to compute the base statistics. Default is all observations. For quality control, this may indicate an initial period considered in control.

Probability level for multivariate tests - alpha probability level which initiates one of two multivariate tests and a normality test on individual variables. If observations are aggregated, Hotelling's T² test on multivariate means is applied. Otherwise Wilk's multivariate outlier test is applied for individual observations.

Colors assigned to 8 positions - with color devices the user can designate a sequence of colors for controlling color (see section 3.4).

Type of spokes - for STARS only, spokes can be restricted to the INSIDE or OUTSIDE of the data polygon, can be of several lengths, can indicate the low-high specs or data min-max ranges, or can segment the profile into vertical and/or horizontal symmetric regions for paired comparisons.

Aggregate obs. in CLASSES - if the CLASS statement is given, it forms an aggregate profile of all observations within a class.

Show the data pts. only, no polygon - this option maximizes the "data ink" (see Tuftte [11]). For STARS, only the data points are shown, no connecting lines forming polygons. For FACES, only those components assigned variables are drawn.

In addition, with the FACES procedure, an option provides for a display of FACES showing individually the control of the variables of the facial features.

3.2 Procedure Information Statements  The following procedure information statements supplying special variable(s) are allowed:

ID  - supplies ID caption for each obs.
CLASS  - indicates a data grouping structure.
COLORS  - specifies color of profile for each obs.
WEIGHT  - controls size of profile relative to 1.
BY  - performs analysis for each BY-group.
VARS  - selects subset of vars. from data set.

3.3 Variable Specifications and Transformations  A specifications ("specs") file can be supplied to specify for each variable a) its assignment to a profile component, b) a transformation code, and c) low, nominal, and high specification limits. This file is required with FACES to supply the variable assignments to the facial features.

With STARS, unless explicit position assignments
are given, variables are uniformly assigned successive positions around the periphery of the circle. Assignment of variables to features of FACES is made through the "specs" file using the four letter keys given in Table 1. Suffixes of "R" and "L" designate right and left assignments only.

The transformation code is a numeric value which gives the sigma range about the mean to be mapped onto the (0,1) unit interval. For example, the default value of 3.0 maps the data within 3 standard deviations of the mean onto (0,1). A special code of zero indicates either the low-high specs, if given, or the min-max data range defines the data interval to be mapped onto (0,1). Negative code reverses a variable's effect on the profile while a "W" suffix (Winsorize) shrinks extreme data values onto the (0,1) interval end-points.

3.4 Color Control
On color devices, color can be employed to distinguish between components of the plot or to highlight important data characteristics. For example, the grid, labels, polygon, etc. of STARS and the hair, eyes, nose, etc. of FACES can be distinguished with colors. "Good", "marginal", and "bad" data can be differentiated and outliers can be highlighted using color. Data groupings can be set off with color and overlaid profiles can be drawn in different colors. The user controls the set of colors used through a COLORS parameter on the PROC statement. The colors selected for components of the plot depend upon the use of the ML (Marginal limit) and OVERLAY options and the CLASS and COLORS statements and finally characteristics of the data. Details are given in Gugel [6].

3.5 Full Screen Menus
Since a large number of options are provided and their settings can have a significant impact on the resulting displays, a menu facility was developed for easy interactive and iterative execution. (The SAS/FSP product was used as suggested by Hardison and Muller [7]). Menu panels are provided for building the "specs" file and for executing both procedures. Fig. 3 shows the two panel menus for executing the STARS procedure.

4. Examples
This section gives an application illustrating each procedure. Space only allows for a very brief discussion — a more complete analysis is found in Gugel [6]. (Color was not applied for this paper since it could not be reproduced.)

4.1 PROC STARS for studying serial data
The first example illustrates the use of the STARS procedure for multivariate paired comparisons and for studying correlations and variation in serial data. It graphically explores the data. Details are given in Gugel [6].

1. Nearly all states have reduced their rates from 1967 through 1975. Reduction in rural rates have been especially uniform (smooth inward spiral) over the nine years in the states of AL, GA, IL, MO, and OR. However, two states, GA and RI are exceptions.

2. Between states there seems to be little correlation between rural and urban rates. That is, states with high rural rates don't necessarily have high urban rates. Similarly for low rates. Within states there is considerable correlation between rural and urban rates. That is, states with high rural rates don't necessarily have high urban rates. Similarly for low rates.

3. The following lists several observations, some taken from the entire display of 48 states (see also Gugel [6]).

1. Rural rates (left) are generally substantially higher than urban rates; the states MD, MA, and NJ are exceptions.

2. Nearly all states have reduced their rates from 1967 through 1975, especially rural rates. Reduction in rural rates have been especially uniform (smooth inward spiral) over the nine years in the states of CO, HD, MH, NC, and TN. Only in the states of GA and RI have the urban rates increased. RI is an anomaly.

3. Between states there seems to be little correlation between rural and urban rates. That is, states with high rural rates don't necessarily have high urban rates. Similarly for low rates. Within states there is considerable correlation between rural and urban rates. That is, states with high rural rates don't necessarily have high urban rates. Similarly for low rates.
RI, show a negative correlation between the two rates.

4.2 PROC FACES for exploring multivariate data
This example re-explores the pollution-mortality data studied by McDonald and Schwing [10]. The data consists of 16 measurements recorded on 60 cities — 4 climate variables, 3 pollutants, and a mortality rate. The cities were grouped into four geographical regions: northeast, south, midwest, and west.

A FACE is constructed for each city. The assignment of variables to the facial features, together with the transformation codes and low-high specs, is given by the "specs" file listed in Fig. 5. Similar variable types were assigned to related features. With the exception of the nose, both left and right features of faces were assigned to the same variable.

The four climate variables control the general contour of the faces. January and July temperatures were assigned to the lower and upper hairlines respectively. Their joint variation controls the amount of hair — much hair indicates a year round warm climate.

The socioeconomic variables were assigned to properties of the eyes and eyebrows. The strongly correlated variables, education (EDUC) and percent white-collar (PCTWHCOL), control the size of the eyes and pupils. Percent nonwhite (PCTNOWHT) and percent poor (PCTPOOR), also strongly correlated, were assigned to the vertical position of the eyes and eyebrows.

The pollutants, HC and NOX, were assigned respectively to the left and right side of the nose. High correlation would result in symmetrical noses. SO2 was assigned to the mouth size and mortality rate to the curvature of the mouth. A low mortality rate causes a smile while high mortality results in a sad face.

The data was sorted by region and mortality rate within region. Fig. 6 shows nine individual profiles from the south followed by seven profiles from the west. Fig. 7 gives the aggregate profile for each of the four regions.

To illustrate the interpretation of the FACES, New Orleans (NWOR) will be examined in some detail. The following observations can be made:
1. High January and July temperatures (wide hair).
2. High precipitation (dark and slanted hair).
3. Low percent sound housing (SNHOUS) (narrow and convex eyebrows).
4. Low educational level (small eyes) compared with relatively high percent white-collar (large pupils).
5. Relatively high percent poor (low eye level) and high percent nonwhite (low eyebrows).
6. Lower than average percent over 65 (PCTPOOR) (eyes slanted down).
7. Low pollution (narrow nose and small mouth).
8. High rate of mortality (sad face).

Examining the 16 profiles the following observations can be made:
1. The south and west have distinctly different profiles.
2. The south is far more homogeneous, has warmer climates (more hair) and greater precipitation (darker hair). Miami is exceptionally warm and wet.
3. Education (eye size) and sound housing (curvature and density of eyebrows), in the south, are positively correlated with each other and negatively correlated with mortality rate.
4. HOUSTON has a very low percent over 65 (eyes-lant) while BRNM has a large nonwhite population (low eyebrows).

Fig. 5. Listing of specifications file.

The pollutants, HC and NOX, were assigned respectively to the left and right side of the nose. High correlation would result in symmetrical noses. SO2 was assigned to the mouth size and mortality rate to the curvature of the mouth. A low mortality rate causes a smile while high mortality results in a sad face.

The data was sorted by region and mortality rate within region. Fig. 6 shows nine individual profiles from the south followed by seven profiles from the west. Fig. 7 gives the aggregate profile for each of the four regions.

To illustrate the interpretation of the FACES, New Orleans (NWOR) will be examined in some detail. The following observations can be made:
1. High January and July temperatures (wide hair).
2. High precipitation (dark and slanted hair).
3. Low percent sound housing (SNHOUS) (narrow and convex eyebrows).
4. Low educational level (small eyes) compared with relatively high percent white-collar (large pupils).
5. Relatively high percent poor (low eye level) and high percent nonwhite (low eyebrows).
6. Lower than average percent over 65 (PCTPOOR) (eyes slanted down).
7. Low pollution (narrow nose and small mouth).
8. High rate of mortality (sad face).

Examining the 16 profiles the following observations can be made:
1. The south and west have distinctly different profiles.
2. The south is far more homogeneous, has warmer climates (more hair) and greater precipitation (darker hair). Miami is exceptionally warm and wet.
3. Education (eye size) and sound housing (curvature and density of eyebrows), in the south, are positively correlated with each other and negatively correlated with mortality rate.
4. HOUSTON has a very low percent over 65 (eyes-lant) while BRNM has a large nonwhite population (low eyebrows).
The west has some very extreme values: high HC and NOX (nose) in LOSA and SNFR, and low relative humidity in DNVR (faceline).

6. The west overall is high in sound housing (dark eyebrows), high on education and percent white-collar workers (large eyes and pupils), low pop/household (eyes far apart), and low on mortality rate (generally smiles).

7. Overall, HC and NOX are strongly positively correlated (symmetric noses) and appear not to be correlated with mortality rate. HC and NOX are exceptionally high in California.

The four aggregate profiles again show the year round warm and wet climate for the south and the sharp differences on many variables between the west and the other regions.

5. Conclusions
The new graphical profiles procedures provide useful tools for simultaneously presenting information on a large number of variables. Color and several analytical options offer added flexiblity and diagnostic capabilities for exploring complex multivariate data. These procedures provide graphical methods which require little or no statistical expertise and can be used by "the man in the street". The examples briefly demonstrated their use in several application areas including: data description and summarization, identification of trends and outliers, multivariate paired comparisons, data clustering, and correlation and variance analysis.

ACKNOWLEDGEMENTS
Special thanks must be extended to Professor Bernhard Flury, Department of Statistics, Purdue University, and Professor Hans Riedwyl, Department of Statistics, University of Berne, Switzerland, for providing the base code of the algorithm for constructing the FACES. Thanks must also go to Dr. Thomas J. Lorenzen of the Mathematics Department, General Motors Research Laboratories, for providing a number of valuable suggestions.

REFERENCES