SCAN: AN INTERACTIVE GRAPHICS SAS PROCEDURE FOR SPECIALIZED CORRELATION ANALYSIS OF BIVARIATE DATA

Robert N. Rodriguez and David A. DeNardis, General Motors Research Laboratories

ABSTRACT

Continuous bivariate data are routinely analyzed with the aid of a scatter plot and a correlation. However, these tools have limitations: a scatter plot is not feasible if the data are weighted or grouped, and a correlation measures only the strength of linear dependence. These limitations are overcome by a new procedure, referred to as SCAN, which was developed for the Statistical Analysis System (SAS).

The SCAN Procedure implements a new graphical technique which displays detailed information about the joint distribution of a bivariate data set. The graphical display is referred to as a "correlation fingerprint", because it can be used to distinguish and identify correlation structures. For example, correlation fingerprints can reveal power transformation relationships and departures from bivariate normality.

The interactive nature of the Procedure itself is also novel. Execution of SCAN can be controlled by procedure options similar to those of most SAS procedures, or by an interactive internal command processor, or by a combination of these modes. Thus the user can invoke the Procedure with a few specifications, and then interactively modify these specifications with commands internal to SCAN.

1. INTRODUCTION

The two most basic tools for the analysis of continuous bivariate data are a scatter plot and a Pearson correlation. Both of these tools are provided in the Statistical Analysis System (SAS). Scatter plots can be generated with the PLOT and GPLOT Procedures, and a variety of correlations can be computed with the CORR Procedure; see Barr et al. (1979).

Since correlation measures only the strength of the linear relationship between two variables, nonlinear relationships (as well as outliers) may be missed if the analyst does not examine a scatter plot. This problem is illustrated in Figures 1 and 2. The Pearson correlations (r) of the scatter plots in these two figures are both equal to -0.04, but the data in Figure 1 are randomly scattered, and the data in Figure 2 exhibit a quadratic relationship.

Unlike the artificial examples in Figures 1 and 2, scatter plots commonly encountered can be difficult to interpret and distinguish. Moreover, situations do occur in which scatter plots are not feasible, even though a Pearson correlation can be computed. For example, in some cases, the bivariate observations \((x_i, y_i)\) in a sample of size \(N\) are not equally important, and their relative importance is indicated by a third variable \(w_i\), which is referred to as the weight of the \(i\)th observation. The WEIGHT statement for the CORR Procedure enables one to compute a weighted product-moment correlation, but this value will not be reflected by a scatter plot of the \((x_i, y_i)\) observations. Likewise, if the data are grouped, a correlation can be obtained with a FREQ statement for the CORR Procedure, but a scatter plot is inappropriate for displaying the data.
The SCAN Procedure described in this report is an interactive graphics SAS procedure which is intended to overcome these limitations. The name SCAN is an acronym for Specialized Correlation Analysis, and, as the name implies, the procedure generates a graphical display which the user can "scan" in order to gain insight about the correlation structure of the two variables. This display is termed a "correlation fingerprint", because it resembles a human fingerprint, and because it can be used to identify or distinguish the dependence structure of a bivariate data set.

Both this graphical technique and the manner in which it has been implemented as a SAS procedure are new. The correlation fingerprint method, which was introduced by Rodriguez (1981), is summarized in Sections 2 and 3 of this paper. The design of the SCAN Procedure, described in Section 4, uniquely combines interactive approaches described by Gugel (1981). Section 5 illustrates the execution of the Procedure with example sessions. Finally, some conclusions are stated in Section 6.

2. DEFINITION OF A CORRELATION FINGERPRINT

The basis for a correlation fingerprint is an extension of the familiar formula for the Pearson correlation $r$, which can be written as

$$
r = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
$$

As usual, $\bar{x}$ and $\bar{y}$ denote the averages, and $s_x$ and $s_y$ denote the standard deviations of the $x$-data and $y$-data, respectively. The "$N$" version of the formula for standard deviation is used to obtain $s_x$ and $s_y$; for example,

$$
s_x = \left( \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right)^{1/2}
$$

For convenience, standardized values will be indicated with primes:

$$
x'_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, \ldots, N
$$

$$
y'_i = \frac{y_i - \bar{y}}{s_y}, \quad i = 1, \ldots, N
$$

We will define a correlation fingerprint as the contour plot of the function

$$
r(\gamma, \lambda) = \frac{1}{N} \sum_{i=1}^{N} |x'_i|^\gamma |y'_i|^\lambda \text{sign}(x'_i y'_i)
$$

over the range $0 \leq \gamma, \lambda \leq 3$. Here "sign(·)" denotes the function

$$
sign(u) = \begin{cases} 
-1 & \text{if } u < 0 \\
0 & \text{if } u = 0 \\
+1 & \text{if } u > 0
\end{cases}
$$

Since $r(1,1) = r$, the function $r(\gamma, \lambda)$ is clearly a continuous extension of the formula for correlation. Moreover, the special values $r(1,3)$, $r(3,1)$, and $r(3,3)$ are the product moments $m_{13}$, $m_{31}$, and $m_{33}$, respectively; see Kendall & Stuart (1977). In general, $r(\gamma, \lambda)$ can be regarded as a modified product moment, based on power transformations of the magnitudes of the data. Thus a correlation fingerprint is a smooth graphical extension of the Pearson correlation which provides a highly detailed picture of the dependence structure of the data.

Correlation fingerprints are also defined for weighted or grouped bivariate data. These are determined with versions of $r(\gamma, \lambda)$ obtained by extending the Pearson correlation formulas for weighted and grouped data. If the data are ungrouped but weighted, then

$$
r(\gamma, \lambda) = \frac{1}{N} \sum_{i=1}^{N} w_i |x'_i|^\gamma |y'_i|^\lambda \text{sign}(x'_i y'_i)
$$

assuming that the weights $w_i$ are nonnegative and sum to unity. (The standardized values, $x'_i$ and $y'_i$, must be computed with weighted means and weighted standard deviations.)

If the data are grouped, then

$$
r(\gamma, \lambda) = \frac{1}{100N} \sum_{m=1}^{M} \sum_{n=1}^{N} f_{mn} |x'_m|^\gamma |y'_n|^\lambda \text{sign}(x'_m y'_n)
$$

Here $x'_m, \ldots, x'_{m+1}$ and $y'_n, \ldots, y'_{n+1}$ denote the standardized midpoints of the intervals for the $x$-data and $y$-data, respectively, and $f_{mn}$ is the percent of observations grouped in the interval centered at $(x'_m y'_n)$.

3. CORRELATION FINGERPRINT INTERPRETATIONS

Two sets of bivariate data with the same value of $r$ but with different correlation structures can be distinguished by their correlation fingerprints. More importantly, correlation fingerprints can be used to identify specific dependence structures:

- **Power transformations.** Figure 3 is a correlation fingerprint for the bivariate normal data
plotted in Figure 1, and Figure 4 is a fingerprint for the transformed data plotted in Figure 2. In contrast to the curved contours of Figure 3, the contours of Figure 4 are parallel lines with a slope approximately equal to \(-1/2\).

**Figure 3.** Correlation fingerprint of data in Figure 1.

**Figure 4.** Correlation fingerprint of data in Figure 2.

In general, if \(x\)-data and \(y\)-data are related by a power function such that \(|y| = |x|^p\) for some power \(p\), then \(r(y, \lambda)\) is constant over the lines with equation

\[
\lambda = \text{constant} - \gamma/p,
\]

whose slope is \(-1/p\). Simulation work (Rodriguez (1981)) indicates that nearly parallel lines will be evident in such a fingerprint even if the power function relationship is weakened by chance variation.

**Bivariate normality.** The assumption of bivariate normality plays such an important role in statistical work, that it is of interest to know what types of fingerprint patterns to expect when examining bivariate normal data.

We define the theoretical analogue of \(r(Y, \lambda)\) for the bivariate normal model as the function

\[
r^*(Y, \lambda) = E[X|Y|^{\rho \lambda} \text{sign}(XY)],
\]

where \(X\) and \(Y\) have a standard bivariate normal distribution with correlation \(p\). This definition is motivated by the fact that \(r(l, \lambda)\) is a nearly unbiased estimator of \(p\) for large samples. The function \(r^*(Y, \lambda)\) can be expressed as

\[
2^{\frac{\rho(1-\rho)}{2}} 2^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right)^2 \Gamma\left(\frac{1}{2} + \frac{\rho}{2}\right) \Gamma\left(\frac{1}{2} + \frac{1}{2}\right) - P\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right),
\]

where \(P(\cdot, \cdot, \cdot, \cdot, \cdot, \cdot)\) is the Gauss hypergeometric function; see Rodriguez (1981). If \(\rho \neq 0\), then one can compute the contours of \(r^*(Y, \lambda)\), which serve as templates for comparison with data fingerprints.

A bivariate normal template (corresponding to \(\rho = 0.25\)) is illustrated in Figure 5. In general, these templates have a symmetric whorled pattern. Simulation with bivariate normal data demonstrates that for sample sizes greater than 500, a close match ensues between the correlation fingerprint and the template for \(\rho\) equal to \(r\).

**Bivariate normal template**

**Figure 5.** Bivariate normal template for \(\rho = 0.25\).
4. THE SCAN PROCEDURE

The correlation fingerprint supplements, rather than replaces, the methods ordinarily used for bivariate data analysis. Furthermore, the interpretation of a correlation fingerprint can influence the direction in which subsequent analysis proceeds. Thus the SCAN Procedure is intended primarily for interactive use, and it is designed to provide the user with both flexibility and power. (See Gugel (1981) for a lucid description of principles for creating interactive SAS procedures.)

Specifically, the SCAN Procedure is an interactively command-driven procedure with three modes of execution:

Standard execution. The user may completely specify the execution by providing options and parameters in a PROC statement; default values are preassigned to options and parameters which are not supplied. This mode of execution is common to most SAS procedures, and it is particularly appropriate for the SCAN Procedure if the user wants to direct the graphical output to a file for postprocessing. (See Example A of Section 5.)

Totally interactive execution. The user may invoke the procedure with the statement PROC SCAN; RUN; and then proceed totally interactively by means of an internal command language. As in PROC EDITOR, the user is prompted for commands, which the procedure parses with its own interpreter. This mode of execution enables the user to generate fingerprints, request special facilities (such as template overlays), and add enhancements (such as optional character styles) in any order. (See Example B in Section 5.)

Mixed execution. The SCAN Procedure can also be executed by a combination of the standard and totally interactive modes. For example, a user at a graphics terminal may initiate a SCAN session with a PROC statement which requests a correlation fingerprint for two variables, X and Y. Then, after examining this fingerprint, internal commands can be issued to request a second fingerprint for X and Y weighted by a third variable W. (See Example C in Section 5.)

The mixed execution mode of the SCAN Procedure provides the user with exceptional flexibility. To the best of our knowledge, no other SAS procedure for data analysis has this capability for combining standard and totally interactive execution.

5. EXAMPLES

The examples in this section illustrate the use of the SCAN Procedure. In each example the same data set (referred to as OCTANE) is analyzed, but a different form of execution is presented.

The observations in OCTANE (see Table 1) are the gasoline octane requirements for N=49 cars, as determined by a trained rater (x) and the gasoline customer (y), who drives the car. These requirements correspond to the 1978 model year cars surveyed in a program conducted by the Coordinating Research Council of the automobile and petroleum industries; see Rodriguez & Taniguchi (1980). The objective of this program was to quantify the relationship between rater and customer octane requirements, in order to simplify the subsequent prediction of customer requirements.

The correlation of the rater and customer requirements, which are plotted in Figure 6, is r=0.57. However, it can be argued that a correlation of the observations weighted by relative sales of the sample cars (see Table 1) is more meaningful; this correlation is equal to 0.62.

The correlation fingerprint supplements, rather than replaces, the methods ordinarily used for bivariate data analysis. Furthermore, the interpretation of a correlation fingerprint can influence the direction in which subsequent analysis proceeds. Thus the SCAN Procedure is intended primarily for interactive use, and it is designed to provide the user with both flexibility and power. (See Gugel (1981) for a lucid description of principles for creating interactive SAS procedures.)

The correlation fingerprint is more meaningful; this correlation is equal to 0.62.

The correlation of the rater and customer requirements, which are plotted in Figure 6, is r=0.57. However, it can be argued that a correlation of the observations weighted by relative sales of the sample cars (see Table 1) is more meaningful; this correlation is equal to 0.62.

In spite of the similarity of these two correlations, sales weighting may have changed some aspect of the relationship between rater and customer requirements, which is not measured by correlation. We will explore the effect of weighting in the examples that follow.

Table 1

<table>
<thead>
<tr>
<th>OCTANE REQUIREMENTS* of 1978 MODEL YEAR VEHICLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater Customer Sales Rater Customer Sales</td>
</tr>
<tr>
<td>93.0 93.0 0.0077 94.0 91.5 0.0069</td>
</tr>
<tr>
<td>91.0 89.0 0.0232 96.0 91.5 0.0241</td>
</tr>
<tr>
<td>90.5 84.0 0.0398 86.0 86.7 0.0007</td>
</tr>
<tr>
<td>96.0 88.0 0.0271 88.0 84.0 0.0085</td>
</tr>
<tr>
<td>93.0 85.0 0.0261 75.0 78.2 0.0127</td>
</tr>
<tr>
<td>88.0 82.0 0.0254 75.0 78.2 0.0127</td>
</tr>
<tr>
<td>95.0 91.0 0.0226 84.0 84.0 0.0127</td>
</tr>
<tr>
<td>89.0 85.0 0.0254 86.0 86.7 0.0007</td>
</tr>
<tr>
<td>91.0 85.0 0.0173 96.0 91.5 0.0078</td>
</tr>
<tr>
<td>93.0 89.5 0.0165 90.0 91.5 0.0073</td>
</tr>
<tr>
<td>94.0 90.0 0.0186 94.0 91.5 0.0093</td>
</tr>
<tr>
<td>97.0 91.0 0.0326 86.0 91.0 0.0186</td>
</tr>
<tr>
<td>92.0 88.0 0.0121 86.0 87.0 0.0059</td>
</tr>
<tr>
<td>92.0 86.0 0.0333 82.0 84.0 0.0025</td>
</tr>
<tr>
<td>96.0 89.0 0.0303 85.0 87.0 0.0030</td>
</tr>
<tr>
<td>89.0 83.0 0.0277 86.0 86.0 0.0034</td>
</tr>
<tr>
<td>91.0 85.0 0.0254 82.0 74.0 0.0173</td>
</tr>
<tr>
<td>89.0 84.0 0.0388 84.0 90.0 0.0076</td>
</tr>
<tr>
<td>97.0 90.0 0.0088 66.0 85.0 0.0228</td>
</tr>
<tr>
<td>91.0 80.0 0.0137 84.0 89.0 0.0093</td>
</tr>
<tr>
<td>90.0 86.0 0.0174 96.0 86.0 0.0074</td>
</tr>
<tr>
<td>90.0 86.0 0.0173 86.0 87.0 0.0053</td>
</tr>
<tr>
<td>88.0 86.0 0.0127 86.0 88.0 0.0058</td>
</tr>
<tr>
<td>89.0 86.0 0.0092 81.0 84.0 0.0137</td>
</tr>
<tr>
<td>94.0 88.0 0.0069</td>
</tr>
</tbody>
</table>

*Units of Research Octane Number (RON)
Source: 1978 CRC Program

The correlation of the rater and customer requirements, which are plotted in Figure 6, is r=0.57. However, it can be argued that a correlation of the observations weighted by relative sales of the sample cars (see Table 1) is more meaningful; this correlation is equal to 0.62.

In spite of the similarity of these two correlations, sales weighting may have changed some aspect of the relationship between rater and customer requirements, which is not measured by correlation. We will explore the effect of weighting in the examples that follow.
Example A: Generating a Fingerprint in Batch

This example illustrates running the SCAN Procedure in batch to generate a correlation fingerprint of rater and customer requirements:

```sas
// EXEC SAS
// IN DD DSN=userid.CARS.DATA, DISP=SHR
// PLFILE DD DSN=userid.COMPRESS, DISP=OLD
PROC SCAN DATA=IN.OCTANE COMPRS X=RATER Y=CUSTOMER;
```

Here the SAS dataset OCTANE is contained in a SAS data base named CARS.DATA, which is allocated to a file named IN. The SCAN Procedure statement identifies RATER as the x-variable and CUSTOMER as the y-variable; the option COMPRS directs the output to a compress file for postprocessing. The fingerprint obtained is shown in Figure 7.

Example B: Adding a Template In the Totally Interactive Mode

In this example the user runs the Procedure totally interactively at a Tektronix terminal to generate a fingerprint of rater and customer requirements and then overlay a bivariate normal template. User entries are in lowercase, and system responses are in uppercase.

```sas
proc scan data=in.octane; run;
enter:
title 'data fingerprint and normal template'
enter:
scan customer*rater
enter:
template
```

Note that the Procedure statement contains limited information; it serves only to invoke the Procedure. A fingerprint (as in Figure 7) with a title is drawn after the internal command "scan customer*rater" is issued; the first variable (in this case, customer) following the SCAN command is taken to be the y-variable. Then the user requests an overlay of a bivariate normal template; by default, the value of \( \rho \) is taken to be the sample correlation \( r=0.57 \). This overlay (see Figure 8) reveals a departure from bivariate normality.

Example C: Adding a Weight Variable in Mixed Mode

This example illustrates the execution of the Procedure in the mixed mode to generate a correlation fingerprint of rater and customer requirements and then regenerate the fingerprint for the

---

**Figure 6.** Scatter plot of octave requirements.

**Figure 7.** Correlation fingerprint of octave requirements.

**Figure 8.** Octane requirement fingerprint overlaid with bivariate normal template.
data weighted by sales.

```r
proc scan x=rater y=customer; run;
```

ENTER:
newgraph
ENTER:
title='fingerprint of octane requirements'
ENTER:
title2='weighted by sales'
ENTER:
tlet alpha=italic style=triplex
ENTER:
longlab
ENTER:
corr
ENTER:
weight sales
ENTER:
scan customer=rater
```

Here the Procedure draws a fingerprint (as in Figure 7) immediately after the "run" command is entered. (Note that the x- and y-variables are identified in the Procedure statement.) Then the Procedure requests additional information. By responding with the internal commands shown, the user obtains a fingerprint (see Figure 9) of the sales-weighted requirements. The "corr" command causes the special contour whose value equals the (weighted) sample correlation to be drawn. The "tlet", "alpha", and "style" commands designate the alphabet and style types for the two titles specified, and the "longlab" command identifies the x- and y-variables on the fingerprint; these are some of the many graphical enhancements available.

A comparison of Figures 7 and 9 reveals the effect of introducing sales weights. The parallel line pattern of the fingerprint in Figure 9 suggests a power transformation relationship of the form

\[ |y'_{1}| = |x'_{1}|^{2/3} \]

between sales-weighted rater \((x_{1})\) and customer \((y_{1})\) requirements which would be difficult to detect from a scatter plot or a correlation.

6. CONCLUSIONS

The following conclusions result from our development and application of the SCAN Procedure:

1. The SCAN Procedure provides a versatile interactive capability in SAS for the display and interpretation of correlation in bivariate data.

2. The new graphical method implemented by the SCAN Procedure can be used to diagnose nonlinear relationships and departures from bivariate normality, information which is not conveyed by a correlation.

3. The design of the SCAN Procedure demonstrates a highly effective approach to the development of user-oriented SAS tools for data analysis.

ACKNOWLEDGEMENT

The authors are indebted to Harold W. Gugel of the GMR Computer Science Department for valuable advice concerning the development of the SCAN Procedure as an interactive graphics tool.

REFERENCES


