PROC IDPLOT: A New Plotting Procedure With Point Labeling Capabilities
Warren F. Kuhfeld, The University of North Carolina at Chapel Hill

ABSTRACT

PROC IDPLOT is a Version 5, base product, line printer plotting procedure that allows points to be labeled with long point labels. PROC IDPLOT has four plotting methods, three of which can guarantee that no point labels will be hidden. This is accomplished by using varying amounts of nonmetric rank and metric distance information when computing row and column point locations. The RANK method locates each point on a separate line of the plot so long point labels will not be obscured. The MODRANK method also locates each point on a separate line, but the number of lines separating points is a function of the differences between adjacent y coordinates. The GENRANK method computes the row number with a formula that uses a variable weighting of y coordinate and y rank information. The METRIC method produces ordinary scatterplots. PROC IDPLOT is currently documented in "IDPLOT and OPRINT Procedures, Preliminary Documentation" and "P-146 Changes and Enhancements to the Version 5 SAS System".

INTRODUCTION

A standard data summarization tool is the ordinary scatterplot. The usual goal when making a scatterplot is to produce an accurate display of shape and scatter. In some situations there is a second goal: accurately displaying the names of the points in the plot. PROC PLOT can be used to produce printer plots that accurately display shape, but PROC PLOT's single character plot symbols are not always sufficient. Sometimes it would be useful to have plots where long point labels are permitted, and no point label is hidden by any other point label. This poster will describe a new Version 5 base product SAS ® procedure, PROC IDPLOT, which has four methods of producing printer plots where the emphasis is placed on adequately labeling the points.

The methods are: RANK, based on the rank model first implemented in Kuhfeld's (1984) RANKPLOT macro; MODRANK, based on the modified-rank model; GENRANK, based on the generalized-rank model; and METRIC which produces an ordinary scatterplot (Kuhfeld, 1986). The modified-rank and generalized-rank models are extensions of the original rank model that use both nonmetric rank, and metric distance information to locate points on the plot. PROC IDPLOT can use the three rank models to produce plots where no ID's are hidden. All four methods use a simple formula to map the original data values to a row or column of a plot. These methods will be discussed in more detail in the next section. Plots of the data in Table 1 will be used as examples. The data are scores on two principal components, computed from monotonic transformations of car preference judgments collected from 25 individuals.

MODELS

Consider the problem of mapping n points to p rows of a graph numbered 1 through p. Assume n > 1, p > 1, and the y coordinates are sorted into descending order. Let y(i) be a y coordinate, and r(i) be a row print position, (1 ≤ i ≤ n). The four models used by PROC IDPLOT for mapping to rows will be presented below. The column models can be derived from the row models.

The METRIC method (Figure 1) uses the following formula to map coordinates to row print positions, producing an ordinary scatterplot:

\[ r(i) = \text{FLOOR} \left( \frac{y(i) - y(1)}{y(1) - y(n)} \right) \cdot \frac{p-1}{0.5} + 1 \]

The result of the FLOOR function is the largest integer less than or equal to the argument. The row is proportional to the distance to the first point and the mapping is linear. The Figure 1 plot was generated by specifying:

PROC IDPLOT METRIC PANELS=1 LO HPOS=45;
TITLE 'FIGURE 1';
PLOT COMP2*COMP1=CAR / VREF=0 HREF=0 VINC=2 HINC=2;

The methods and procedure described in this poster were developed by the author at the University of North Carolina as a part of a software development project funded by SAS Institute Inc. The author wishes to thank Warren Sarle and Forrest Young who provided many valuable suggestions during the development of PROC IDPLOT.

SAS® is a registered trademark of SAS Institute Inc. Cary, NC.
TABLE 1

<table>
<thead>
<tr>
<th>COMPONENT ONE</th>
<th>COMPONENT TWO</th>
<th>CAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.59</td>
<td>-3.38</td>
<td>CADILLAC ELDORADO</td>
</tr>
<tr>
<td>-1.95</td>
<td>-2.38</td>
<td>CHEVROLET CHEVETTE</td>
</tr>
<tr>
<td>-1.52</td>
<td>-2.19</td>
<td>CHEVROLET CITATION</td>
</tr>
<tr>
<td>-1.87</td>
<td>-1.61</td>
<td>CHEVROLET MALIBU</td>
</tr>
<tr>
<td>-1.72</td>
<td>-2.16</td>
<td>FORD FAIRMONT</td>
</tr>
<tr>
<td>-1.19</td>
<td>-1.40</td>
<td>FORD MUSTANG</td>
</tr>
<tr>
<td>-2.40</td>
<td>-2.47</td>
<td>FORD PINTO</td>
</tr>
<tr>
<td>4.84</td>
<td>0.05</td>
<td>HONDA ACCORD</td>
</tr>
<tr>
<td>3.12</td>
<td>-1.38</td>
<td>HONDA CIVIC</td>
</tr>
<tr>
<td>-3.31</td>
<td>6.91</td>
<td>LINCOLN CONTINENTAL</td>
</tr>
<tr>
<td>-2.00</td>
<td>0.53</td>
<td>PLYMOUTH GRAN FURY</td>
</tr>
<tr>
<td>-1.81</td>
<td>-2.02</td>
<td>PLYMOUTH HORIZON</td>
</tr>
<tr>
<td>-1.97</td>
<td>-1.86</td>
<td>PLYMOUTH VOLARE</td>
</tr>
<tr>
<td>-1.13</td>
<td>1.12</td>
<td>PONTIAC FIREBIRD</td>
</tr>
<tr>
<td>3.77</td>
<td>0.63</td>
<td>VOLKSWAGEN DASHER</td>
</tr>
<tr>
<td>4.59</td>
<td>-0.41</td>
<td>VOLKSWAGEN RABBIT</td>
</tr>
<tr>
<td>8.14</td>
<td>2.71</td>
<td>VOLVO DL</td>
</tr>
</tbody>
</table>

The major difference between PROC IDPLOT's plot requests and PROC PLOT's plot requests is that a point ID variable be specified. Any valid PROC IDPLOT request will produce the same plots. PROC IDPLOT has a few additional forms that make it easier to request certain commonly needed patterns of plots. More will be said on this point later.

The options on the plot statement request: a vertical reference line be drawn at zero, a horizontal reference line be drawn at zero, a vertical axis tick increment of 2.0, and a horizontal axis tick increment of 2.0.

All options (except DATA) may appear on either the PLOT or PROC statement, there is no limit on the number of PLOT statements, or the number of plot requests that may appear on a single PLOT statement. Options specified on PLOT statements apply only to that PLOT statement. Options specified on a PROC statement apply to all PLOT statements unless overridden on a PLOT statement.

It can be seen from Figure 1 that some point labels may obscure other point labels in metric plots. This fact prompted the development of the three rank methods.

The simplest plotting model that guarantees that no point or point description will be hidden is the rank model. This is a nonmetric model where the row of each plotted point is its rank, with arbitrary assignment of ranks within ties (see Figure 2). Each point appears on a separate line, so long point ID's can be printed without any ID being hidden by any other ID.

FIGURE 1

METRIC PLOT OF COMPONENT TWO BY COMPONENT ONE CAR

COMPONENT ONE SCORES

-6.00 -4.00 -2.00 0.00 2.00 4.00 6.00 8.00 10.00

COMPONENT TWO SCORES

-4.00 -2.00 0.00 2.00 4.00 6.00 8.00

THE FOLLOWING POINTS WERE AT LEAST PARTIALLY OBSCURED IN METRIC PLOT OF COMPONENT TWO BY COMPONENT ONE CAR.

COMP2  COMP1  CAR

-1.40    -1.19    FORD MUSTANG
-2.16    -1.72    FORD FAIRMONT
-2.02    -1.91    PLYMOUTH HORIZON
-1.61    -1.97    CHEVROLET MALIBU
-1.86    -1.97    PLYMOUTH VOLARE
-0.53    -2.00    PLYMOUTH GRAN FURY
-2.47    -2.40    FORD PINTO
-6.98    -3.59    CADILLAC ELDORADO

METRIC specifies the type of plot (MODRANK is the default). PANELS=1 specifies that the plot should only span one horizontal page. (Large plots may be generated in many panels or sections and then cut and taped together.) LO, an abbreviation of LOCATE, (all options can be abbreviated to two characters) requests that a list of obscured points be printed. HPOS=45 specifies that only 45 print positions be used on the horizontal axis (to make a plot small enough to fit in the proceedings).
Define \( m \) to be a "separation parameter", equal to the number of rows that will separate each point. Let \( m \) be a positive integer and \( p = mn \). The row for the \( ith \) point in a rank plot is:

\[
r(i) = m(i - 1) + 1
\]

The row is proportional to the rank of the coordinate. Figure 2 contains a rank plot with a vertical separation of one and a horizontal separation of two. Distances are distorted but each point is properly labeled. Unlike the metric plot, which labels the axes with "round" numbers, the rank plot labels the axes with actual coordinate values. The Figure 2 plot was generated by specifying:

```
PROC IDPLOT RANK PANELS=1 HT=2 VT=1;
TITLE 'FIGURE 2';
PLOT COMP2*COMP1=CAR / VR=0 HR=0;
```

HT and VT specify the number of print positions per tick interval on the horizontal and vertical axes.

The modified-rank model differs from the rank model in that it allows for a variable amount of separation, \( m(i) \). Assume \( p > n \). (If \( p = n \), an ordinary rank plot with \( m(i) = 1 \) for all \( i \) is generated.) Let \( r(0) = 0, m(1) = 1, \) and for \( i \) greater than 1, define:

\[
m(i) = \text{FLOOR} \left( \frac{y(i)-y(i-1)}{y(1)-y(n)} \right) + 0.5 + 1
\]

The row for the \( ith \) point is:

\[
r(i) = r(i-1) + m(i)
\]

The variable separation value, \( m(i) \), is a function of the unidimensional vertical distance to the preceding point, rather than a constant amount, \( m \). It can be seen from Figure 3, that like the rank plot, the axes are labeled with actual coordinate values, and each point is plotted in a different row and column. Unlike the rank plot, the number of lines separating points is a function of both nonmetric rank and metric distance information. The Figure 3 modified-rank plot was generated by specifying:

```
PROC IDPLOT MODRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 3';
PLOT COMP2*COMP1=CAR / VR=0 HR=0;
```

The generalized-rank model is the most general of the four models, hence the GENRANK method is the most flexible method in PROC IDPLOT. In the generalized-rank model, the separation parameter \( m \), is a constant that represents the minimum number of lines that can separate observations. Assume that \( (p-1-m(n-1)) \) is nonnegative. The row for the \( ith \) point is:

\[
FLOOR \left( \frac{y(i)-y(i-1)}{y(1)-y(n)} \right) + m(i-1) + 0.5 + 1
\]

If \( m = 0 \), the formula reduces to the metric model. If \( m(n-1) = (p-1) \), the model reduces to the rank model. Otherwise both rank and unidimensional vertical distance information are used. When both separation parameters are one, the default, the generalized-rank model is very similar to the modified-rank model. In the generalized-rank model,
the distance to the first point is modeled. In the modified-rank model, the distance to the preceding point is used. In both models, the number of lines separating points is a function of both nonmetric rank and metric distance information.

Figure 4 contains a generalized-rank plot with both separation parameters equal to one. Figures 3 and 4 are nearly, but not exactly identical. This result is typical. The modified-rank model and the generalized-rank model with separations of one are two closely related ways of operationalizing the goal of putting each point on a separate line, and distributing the remaining lines so that metric distance information is reflected.

The separation parameters may have any nonnegative value in the generalized-rank model. Values less than one allow more than one point to map to a row so ID’s may be hidden, but the number of possible obfuscations is limited. Varying the separation parameters allows for the use of varying amounts of metric distance and nonmetric rank information. Varying the separation parameters varies the trade-off between presenting shape and scatter accurately, and point labels completely. The vertical separation parameter has the greatest effect on obscuration since it dictates how many points may map to a single line.

The Figure 4 generalized-rank plot was produced using default separation parameters (both one) as follows:

PROC IDPLOT GENRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 4';
PLOT COMP2. COMP1=CAR / VR=0 HR=0;

Figure 5 contains a generalized-rank plot with a horizontal separation parameter of zero and a vertical separation of 0.6. The plot was generated as follows:

PROC IDPLOT GENRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 5';
PLOT COMP2. COMP1=CAR / VREF=0 HREF=0 VSEP=0.6;

The horizontal separation parameter of zero specifies that the mapping for column coordinates should be completely metric. Stated differently, the minimum difference between the values of the argument to the FLOOR function for the ith and (i+1)th column is zero. The vertical separation of 0.6 means that the minimum difference between the values of the argument to the FLOOR function for the ith and (i+1)th column is 0.6, so no more than two points may map to the same row. The Figure 5 generalized-rank plot is more like the Figure 1 metric plot than either the Figure 4 generalized-rank plot or the Figure 3 modified-rank plot, but the Volkswagen Rabbit’s and the Chevrolet Citation’s point labels are partly obscured.

The Figure 6 contains a generalized-rank plot with a horizontal separation parameter of zero and a vertical separation of 0.25. The plot was generated as follows:

PROC IDPLOT GE PA=1 HP=45;
TITLE 'FIGURE 6';
PLOT COMP2. COMP1=CAR / VR=0 HR=0 HS=0 VS=0.25;

Figure 6 contains a generalized-rank plot with both separation parameters equal to zero. This result is typical. The modified-rank model and the generalized-rank model with separations of zero are two closely related ways of operationalizing the goal of putting each point on a separate line, and distributing the remaining lines so that metric distance information is reflected.

The Figure 6 contains a generalized-rank plot with both separation parameters equal to one. Figures 3 and 4 are nearly, but not exactly identical. This result is typical. The modified-rank model and the generalized-rank model with separations of one are two closely related ways of operationalizing the goal of putting each point on a separate line, and distributing the remaining lines so that metric distance information is reflected.

The separation parameters may have any nonnegative value in the generalized-rank model. Values less than one allow more than one point to map to a row so ID’s may be hidden, but the number of possible obfuscations is limited. Varying the separation parameters allows for the use of varying amounts of metric distance and nonmetric rank information. Varying the separation parameters varies the trade-off between presenting shape and scatter accurately, and point labels completely. The vertical separation parameter has the greatest effect on obscuration since it dictates how many points may map to a single line.

The Figure 4 generalized-rank plot was produced using default separation parameters (both one) as follows:

PROC IDPLOT GENRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 4';
PLOT COMP2. COMP1=CAR / VR=0 HR=0;

The Figure 5 contains a generalized-rank plot with a horizontal separation parameter of zero and a vertical separation of 0.6. The plot was generated as follows:

PROC IDPLOT GENRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 5';
PLOT COMP2. COMP1=CAR / VREF=0 HREF=0 VSEP=0.6;

The horizontal separation parameter of zero specifies that the mapping for column coordinates should be completely metric. Stated differently, the minimum difference between the values of the argument to the FLOOR function for the ith and (i+1)th column is zero. The vertical separation of 0.6 means that the minimum difference between the values of the argument to the FLOOR function for the ith and (i+1)th column is 0.6, so no more than two points may map to the same row. The Figure 5 generalized-rank plot is more like the Figure 1 metric plot than either the Figure 4 generalized-rank plot or the Figure 3 modified-rank plot, but the Volkswagen Rabbit’s and the Chevrolet Citation’s point labels are partly obscured.

The Figure 6 contains a generalized-rank plot with a horizontal separation parameter of zero and a vertical separation of 0.25. The plot was generated as follows:

PROC IDPLOT GE PA=1 HP=45;
TITLE 'FIGURE 6';
PLOT COMP2. COMP1=CAR / VR=0 HR=0 HS=0 VS=0.25;

Figure 6 contains a generalized-rank plot with both separation parameters equal to zero. This result is typical. The modified-rank model and the generalized-rank model with separations of zero are two closely related ways of operationalizing the goal of putting each point on a separate line, and distributing the remaining lines so that metric distance information is reflected.

The separation parameters may have any nonnegative value in the generalized-rank model. Values less than one allow more than one point to map to a row so ID’s may be hidden, but the number of possible obfuscations is limited. Varying the separation parameters allows for the use of varying amounts of metric distance and nonmetric rank information. Varying the separation parameters varies the trade-off between presenting shape and scatter accurately, and point labels completely. The vertical separation parameter has the greatest effect on obscuration since it dictates how many points may map to a single line.

The Figure 4 generalized-rank plot was produced using default separation parameters (both one) as follows:

PROC IDPLOT GENRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 4';
PLOT COMP2. COMP1=CAR / VR=0 HR=0;

The Figure 5 contains a generalized-rank plot with a horizontal separation parameter of zero and a vertical separation of 0.6. The plot was generated as follows:

PROC IDPLOT GENRANK PANELS=1 HPOS=45;
TITLE 'FIGURE 5';
PLOT COMP2. COMP1=CAR / VREF=0 HREF=0 VSEP=0.6;

The horizontal separation parameter of zero specifies that the mapping for column coordinates should be completely metric. Stated differently, the minimum difference between the values of the argument to the FLOOR function for the ith and (i+1)th column is zero. The vertical separation of 0.6 means that the minimum difference between the values of the argument to the FLOOR function for the ith and (i+1)th column is 0.6, so no more than two points may map to the same row. The Figure 5 generalized-rank plot is more like the Figure 1 metric plot than either the Figure 4 generalized-rank plot or the Figure 3 modified-rank plot, but the Volkswagen Rabbit’s and the Chevrolet Citation’s point labels are partly obscured.

The Figure 6 contains a generalized-rank plot with a horizontal separation parameter of zero and a vertical separation of 0.25. The plot was generated as follows:

PROC IDPLOT GE PA=1 HP=45;
TITLE 'FIGURE 6';
PLOT COMP2. COMP1=CAR / VR=0 HR=0 HS=0 VS=0.25;
Now, no more than four points may map to a row, so the true shape is more accurately represented, but there are more obscurations.

PROC IDPLOT FEATURES

PROC IDPLOT's syntax is similar to PROC PLOT's.

PROC IDPLOT (DATA=sasdatasetname options);  
  PLOT (list1)*(list2)=list3 (/ options);

Below is a brief summary of the PROC IDPLOT options. The first two letters of each option name is an acceptable abbreviation. Not all options are valid with all methods.

- RANK  produce a rank plot.
- MODRANK  produce a modified-rank plot.
- GENRANK  produce a generalized-rank plot.
- METRIC  produce a metric plot.
- MATCHED  produce a metric plot.
- LOCATE  list obscured metric plot points.
- PLOTSYMBOL='c' set plot symbol.
- PANELS=n number of horizontal pages.
- VOVERH=n number of rows per inch over number of columns per inch.
- IDLENGTH=n maximum id length.
- VREF=list vertical reference lines.
- HREF=list horizontal reference lines.
- HPOS=n horizontal axis print positions.
- VPOS=n vertical axis print positions.
- HTICKPOS=n horizontal print positions per tick interval.
- VTICKPOS=n vertical print positions per tick interval.
- HNTICK=n horizontal number of tick intervals.
- VNTICK=n vertical number of tick intervals.
- HINC=n horizontal tick value increments.
- VINC=n vertical tick value increments.
- HSEP=n horizontal separation parameter.
- VSEP=n vertical separation parameter.

USAGE NOTES

PROC IDPLOT ignores the page size when it constructs plots, but not when it prints them. Set a long page size (say, OPTIONS PAGESIZE=120;) to avoid page breaks in plots.

To produce all two-dimensional plots from a multidimensional configuration, specify:

PLOT (variablelist) = idvar / MATCHED;

For example the following are equivalent:

PLOT (Y1-Y6)*(X1-X6)=ID / MATCHED;

CONCLUDING REMARKS

PROC IDPLOT should be viewed as a supplement to, not a replacement for PROC PLOT. PROC IDPLOT was designed to be used in those situations where being able to label the points is most important.

PROC PLOT is more useful than PROC IDPLOT in many other situations. If you use PROC IDPLOT, and have any comments, criticisms, or suggestions, the author would be interested in hearing them.

Warren F. Kuhfeld
The Psychometric Laboratory
UNC, Davie Hall 013-A
Chapel Hill NC 27514
(919) 966-3540

REFERENCES


