Corrgrams: Exploratory Displays for Correlation Matrices

Michael FRIENDLY

Correlation and covariance matrices provide the basis for all classical multivariate techniques. Many statistical tools exist for analyzing their structure but, surprisingly, there are few techniques for exploratory visual display, and for depicting the patterns of relations among variables in such matrices *directly*, particularly when the number of variables is moderately large. This article describes a set of techniques we subsume under the name "corrgram," based on two main schemes: (a) Rendering the value of a correlation to depict its sign and magnitude. We consider some of the properties of several iconic representations, in relation to the kind of task to be performed. (b) Reordering the variables in a correlation matrix so that "similar" variables are positioned adjacently, facilitating perception. In addition, the extension of this visualization to matrices for conditional independence and partial independence is described and illustrated, and we provide an easily used SAS implementation of these methods.

KEY WORDS: Conditional independence; Effect ordering; Independence; Partial correlation; SAS; Value rendering; Variable sorting; Visual thinning; Visualization.

1. INTRODUCTION

Correlation and covariance matrices provide the basis for all classical multivariate techniques, because (together with mean vectors) they provide sufficient statistics under multivariate normal linear models. Many statistical tools exist for analyzing multivariate structure: principal component analysis, factor analysis, canonical correlation analysis, and so forth. All of these have the goal of reducing high-dimensional multivariate structure to a smaller number of dimensions, so that the relationships among the variables may be more readily apprehended.

Some visualization techniques for these dimension-reduction methods have also been developed to help reveal or explain the pattern of relations among variables: biplots (Gabriel 1971); factor pattern plots, canonical structure plots (Friendly 1991), and so forth. Dynamic graphics, including techniques such as exploratory projection-pursuit (Friedman 1987) and grand tours (Asimov 1985) can also lead to simplified views of relations among variables, in terms of linear combinations and projections. Surprisingly, there are few techniques for exploratory visual display, and for depicting the patterns of relations among variables *directly* from correlation matrices.

For a relatively small number of variables, p, ($p \leq 10$, say), the scatterplot matrix provides an excellent visual representation of the relations among variables. It shows all the data and may be considerably enhanced by the addition of linear regression lines, (loess) smoothed curves, data ellipses, and so forth. Particularly with the addition of nonparametric smoothed curves, the scatterplot matrix display can help determine if relationships are linear, or if transformations are useful for some of the variables. From here on, we assume that all such problems have been dealt with, and that all variables may be reasonably assumed to be linearly related on some possibly transformed scales.

When we go beyond a relatively small number of variables, it becomes progressively more difficult to show all the data directly. The main approach, as indicated earlier, has been the application of dimension-reduction techniques.

Here, we consider techniques to display the pattern of relations among a possibly large set of variables directly, in terms of their correlations. To do so in a comprehensible way, even for a moderately large number of variables requires some schematic visual summary—an effective visual thinning, as in the boxplot (Tukey 1977), which sacrifices detail in the middle to provide more essential information on univariate shape, center, spread, and outliers.

This article focuses on techniques to display the *pattern* of correlations in terms of their signs and magnitudes using visual thinning and correlation-based variable ordering. Some of the specific ideas and techniques we illustrate have been suggested before, and some are novel. The main contributions of this article are to integrate these methods within a coherent framework based on the principles of *correlation rendering* and *correlation ordering*, with details, comparisons, and software.

In particular, we compare a variety of visual encodings for schematic rendering of bivariate relations among quantitative variables and illustrate the perceptual differences among them for various data-analytic tasks. We also introduce a new method for arranging the variables in such displays so that the pattern of relations among variables may be more easily discerned. Finally, extensions of this framework lead directly to useful new displays for exploring conditional independence and partial independence.

For simplicity, we consider the case for p variables, $Y_1, Y_2, \ldots, Y_p$, assumed to be at least approximately multivariate normal, so that a correlation is a reasonable numerical sum-

Michael Friendly is Professor, Psychology Department, York University, Toronto, Ontario, M3J 1P3, Canada (E-mail: friendly@yorku.ca). This research was supported by the National Sciences and Engineering Research Council of Canada, Grant OGP-0138748. We thank Forrest Young for the initial stimulus to this work, John Fox for insightful comments and suggestions, and the referees and Associate Editor for helpful critical reviews.

  DOI: 10.1198/000313002533

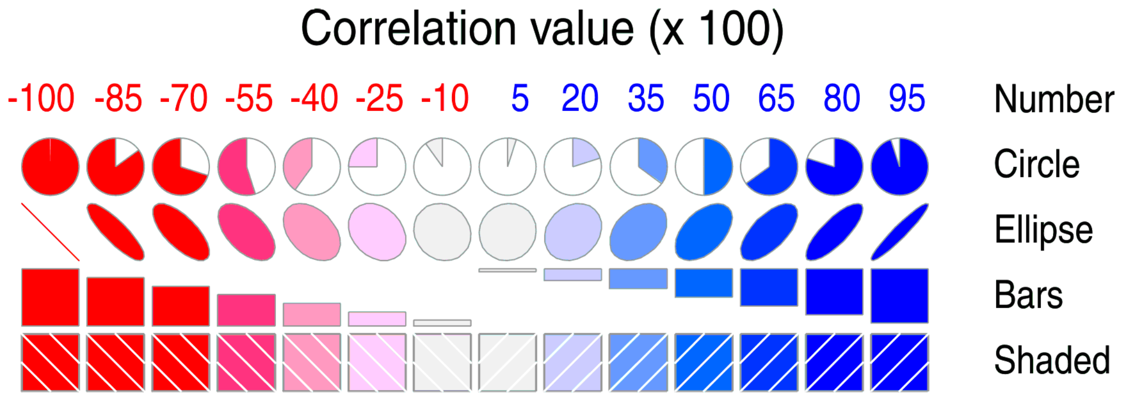

Figure 1. Some renderings for correlation values.

mary, and expressed in standardized form ($\mu_i = 0, \sigma_i = 1$), so that we may focus on a correlation matrix rather than a covariance matrix.

Section 2 describes several methods for visually encoding a correlation value to show both its sign and magnitude, with the goal of depicting the pattern of relations among variables in a potentially large matrix of correlations. Section 3 describes and illustrates a method for reordering the variables in a correlation matrix so that "similar" variables are positioned adjacently, in order to make such patterns more apparent. Section 4 extends these ideas to displays designed to show conditional and partial independence relations among variables. Section 5 describes some related methods, and Section 6 describes software implementing our procedures and some others. Section 7 presents some conclusions and questions for further work on this topic.

2. CORRELATION RENDERING

A matrix of correlations can be displayed schematically in a variety of forms: as numbers, shaded squares, bars, ellipses, or as circular "pac-man" symbols, as shown in Figure 1. These schemes all attempt to show both the sign and magnitude of the correlation value, using a color mapping of two hues in varying lightness (Cleveland 1993), where the intensity of color increases uniformly as the correlation value moves away from 0. Color (blue for positive values, red for negative values) is used to encode the sign of the correlation, but the renderings are designed so that the sign may still be discerned when reproduced in black and white.

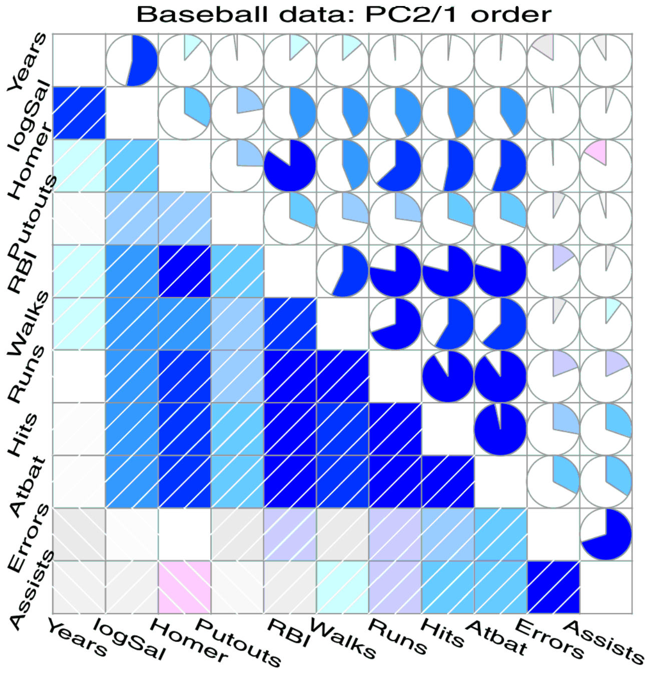

Figure 2. Corrgram for baseball data. Variables ordered by vector angles from Figure 3. Lower triangle: correlations shown by color and intensity of shading; upper: circle symbols.

In the shaded row, each cell is shaded blue or red depending on the sign of the correlation, and with the intensity of color scaled 0–100% in proportion to the magnitude of the correlation. (Such scaled colors are easily computed using RGB coding from red, $(1, 0, 0)$, through white $(1, 1, 1)$, to blue $(0, 0, 1)$. For simplicity, we ignore the nonlinearities of color reproduction and perception, but note that these are easily accommodated in the color mapping function.) White diagonal lines are added so that the direction of the correlation may still be discerned in black and white. This bipolar scale of color was chosen to leave correlations near 0 empty (white), and to make positive and negative values of equal magnitude approximately equally intensely shaded. Gray scale and other color schemes are implemented in our software (Section 6), but not illustrated here.

The bar and circular symbols also use the same scaled colors, but fill an area proportional to the absolute value of the correlation. For the bars, negative values are filled from the bottom, positive values from the top. The circles are filled clockwise for positive values, anti-clockwise for negative values. The ellipses have their eccentricity parametrically scaled to the correlation value (Murdoch and Chow 1996). Perceptually, they have the property of becoming visually less prominent as the magnitude of the correlation increases, in contrast to the other glyphs.

We use these iconic encodings to display the pattern of correlations among variables in the entire matrix, as shown in Figure 2, which depicts the matrix of correlations among 11 measures of performance and salary for 263 baseball players in the 1986 season [from the 1988 Data Expo at the ASA meetings, as corrected by Hoaglin and Velleman (1994); see http://lib.stat.cmu.edu/data-expo/1988.html]. To illustrate the differences among these encodings, we have used shading for the lower triangle, and circles for the upper triangle. The diagonal cells, which have values of 1.0 are intentionally left empty.

The interpretation for this example, and the method used to order the variables are described in Section 3.

The choice of visual representation for graphics always depends on the task to be carried out by the viewer. From Figure 1 and Figure 2 we note that it appears easiest to "read" the numerical value from the number itself, next from the circular symbols, then from the ellipses and the bars, and last for the pure shadings. (Note: The order of the pies and bars may be up for grabs, but we put our money on the much-maligned Camembert, when the purpose is to be able to say "which is more," or estimate the correlation value.) For exploratory visualization, where the task is to detect patterns of relations, and anomalies, this ordering may well be reversed—from shaded boxes as the best to numerical values as the worst.

Other forms of encoding may also be useful, or those shown here may be enhanced for certain purposes. For example, it is straightforward to add visual indications of the significance level, or of the value of a correlation required for significance. We do not consider these extensions here, because our emphasis is on exploratory display.

3. CORRELATION ORDERING

For exploratory visualization, the task of detecting patterns of relations, trends, and anomalies is made considerably easier when "similar" variables are arranged contiguously and ordered in a way that simplifies the pattern of relations among variables. This is an instance of a simple general principle, called "effect-ordered data display" (Friendly and Kwan in press) which says simply that in any data display (table or graph), unordered factors or variables should be ordered according to what we wish to show or see. This principle extends the idea of "main effect ordering" (e.g., Cleveland 1993)—sort quantitative, multiway data by means or medians—and is grounded in the perceptual ideas of similarity and grouping which stem from Gestalt psychology.

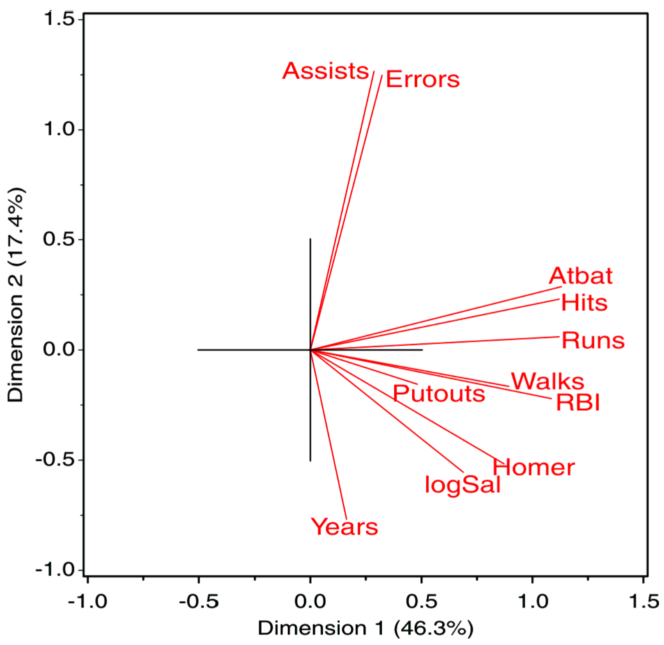

Figure 3. Eigenvector plot for baseball data. Each variable is represented by a vector whose endpoint is the coordinates of the first two eigenvectors.

Of course, for correlations, there are many ways of specifying what we mean by "similar." Several variations have interpreted this criterion in terms of a clustering, typically hierarchical, of the variables whose correlations are displayed. These procedures induce only a partial-ordering on the variables: variables within clusters are contiguous, but those within clusters at any level may be permuted in any order with the same visual interpretation. [Gruvaeus and Wainer (1972) provided a method to make the ordering of variables unique, but this method is ad hoc and not necessarily optimal.]

Here, we take a different tack and opt for a slightly stronger criterion: doing a reasonable job of placing the variables in a well-defined optimal unidimensional order. We confine ourselves here to methods based on the eigenvalues and eigenvectors of the correlation matrix, $\mathbf{R}$, or some function of it; we consider only methods based on the eigenvectors associated with the largest k eigenvalues, $k = 1, 2, 3$. Friendly and Kwan (in press) showed that this approach provides solutions for a wide range of visualization methods. Second, we do this with the hope that this approach may be more useful in some cases, and give results which should not be substantively different from those obtained by the weaker clustering interpretation.

When the structure of correlations is well described by a single, dominant dimension (as in a unidimensional scale or a simplex), ordering variables according to their positions on the first eigenvector, $\mathbf{e}_1$, of the correlation matrix, $\mathbf{R}$, will suffice. Geometrically, this implies that all variable-vectors are contained within a 90° segment of p-space, and all (or most) correlations are positive or near zero. This is not usually the case, and in general, more satisfactory solutions are obtained by ordering variables according to the angles formed by the first two (or three) eigenvectors (principal components).

For example, Figure 3 plots the first two eigenvectors of the correlation matrix among variables in the baseball data. Dimension 1 relates mostly to measures of batting performance, while Dimension 2 relates to two measures of fielding performance and to longevity in the major leagues. However, the lengths of the projections on these dimensions is determined by the adequacy (percent of variance) of the two-dimensional representation. On the other hand, the (cosines of) angles between vectors approximate the correlations between these variables, and so an ordering based on the angular positions of these vectors naturally places the most similar variables contiguously. Friendly and Kwan (in press) referred to this as "correlation ordering," a general method for arranging variables in multivariate data displays.

In Figure 2 the variables have been arranged in the angular order of the eigenvectors from Figure 3. More precisely, the order of the variables is calculated from the order of the angles, α_i,

$$\alpha_i = \begin{cases} \tan^{-1}(e_{i2}/e_{i1}) & e_{i1} > 0 \\ \tan^{-1}(e_{i2}/e_{i1}) + \pi & \text{otherwise} \end{cases}, \quad (1)$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are the $p \times 1$ eigenvectors associated with the largest two eigenvalues. This circular order is unfolded to a linear order by splitting at the largest gap between adjacent vectors. Falissard (1996) described a method for representing

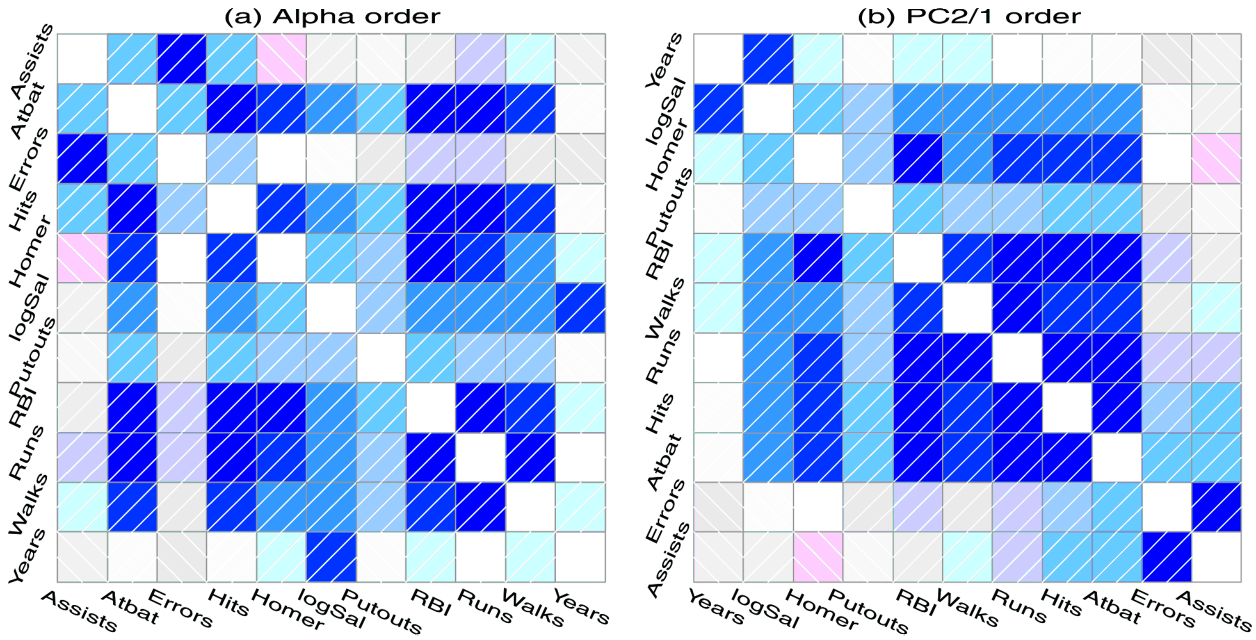

Figure 4. Corrgrams for baseball data. Each correlation is shown by color and intensity of shading. Left: variables in alphabetic order; right: variables ordered by angles of first two eigenvectors.

the variables in a correlation matrix on a unit sphere, using the first three principal components (PCs).

3.1 Examples

We continue the analysis of the baseball data and illustrate these methods with additional data on characteristics of automobile models. Both examples are illustrated in other forms in the corrgram extensions in Section 4.

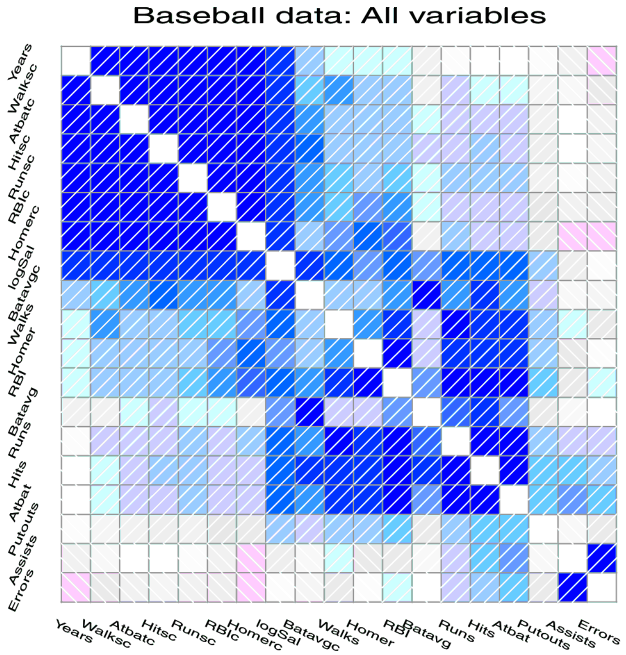

Figure 5. Corrgram for baseball data, using season and career variables.

3.1.1 Baseball Data

Figure 4 compares an arbitrary, alphabetic ordering of variables with ordering based on the angles of the first two PCs for the baseball data using the shaded encoding. In the left panel (ordered alphabetically), it is difficult to see any overall pattern of relations among these variables, despite the fact that most correlations are positive, and the relations here are fairly simple. The right panel shows clearly that (a) Assists and Errors stand out as a separate cluster; (b) RBIs, Walks, Runs, Hits and Atbats form a relatively homogeneous grouping with high positive correlations; (c) Putouts has weaker positive correlations with these; and (d) there are a few correlations which stand out as higher or lower than their neighbors. (The variable Years actually has a nonlinear relation with (log) Salary, and is better represented in a linear model as a piecewise linear function, $y_{\leq 7} = \min(\text{Years}, 7)$, which is linear up to seven years and flat thereafter. Similarly, a number of the counted variables, such as Hits, Runs, Homer, etc., are better represented on a square-root scale. These transformations do not affect the general nature of the interpretations drawn here.)

In this case, these observations could arguably be made more easily from the eigenvector display in Figure 3. For larger or more complex datasets, the corrgram may have some advantage for exploratory purposes, because it shows all the correlations, rather than just a low-dimensional summary.

The baseball dataset actually contains performance statistics for the 1986 season, and similar measures for the player's career (whose names have a `c' suffix). Figure 5 shows the corrgram for all 19 variables (including season and career batting averages, calculated from Hits and Atbats).

Here we see, among other things, that: (a) the career hitting statistics are all nearly uniformly highly positively correlated and, not surprisingly, highly correlated with Years; (b) Salary (logSal) is most highly related to the career totals (but it turns

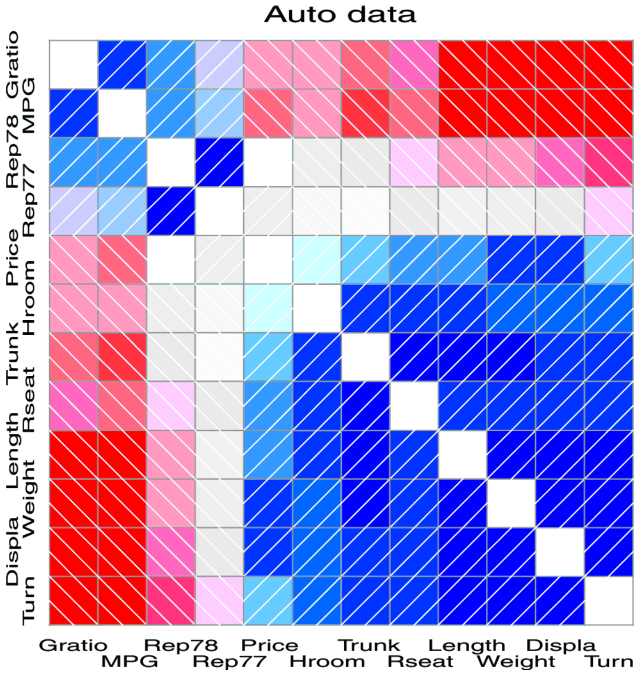

Figure 6. Corrgram for auto data, variables ordered according to Equation (1).

out that Years is an efficient proxy for most of these); (c) the season and career batting average statistics have a moderately strong correlation, but are weakly associated with most other variables; and (d) the fielding variables, Putouts, Assists, and Errors (all seasonal) form a separate group, with weak correlations to most other variables (perhaps the correlation between Assists and Errors stands out).

3.1.2 Auto Data

Figure 6 shows a corrgram of data on 74 automobile models from the 1979 model year (Chambers, Cleveland, Kleiner, and Tukey 1983, pp. 352–355) . The variables are various physical measures (gear ratio, head-room, trunk space, rear-seat, length, weight, engine displacement (Displa), turning circle diameter (Turn)) as well as Price, gas mileage (MPG), and repair-records for each of 1978 and 1979.

It is immediately clear that there are two separate groups of variables: those related to overall size and weight (which have a positive correlation with Price), and the others, which include Gratio, MPG, and the two repair record variables. Within the first group, Length, Weight, Displa, and Turn are most positively correlated; within the second group, MPG and Gratio are highly correlated, as are the two repair record variables. We also see strong negative correlations between Gratio and MPG on the one hand, and the size variables on the other.

4. EXTENSIONS

The corrgram is designed to display patterns of (linear) dependence among variables, as well as patterns of independence. This display is easily adapted to conditional or partial dependence and independence. See Friendly (1999) and Whittaker (1990) for some relations among these forms of independence for both quantitative and qualitative variables.

4.1 Conditional Independence

From Dempster (1969) and from the theory of graphical models (e.g., Whittaker 1990) it is well known that the elements of the inverse of the correlation matrix, $\mathbf{R}^{-1}$, expresses conditional dependence and independence relations in the same way that corresponding elements of $\mathbf{R}$ express ordinary (linear) dependence and independence. More precisely,

$$\begin{aligned} r_{ij} = 0 &\iff Y_i \perp Y_j \\ r^{ij} = 0 &\iff Y_i \perp Y_j \mid \text{others}, \end{aligned} \tag{2}$$

where r_{ij} is the (i, j)th element of $\mathbf{R}$, r^{ij} is the (i, j)th element of $\mathbf{R}^{-1}$, $\perp$ means "is independent of," and "others" refers to the complementary set excluding variables i and j. Thus, near 0 elements in $\mathbf{R}$ signify (bivariate, marginal) independence while near 0 elements in $\mathbf{R}^{-1}$ signify conditional independence, given all other variables in the set.

When the negative of $\mathbf{R}^{-1}$ is appropriately rescaled to have unit diagonals, the off-diagonal elements are all pairwise partial correlations, each of the form $r_{ij|\text{others}}$. Thus, a corrgram of $-\mathbf{R}^{-1}$ provides a visualization of conditional independence and dependence, just as the corrgram of $\mathbf{R}$ does for marginal independence and dependence. In a corrgram of $-\mathbf{R}^{-1}$, we should therefore pay particular attention to empty off-diagonal cells, as well as those which are strongly shaded.

Figure 7 gives an example, for the baseball data seasonal variables. For ease of comparison, the variables have been ordered in the same way as in Figure 4. We see that most of the partial correlations defined from $-\mathbf{R}^{-1}$ are small in magnitude, but there are a few notable exceptions: controlling for all other variables, there are still sizeable correlations between Years and logSal, Homers and RBIs, Hits and Atbats, and Errors and Assists.

All of these have sensible interpretations. For example, comparing Figure 7 with Figure 4, the positive relations between

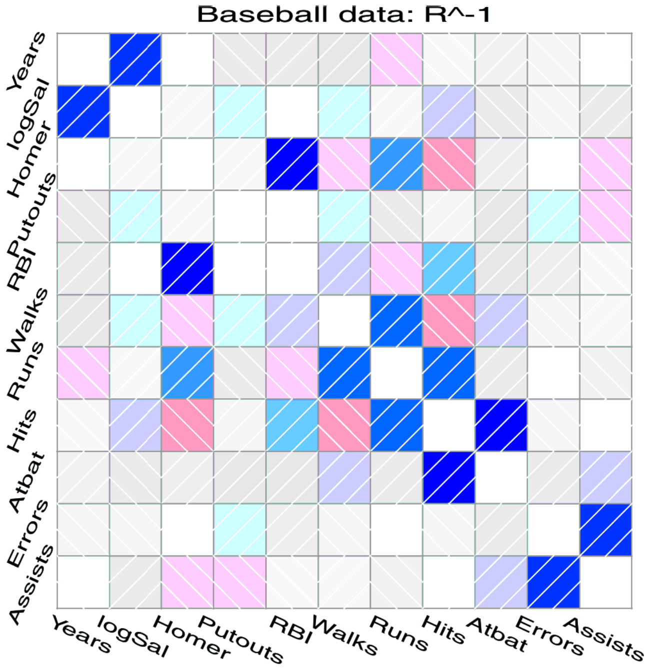

Figure 7. Conditional independence corrgram for baseball data, a visualization of pairwise partial correlations, $r_{ij|\text{others}}$.

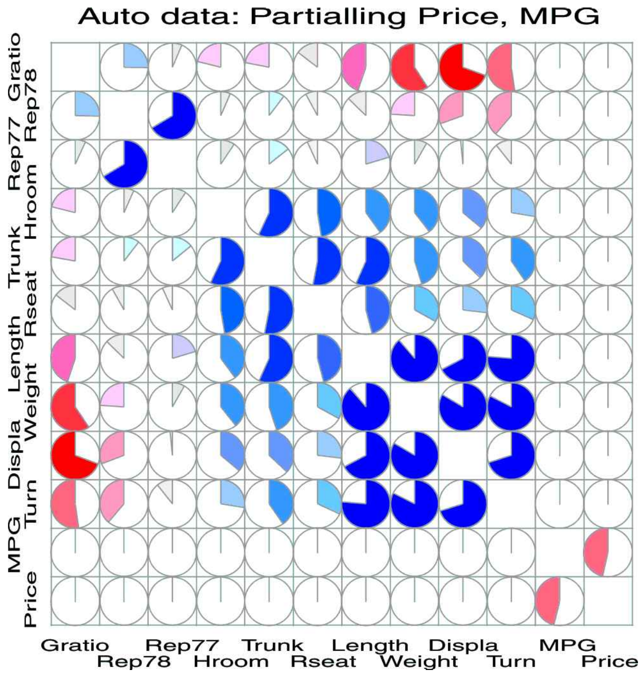

Figure 8. Partial independence corrgram for auto data, MPG and Price partialed out.

Years and logSal, and between Homers and RBIs remain when all other variables are controlled. On the other hand, although logSal was positively related (marginally) to all of the hitting performance statistics in Figure 4, we see in Figure 7 that these (conditional) relations are negligible when the other variables are taken into account. Figure 7 may therefore be interpreted to say that the relation between logSal and the hitting performance measures is largely a reflection of Years in the major leagues.

4.2 Partial Independence

Partial correlations, $R(Y|X)$, give the correlations among one

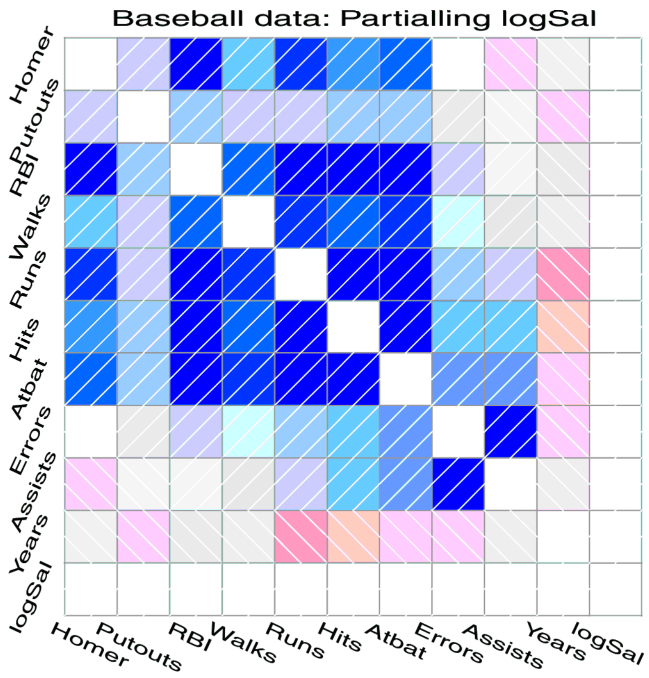

Figure 9. Partial independence corrgram for baseball data.

set of variables (Y), when another set (X) have been statistically controlled ("held constant"), adjusted for, or "partialled out." Conceptually, they differ from the conditional correlations just discussed only in that the set of X variables is fixed, rather than all the others, for each pair (Y_i, Y_j). Computationally, they may be viewed as the correlations among the residuals in the regressions of each of the Ys on all of the Xs.

Several interpretations of partial correlations are useful for visual display by corrgrams.

- If ordinary (zero-order) correlations between two Ys are large in magnitude, but the partial correlations, given one or more are near zero, then the Xs may be said "to account" for the correlation between the corresponding Ys.
- Interchanging the typical roles of X and Y, if the one or more Xs are considered responses, and the Ys explanatory, then the partial correlations $R(Y|X)$ may be interpreted as correlations among the explanatory variables "focused on" (or partialling out) their relations to the response variables (Falissard 1999).

In both cases, suppose that X is a subset of the variables to be partialled out. Then, we show a corrgram of the partitioned matrix,

$$\begin{bmatrix} \mathbf{R}_{Y|X} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{XX} \end{bmatrix}, \tag{3}$$

where $\mathbf{R}_{Y|X} = \mathbf{R}_{YY} - \mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY}$ is the matrix of partial correlations. If there is only one X, that row and column will be the representation of zeros; otherwise, the $\mathbf{R}_{XX}$ portion will portray the correlations among the Xs.

To illustrate, using the auto data, we might wish to explore the partial correlations among the remaining variables when Price and MPG are partialled out; for example, to determine whether the dependencies among the remaining variables can be accounted for by Price and gas mileage. Figure 8 shows the corrgram display, using circular encodings to better depict the numerical values of the partial correlations. Here, the $\mathbf{R}_{XX}$ portion (bottom right) shows the moderately strong negative zero-order correlation between Price and gas mileage (MPG).

The $\mathbf{R}_{Y|X}$ portion shows that, controlling for both of these, the size variables are all positively correlated, but particularly so for Length, Weight, Displacement and Turn. The remaining variables (Gratio, Rep77, and Rep78) are generally positively correlated with each other, although the two repair-record variables stand out most strongly. Between these two subsets, there is a consistent pattern for Gratio vs. the others, but somewhat weaker for the two repair-record variables.

Figure 9 illustrates the second case of a partial correlation display, showing the correlations among the (seasonal) predictors of logSal in the baseball data, focused on the salary response variable. Controlling for salary, years in the major leagues has (weak) negative correlations with all other variables. Errors and Assists are still highly correlated with each other, but weakly correlated with the batting variables. Even taking salary into account, the batting variables are still highly positively related, and again the relation between Homers and RBIs stands out against the relations of the other variables in the upper left corner.

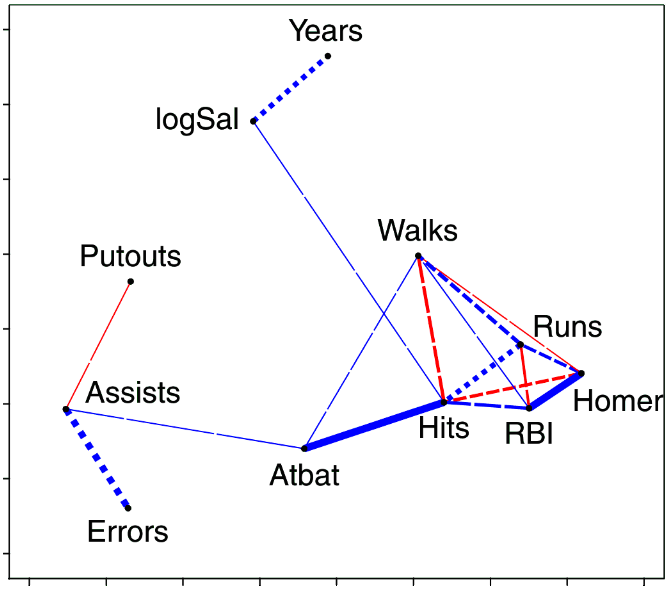

Figure 10. Conditional independence network graph for the baseball data. The locations of variables were determined by an MDS analysis of the partial correlation matrix derived from $-\mathbf{R}^{-1}$. Line style and thickness are proportional to the partial correlation.

5. RELATED METHODS

Several graphic methods, often ad hoc, for depicting the structure of matrices and rendering their values have been proposed in a variety of contexts. The following brief review attempts to relate the present methods to this other work.

5.1 Ordering

For example, Paolini and Santangelo (1991) used similar displays for visual analysis of the pattern of sparsity in the $p \times p$ coefficient matrix, $\mathbf{A}$, in large linear systems of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$, where p may be of size 10^4 or more. Permutation of the matrix rows and columns is used to search for block structure in the nonzero elements, enabling specialized solvers to find solutions for these systems far more efficiently.

In earlier work, (Hills 1969) proposed two techniques for graphical analysis of large correlation matrices: a half-normal plot of Fisher's z-transforms to identify correlation values too large to have come from zero population values, and an application of metric multidimensional scaling (MDS) to identify clusters of variables "such that members of the same group are all fairly positively correlated with each other, and behave similarly in their relations with other variables."

In the metric MDS analysis, the variables are represented as points in k-dimensional Euclidean space determined from the first k eigenvectors of the double-centered matrix with elements $(r_{ij} - \bar{r}_{i\cdot} - \bar{r}_{\cdot j} + \bar{r}_{\cdot\cdot})$. Distances between pairs of points approximate $2(1 - r_{ij})$ (to the extent that the first k eigenvalues are large), so close points represent variables with high positive correlations.

For comparison with the present methods, Figure 10 shows a network graph representation of the conditional independence relations in the baseball data depicted in Figure 7. In the network graph, the positions of the variables were derived from a nonmetric MDS analysis of the partial correlation matrix, $r_{ij|\text{others}}$ derived from $-\mathbf{R}^{-1}$, which allows a monotonic, but not necessarily linear relation between the partial correlation value and distance in the 2D spatial representation. Note that the spacing of the points, by themselves, does not lead to identification of similar clusters of variables, nor does it provide any coherent interpretation.

To show a graph representation of the conditional independencies in these data, we follow Friendly (1999) to extend the simple 0/1 graph diagrams of Whittaker (1990). In Figure 10 we have added lines between pairs of variables, for all cases where $r_{ij|\text{others}} > 0.18$, the smallest value required to make the graph connected. In this graph, line-style and thickness encode the magnitude, and color encodes the sign of the conditional correlation. Comparing Figure 10 with Figure 7, we can see the same conditional relations identified earlier: Strong positive relations between 5 and logSal, Homers and RBIs, Hits and Atbats, and Errors and Assists, when all other variables are controlled. In addition several negative conditional relations attract attention; for example, RBIs and Runs, Hits and Homers, Putouts and Assists. In this network graph, it is perhaps somewhat easier to see these relations than in the corresponding corrgram. We have found, however, that the usefulness of such graphs depends critically on the use of a (somewhat arbitrary) threshold for drawing lines, and the encoding of correlation value by line-style and thickness is often less effective than in the corrgram.

Other related techniques stem from the method of McQuitty (1968), which involves iteratively recalculating correlations among the columns of the correlation matrix itself. If $\mathbf{R}^{(0)} = \text{corr}\,(\mathbf{Y})$ is the original correlation matrix, the sequence, $\mathbf{R}^{(1)}, \mathbf{R}^{(2)}, \ldots$ is calculated as $\mathbf{R}^{(i)} = \text{corr}\,(\mathbf{R}^{(i-1)})$. McQuitty showed that this sequence often converges to a matrix whose elements are all $+1$ or -1, which allows the variables to be partitioned into two groups. Recursive application of this method to each group may then be used to generate a hierarchical clustering of the variables. This method was apparently rediscovered by Breiger, Boorman, and Arabie (1975) as the CONCOR algorithm, who applied it to proximity matrices and social network analysis.

Most recently, Chen (1996, 1999) used these ideas to develop a "generalized association plot" in which (a) the iterative procedure is continued until $\mathbf{R}^{(i)}$ becomes (nearly) rank two, (b) the eigenvectors of $\mathbf{R}^{(i)}$ have an elliptical structure, whose order is used to seriate the variables, and (c) shadings of the reordered matrix are used to display its structure.

5.2 Rendering

Finally, it is of some interest to compare the techniques presented here with other methods for schematic rendering of correlation values. Murdoch and Chow (1996) adopted a minimalist approach by using elliptical glyphs whose eccentricity is scaled to the signed correlation value, an approach which is suitable for large ($p > 25$, say) matrices. However, they eschewed the use of color and shading, and presented no general scheme for variable ordering.

The current techniques may also be compared with methods based on the scatterplot matrix and its' enhancements cited in the introduction. Fox (personal communication, 2001) suggested the combination of concentration ellipses and loess smooths as schematic visual summaries of linear and possibly nonlinear association.

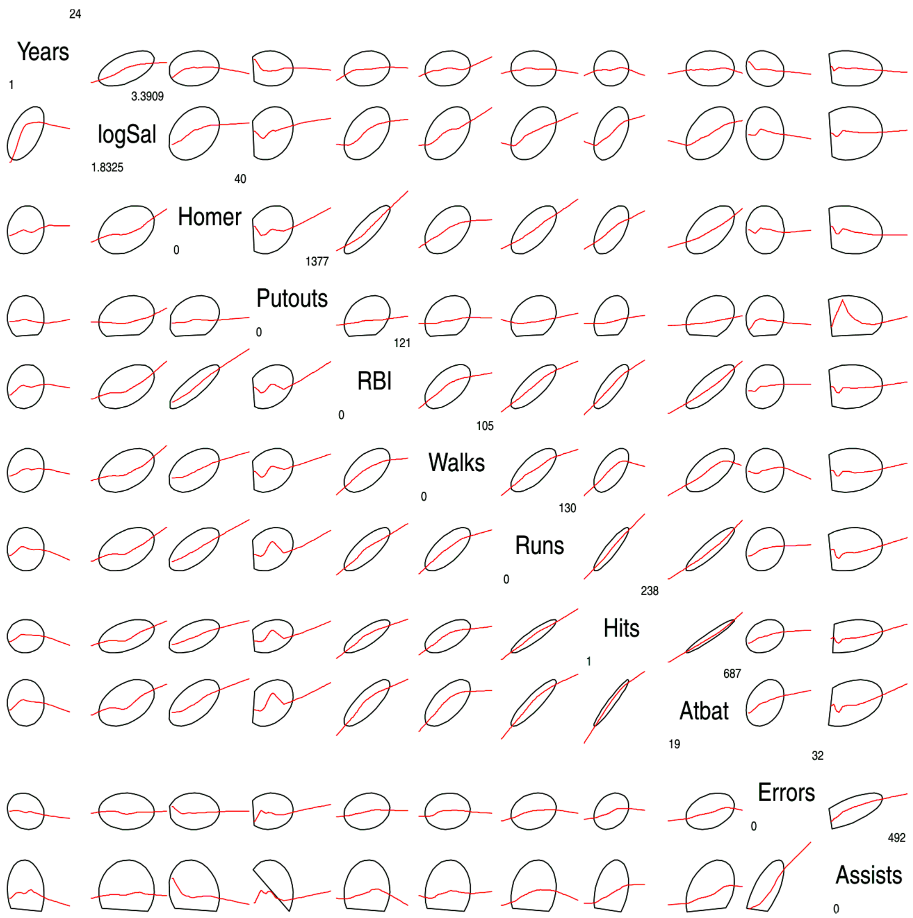

Figure 11. Schematic scatterplot matrix for baseball data. Each panel shows the bivariate 68% concentration ellipse (truncated at the data bounding box) and a loess smoothed curve.

For the baseball data, our version of such a plot is shown in Figure 11, with a "1 standard deviation" ellipse of 68% coverage centered at the means, and the variables ordered as in Figure 4 (right). To highlight the patterns of association, all extraneous ink has been suppressed—points, plot frames, tick marks, and so on. For $p = 11$ variables, this display shows far more detail than the corrgrams presented here, but it also suggests that some of the relation we have assumed to be linear are actually nonlinear. While this form of rendering may be better for some tasks, it would be difficult to accommodate many more variables, and it is clearly more difficult to see the overall pattern of relations among variables in Figure 11 than in the corresponding corrgram in Figure 4.

6. SOFTWARE

The corrgrams shown here are all drawn by a general SAS macro program, `corrgram.sas`, described in http://www.math.yorku.ca/SCS/sasmac/corrgram.html, from which the source code may be downloaded. The program has a large variety of options and is easily used. For example, Figure 2, using shaded encodings below the diagonal, and circular encodings above, is produced by the macro call,

```
title ´Baseball data: PC2/1 order´;
%corrgram(data=baseball,
  var=logSal   5 Homer Runs Hits RBI
  Atbat Walks Putouts Assists Errors,
  fill=S E C);
```

The analogous partial independence corrgram in Figure 9 is obtained by adding the keyword option `partial=logSal` and removing the `fill=S E C` option. The program requires SAS/IML and SAS/GRAPH in addition to the basic SAS System.

A program for MATLAB, `corrmap.m` was developed by Barry Wise, and is available at http://www.eigenvector.com/MATLAB/cormap.html. This program uses a version of the k-nearest neighbor algorithm to reorder the variables. It appears to encode correlation values by a "pseudo-color map" which ranges from white for $r = +1$ through yellow and red, to black for $r = -1$. This is not a good choice, but other color mappings may be readily used, and the source code is available.

Finally, SYSTAT provides a variety of matrix clustering algorithms, clustering both rows and columns. When applied to a correlation matrix, the rows and columns are permuted according to the cluster structure, and the correlation values are depicted by colored squares, but using an apparently arbitrary and fixed color scheme.

7. CONCLUSIONS

In a wide sense there is not much that is *absolutely* novel here—various methods for visually depicting correlation matrices have been proposed (or just used, e.g., Dobkins, Gunther, and Peterzell 2000, fig. 2), and various schemes for reordering variables in such matrices have also been suggested. Yet, surprisingly, there has been no published work we have found treating these methods in any coherent way.

We *can* claim to have presented a more general and comprehensive account of the possibilities than has appeared previously. We have also (a) suggested a new scheme for ordering variables in such displays, (b) extended the idea of correlation mapping to more general concepts of dependence and independence, and (c) illustrated (we hope convincingly) why they might be useful. We also provide a flexible implementation of these ideas (Section 6) with which others can work, and perhaps extend.

In particular, the details of the various rendering techniques suggested here bear further study: continuously scaled versus classed colors, accounting for the nonlinearity of color reproduction and perception, circles or bars versus shaded boxes, and so forth. It was not until we had tried several alternatives that the differences among them became apparent.

For large matrices, these techniques scale relatively well, but the results are most often successful when the level of detail in the rendering is minimized (e.g., using shading, elliptical glyphs, etc.). Labeling of the variables, important for interpretation, also becomes more difficult, but this is easily solved by flexible font orientation and scaling.

Finally, we note that these graphical techniques are applicable to the wider class of symmetric matrices, including distance and proximity matrices. In addition, the method of correlation-based variable ordering described here has been shown (Friendly and Kwan in press) to facilitate perception of relations in other multivariate data displays (e.g., parallel coordinate plots, star plots).

[Received February 2001. Revised August 2002.]

REFERENCES

Asimov, D. (1985), "Grand Tour," *SIAM Journal of Scientific and Statistical Computing*, 6, 128–143.

Breiger, R. L., Boorman, A. S., and Arabie, P. (1975), "An Algorithm for Clustering Relational Data With Applications to Social Network Analysis and Comparison With Multidimensional Scaling," *Journal of Mathematical Psychology*, 12, 328–383.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth.

Chen, C. H. (1996), "The Properties and Applications of the Convergence of Correlation Matrices," in *Proceedings of the Statistical Computing Section*, Alexandria, VA: American Statistical Association, pp. 49–54.

——— (1999), "Extensions of Generalized Association Plots (GAP)," in *Proceedings of the Statistical Computing Section*, Alexandria, VA: American Statistical Association, pp. 111–116.

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press.

Dempster, A. P. (1969), *Elements of Continuous Multivariate Analysis*, Reading, MA: Addison-Wesley.

Dobkins, K. R., Gunther, K. L., and Peterzell, D. H. (2000), "What Covariance Mechanisms Underlie Green/Red Equiluminance, Luminance Contrast Sensitivity and Chromatic (Green/Red) Contrast Sensitivity?" *Vision Research*, 40, 613–628.

Falissard, B. (1996), "A Spherical Representation of a Correlation Matrix," *Journal of Classification*, 13, 267–280.

——— (1999), "Focused Principal Component Analysis: Looking at a Correlation Matrix With a Particular Interest in a Given Variable," *Journal of Computational and Graphical Statistics*, 8, 906–912.

Friedman, J. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.

Friendly, M. (1991), *SAS System for Statistical Graphics*, Cary, NC: SAS Institute.

——— (1999), "Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data," *Journal of Computational and Graphical Statistics*, 8, 373–395.

Friendly, M., and Kwan, E. (in press), "Effect Ordering for Data Displays," *Computational Statistics and Data Analysis*, 37.

Gabriel, K. R. (1971), "The Biplot Graphic Display of Matrices With Application to Principal Components Analysis," *Biometrics*, 58, 453–467.

Gruvaeus, G. and Wainer, H. (1972), "Two Additions to Hierarchical Cluster Analysis," *The British Journal of Mathematical and Statistical Psychology*, 25, 200–206.

Hills, M. (1969), "On Looking at Large Correlation Matrices," *Biometrika*, 56, 249–253.

Hoaglin, D. C., and Velleman, P. F. (1994), "A Critical Look at Some Analyses of Major League Baseball Salaries," *The American Statistician*, 49, 277–285.

McQuitty, L. L. (1968), "Multiple Clusters, Types, and Dimensions From Iterative Iintercolumnar Correlational Analysis," *Multivariate Behavioral Research*, 3, 465–477.

Murdoch, D. J., and Chow, E. D. (1996), "A Graphical Display of Large Correlation Matrices," *The American Statistician*, 50, 178–180.

Paolini, G. V., and Santangelo, P. (1991), "An Interactive Graphic Tool to Plot the Structure of Large Sparce Matrices," *IBM Journal of Research and Development*, 35, 231–237.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison Wesley.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.