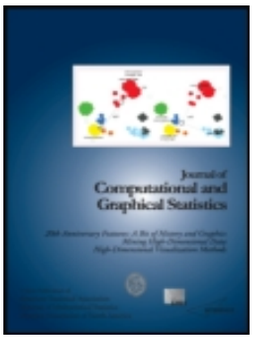

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Determining and Depicting Relationships Among Components in High-Dimensional Variable Selection

Peter Hall & Hugh Miller

To cite this article: Peter Hall & Hugh Miller (2011) Determining and Depicting Relationships Among Components in High-Dimensional Variable Selection, Journal of Computational and Graphical Statistics, 20:4, 988-1006, DOI: 10.1198/jcgs.2011.09179

To link to this article: https://doi.org/10.1198/jcgs.2011.09179

Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Determining and Depicting Relationships Among Components in High-Dimensional Variable Selection

Peter HALL and Hugh MILLER

Many modern treatments of high-dimensional datasets involve reducing the initial collection of features to a much smaller set, from which a predictive model may be built. However, strong relationships between the remaining variables can limit the parsimony or even the predictive performance of such a model. We propose a semi-automatic approach using generalized correlation to detect and quantify these relationships, as well as exploring ways to represent this information graphically. The method can detect both symmetric and asymmetric relationships, as well as nonlinear patterns. Its utility is demonstrated on a range of real and simulated datasets. Supplemental material for performing the real-data analyses in this article is available online.

Key Words: Classification; Generalized correlation; High-dimensional statistics; Massive dimension reduction; Regression; Variable interaction; Variable redundancy.

1. INTRODUCTION

A standard statistical approach to solving variable-selection problems involves determining, for a wide data vector X, a relatively small number of variables on which correct classification or prediction depends. The identified variables might represent a small number of genes, typically between a few and a few tens, out of thousands or tens of thousands of variables in X. In some instances the influential genes might be selected primarily because they each represent, in different ways, the same phenomenon. Some of the variables could, in the presence of others, be essentially redundant, in that they might be deleted without appreciably changing the performance of a classifier or predictor.

In theoretical terms the potential for this phenomenon is clear. For example, if two variables are highly correlated, where correlation is measured conventionally, then one of them can often be deleted without the final result being greatly influenced. More generally,

Peter Hall is Professor and Hugh Miller is Honorary Associate (E-mail: *H.Miller@ms.unimelb.edu.au*), Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.

Journal of Computational and Graphical Statistics, Volume 20, Number 4, Pages 988–1006
DOI: 10.1198/jcgs.2011.09179

if one of the variables is a function of the other, then it might be possible to drop either of the variables without affecting the overall performance of a classifier or predictor. These occurrences are not limited to very high-dimensional variable selection problems; even when the number of variables is only moderate there is potential for highly influential variables to be closely related, and potentially redundant.

However, while avoiding redundancy is important, there are still more compelling motivations for understanding the relationships among the "significant" variables in X. Indeed, in genomic problems there are good scientific reasons for wishing to understand the joint behavior of different, influential variables. For example, it is important to comprehend the manner in which the variables identified by variable selection operate together to exert their influence. We might ask whether the expression levels of two particular genes tend to increase and decrease together, or whether there is a more subtle and complex connection between them. (Gene expression levels indicate the extent to which a gene is "switched on," i.e., its activity level.) Can we quantify and explain this complexity, thereby gaining a better understanding of the problem than just the fact that the r selected variables seem to be more "significant" than others for classification or prediction? In particular, insight into the way in which gene expression levels vary jointly, to produce an overall significant effect, can be of greater value than simply knowing which set of genes is the most significant.

This relates to another context where variable relationships can be quite significant. Suppose that the r observed variables are functions of s unobserved, or latent, variables $U = \{U_1, \ldots, U_s\}$. The observed variables that share common underlying latent variables will probably have observable dependencies which can be viewed to give insight into the underlying behaviors. Further, nonlinear relationships between W and X will cause nonlinear relationships in the observed variables, and allow some variables to give superior insights into latent effects. Again, means of computing and visualizing these relationships can be powerful, and simulation examples of latent effect models are explored in Section 4.

In this article we suggest a simple way of answering questions such as these. We propose techniques based on generalized correlation (Grindea and Postelnicu 1977) and show how to apply them to explore relationships among different variables. Generalized correlation is particularly effective for capturing nonlinear functional relationships between variables, although other patterns, such as clustering of observations, are not necessarily detected. Here we extend and apply the methodology of Grindea and Postelnicu (1977) to multidimensional settings. We introduce graphical methods that enable us to access this sort of information quickly and reliably, thereby guiding the experimenter toward further pertinent questions that could be asked of the data.

For example, Figure 1 shows two ways of depicting the relative strengths of pairwise relationships for genes that appear to influence the development of acute leukemia. The left panel represents relationship strengths in terms of gray shade or color temperature, while the right panel achieves a similar end using a graph where both the darkness and the width of edges reflect the strengths of relationships.

Real-data examples where these issues are important arise in a diverse range of situations. We illustrate this point using the real-data example referred to previously, drawn from

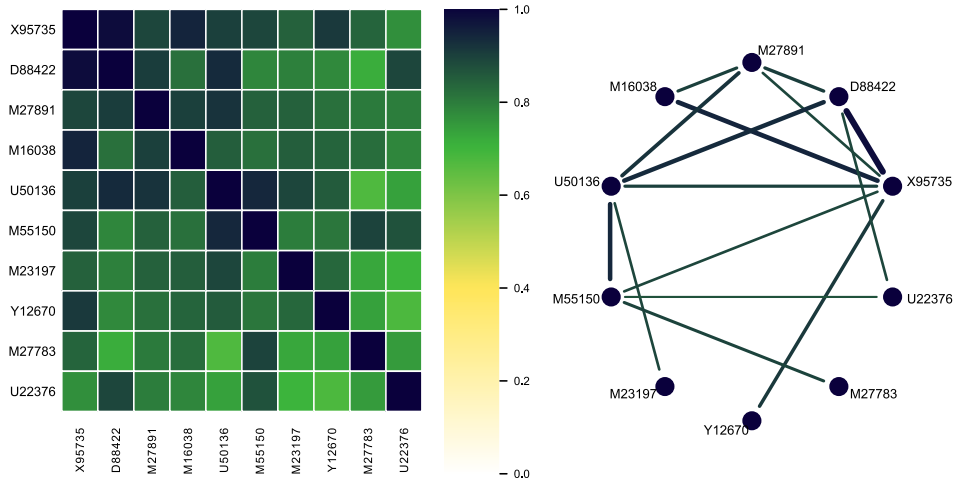

Figure 1. Associative potential for AML/ALL genes, with threshold equal to 0.85 in the network plot. The online version of this figure is in color.

the work of Golub et al. (1999), and the hepatitis survival dataset studied by Diaconis and Efron (1983) and Cestnik, Konenenko, and Bratko (1987). The latter dataset is available online from the UCI machine learning database, while the former is also widely available. In each case we shall apply our methodology and discuss the insight that can quickly be gained.

The remainder of the article is structured as follows. Section 2 introduces methodology for finding generalized correlation estimates and graphical representations for correlations, and discusses the implementation and scalability of our techniques. Section 3 compares different aspects of our approach with other, related ideas in the literature. In particular we include a discussion of partial correlation, which has gained popularity for genomic data analysis. Finally, Section 4 provides related numerical work, including the real-data analysis mentioned earlier.

2. METHODOLOGY

2.1 Generalized Correlation for Measuring Strength of Association and the Potential for Prediction

Assume that a variable selection method has narrowed the number of "significant" variables from a very large value, equal to the length p of X, to just r. For simplicity we designate these by $X^{(1)}, \ldots, X^{(r)}$. We define the generalized correlation between $X^{(j_1)}$ and $X^{(j_2)}$ to be

$$\rho(j_1, j_2) = \sup_{g_1, g_2 \in \mathcal{G}} \operatorname{corr}\bigl\{g_1\bigl(X^{(j_1)}\bigr), g_2\bigl(X^{(j_2)}\bigr)\bigr\}, \tag{2.1}$$

where $\mathcal{G}$ represents a class of functions (e.g., the class of polynomials of a given degree), and $\operatorname{corr}(U, V)$ denotes the standard correlation coefficient between random variables U

and V. We interpret $\rho(j_1, j_2)$ as the extent of association between $X^{(j_1)}$ and $X^{(j_2)}$, in the sense of generalized correlation. Note that, if $\mathcal{G}$ denotes the class of linear functions, then $\rho(j_1, j_2) = |\operatorname{corr}(X^{(j_1)}, X^{(j_2)})|$.

We also define

$$\rho_{j_1 j_2} = \sup_{g \in \mathcal{G}} \operatorname{corr}\{X^{(j_1)}, g(X^{(j_2)})\}, \tag{2.2}$$

which can be interpreted as a measure of the potential for predicting $X^{(j_1)}$ from a function of $X^{(j_2)}$, when that function comes from $\mathcal{G}$. In particular, if $\mathcal{G}$ is closed under addition of scalars and under scalar multiplication, then $\rho_{j_1 j_2} = 1$ if and only if $X^{(j_1)} = g(X^{(j_2)})$ with probability 1, for some $g \in \mathcal{G}$; and $\rho_{j_1 j_2} = 0$ if $X^{(j_1)}$ and $X^{(j_2)}$ are statistically independent. Note that, if $\mathcal{G}$ is the class of linear functions, then $\rho(j_1, j_2) = \rho_{j_1 j_2} = \rho_{j_2 j_1}$.

The predictive correlation $\rho_{j_1 j_2}$, although not $\rho(j_1, j_2)$, has connections to a particular methodology proposed by Hall and Miller (2009) for variable selection. However, $\rho_{j_1 j_2}$ and $\rho(j_1, j_2)$ can be used to explore relationships among variables in very general settings, for example, when $X^{(1)}, \ldots, X^{(r)}$ are selected using a conventional linear model-based method such as the lasso (Tibshirani 1996; Chen and Donoho 1998) or the Dantzig selector (Candes and Tao 2007), or when $X^{(1)}, \ldots, X^{(r)}$ are identified in terms of their leverage for classification (e.g., Hall, Titterington, and Xue 2009) rather than via more conventional variable selection.

The use of the word "predictive" to describe (2.2) may suggest that causal relationships are being detected, particularly when $\rho_{j_1 j_2}$ is strong but the converse relationship is weak. However, the words "prediction" and "predictive" are widely used in statistics without there being any implied notion of causality, and of course predictive relationships are generally only suggestive of causality, rather than a convincing demonstration of it.

The generalized correlation scores suggested by (2.1) and (2.2) above are susceptible to outliers in noisy data situations, in much the same way that standard correlation is. In such circumstances it may be appropriate to use a robustified version of generalized correlation, although details are not pursued here. It is also worth comparing the use of generalized correlations with other measures of robust correlation, such as rank correlation. Such measures can be useful in noisy data situations such as the ones considered here, but are limited by requiring monotonic relationships between variables. Generalized correlation has the potential to detect non-monotonic effects and is thus preferred in circumstances where such patterns are likely.

2.2 ESTIMATORS OF $\rho(j_1, j_2)$ AND $\rho_{j_1 j_2}$

To simplify discussion in this section we shall assume that the initial p-vector X has already been reduced to a much shorter vector of length r, using a variable selection method such as those related to prediction or classification. However, in an abuse of notation we shall refer to the shorter vector as X, and in particular we shall suppose that we observe data vectors $X_i = (X_i^{(1)}, \ldots, X_i^{(r)})$ for $1 \le i \le n$, where each X_i is distributed as $X = (X^{(1)}, \ldots, X^{(r)})$. Define $\bar{X}^{(j)} = n^{-1} \sum_i X_i^{(j)}$ and $\bar{X}^{(j)}(g) = n^{-1} \sum_i g(X_i^{(j)})$. Estimators

of $\rho_{j_1 j_2}$ and $\rho(j_1, j_2)$ are given respectively by

$$\hat{\rho}(j_1, j_2) = \sup_{g_1, g_2 \in \mathcal{G}} \frac{\sum_i \{g_1(X_i^{(j_1)}) - \bar{X}^{(j_1)}(g_1)\}\{g_2(X_i^{(j_2)}) - \bar{X}^{(j_2)}(g_2)\}}{[\sum_i \{g_1(X_i^{(j_1)}) - \bar{X}^{(j_1)}(g_1)\}^2 \sum_i \{g_2(X_i^{(j_2)}) - \bar{X}^{(j_2)}(g_2)\}^2]^{1/2}}, \tag{2.3}$$

$$\hat{\rho}_{j_1 j_2} = \sup_{g \in \mathcal{G}} \frac{\sum_i (X_i^{(j_1)} - \bar{X}^{(j_1)})\{g(X_i^{(j_2)}) - \bar{X}^{(j_2)}(g)\}}{[\sum_i (X_i^{(j_1)} - \bar{X}^{(j_1)})^2 \sum_i \{g(X_i^{(j_2)}) - \bar{X}^{(j_2)}(g)\}^2]^{1/2}}. \tag{2.4}$$

Compare (2.1) and (2.2). If the class $\mathcal{G}$ is determined by only a finite number of parameters (e.g., as in the case where $\mathcal{G}$ is the set of all polynomials of given degree), then generally, under mild additional assumptions (for instance, moment conditions in the polynomial example), the estimators $\hat{\rho}(j_1, j_2)$ and $\hat{\rho}_{j_1 j_2}$ are root-n consistent for $\rho(j_1, j_2)$ and $\rho_{j_1 j_2}$, respectively. (Since both $\hat{\rho}(j_1, j_2)$ and $\hat{\rho}_{j_1 j_2}$ can be written as conventional correlation coefficients, then a derivation of root-n consistency is standard. It uses the fact that variances and covariances are root-n consistent estimators of the respective population quantities.)

2.3 GRAPHICAL METHODS FOR DEPICTING $\hat{\rho}(j_1, j_2)$ AND $\hat{\rho}_{j_1 j_2}$

There are several ways of depicting graphically the values of $\hat{\rho}(j_1, j_2)$ and $\hat{\rho}_{j_1 j_2}$. We describe two of them here, illustrated in Figures 1 and 2 using the Leukemia example discussed in Section 4. First, we suggest representing $\hat{\rho}(j_1, j_2)$ and $\hat{\rho}_{j_1 j_2}$ in terms of the darkness of a gray shade, or the hue of a color in a spectrum, using a square matrix, with each square split triangularly to contain both $\hat{\rho}_{j_1 j_2}$ and $\hat{\rho}_{j_2 j_1}$. In our plots we have adopted a range of hues from dark blue (for strong relationships), through green and finally light yellow fading to white for the weakest scores. Specifically, construct an $r \times r$ array of square boxes, and color the lower left triangle of box (j_1, j_2) to reflect the value of $\hat{\rho}_{j_1 j_2}$, and upper right to denote $\hat{\rho}_{j_2 j_1}$, where j_1 and j_2 are indicated on the vertical and horizontal axes, respectively. Of course in this depiction squares (j_1, j_2) and (j_2, j_1) contain the

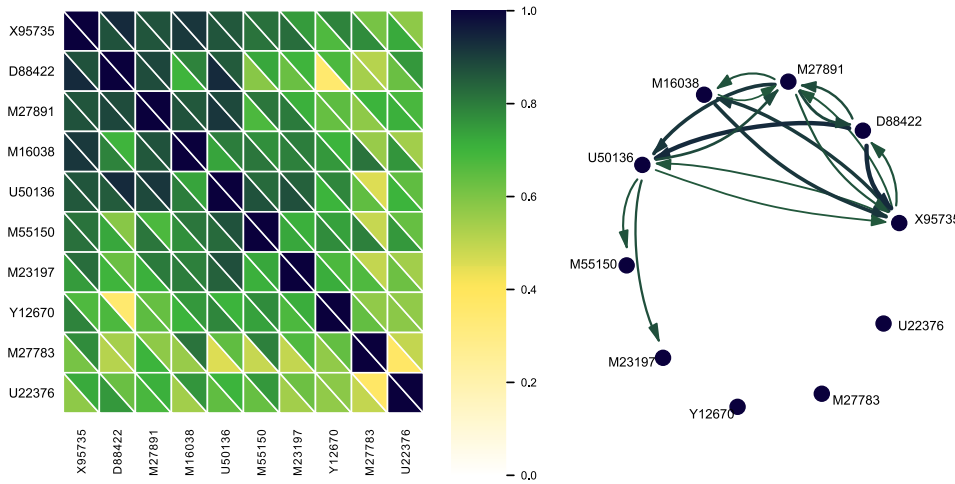

Figure 2. Predictiveness potential for AML/ALL genes, with threshold equal to 0.85 in the network plot. The online version of this figure is in color.

same information, but we feel the triangular split allows asymmetries in relationships to be spotted easily.

In this depiction the boxes down the main diagonal would be black, if using gray shade to represent the potential for prediction, or dark blue, if using color for that purpose. (Of course, $\hat{\rho}_{jj} = 1$ for each j.) A representation of $\hat{\rho}(j_1, j_2)$ would be similar, except that this quantity is symmetric and so does not require the triangular splitting of matrix squares nor require entries both above and below the main diagonal. However, it is helpful to be able to view the relationships between one variable and all the others in a single row, or single column, and so we suggest retaining all the boxes both above and below the diagonal.

A second way of depicting $\hat{\rho}_{j_1 j_2}$ graphically is to place r points in the plane, numbered from 1 to r, and link points j_1 and j_2 by an arrow leading from j_2 to j_1. In the resulting diagram the value of $\hat{\rho}_{j_1 j_2}$ can be represented by the thickness of the arrow as well as the darkness of the gray shade or the color of the arrow. Relatively weak predictive potential can be ignored, for example by confining attention to pairs (j_1, j_2) for which $\hat{\rho}_{j_1 j_2}$ exceeds a given threshold. Strength of association can be represented similarly, on a separate diagram. In this instance, in recognition of the symmetry of $\hat{\rho}(j_1, j_2)$, the arrows should either be double-ended, or simply omitted, as in our plots. We have elected to space the points equally around a circle. While other choices are possible, this arrangement enforces a sense of a priori equality among variables, as well as reducing the chances of coincident lines.

The choice of a threshold can of course have great impact on both the appearance of the network plot and the resulting conclusions. While our principle for choosing a threshold is balancing information and interpretability, we believe that in practical situations the analyst would experiment with a range of thresholds and thus draw conclusions from more than one plot. Reflecting this, we have included a rudimentary Java applet in the online supplemental material, showing how Figure 2 varies with choice of threshold.

For both the matrix and network representations it is helpful to order the variables strategically so that variables that are very similar are close. This reduces clutter as well as aiding interpretability. A simple and natural way to do this is to perform hierarchical clustering on the variables, using either $1 - \hat{\rho}(j_1, j_2)$ or $1 - \max(\hat{\rho}_{j_1 j_2}, \hat{\rho}_{j_2 j_1})$ to represent the distance between variable pairs. We have done this for all our real-data examples, using Ward agglomeration.

There has been a significant amount of research exploring ways to plot matrices of numbers of the general type considered here. We discuss other possible approaches in Section 3. Also, we make a comment on grayscale printing; displaying standard correlation matrices in grayscale can be problematic, since it is usual to have darker shades for both strong negative and strong positive scores. However, in the case of generalized correlation the score is always positive, so this issue is avoided.

2.4 GRAPHING PREDICTIVE RELATIONSHIPS

Provided sample size, n, is not too small, it is possible to go beyond the simple numerical descriptor $\hat{\rho}_{j_1 j_2}$ when assessing the potential that $X^{(j_2)}$ has for predicting $X^{(j_1)}$. For

example, we can investigate a simple regression model,

$$X^{(j_1)} = g\bigl(X^{(j_2)}\bigr) + \text{error}, \tag{2.5}$$

where $g = g_{j_1 j_2}$ is estimated nonparametrically rather than through being constrained to lie in a predetermined function class $\mathcal{G}$. If we observe data vectors X_i distributed as $X = (X^{(1)}, \ldots, X^{(r)})$ (see Section 2.2), then we can estimate $g_{j_1 j_2}$ using standard methods, for instance, employing local-linear regression with a bandwidth chosen by cross-validation, and construct a lattice, or set of scatterplots, of graphs of function estimates. Examples will be given in Section 3. The concept of a trellis plot has its roots in work of Chambers and Hastie (1992) and Becker, Cleveland, and Shyu (1996).

While this method has the potential to provide a relatively high amount of information in visual form, in genomic applications it can be seriously inhibited by the small sample sizes that commonly occur in practice, or by the difficulty that a reader has absorbing, and placing into context, all the information that arises when r is relatively large. In such cases the cruder representations discussed in Section 2.3 can be more useful.

2.5 Computation and Scalability

For some very general choices of $\mathcal{G}$, optimization of expressions (2.3) and (2.4) is computationally challenging. However, for a wide class of choices of $\mathcal{G}$, where transformations are spanned by a finite set of basis functions, the solution is actually fast and easy to implement. Further, in the case of the predictive form it is even analytic. Let us first consider this case, represented by (2.4). Suppose that $g(x) = \sum_{1 \le k \le K} \beta_k h_k(x)$, where the h_k are a chosen set of (possibly nonlinear) transformations, for instance, a set of polynomial terms or a set of natural cubic splines. Then it is possible to show that the optimization in (2.4) is equivalent to finding the least squares solution minimizing

$$\sum_{i=1}^{n} \biggl[X_i^{(j_1)} - \biggl\{ \sum_{k=1}^{K} \beta_k h_k\bigl(X_i^{(j_2)}\bigr) \biggr\} \biggr]^2 .$$

[A formal statement of this result, with assumptions, is given in theorem 1 of the article by Hall and Miller (2009).] The estimated value of $g(X_i^{(j_2)})$ can then be computed from parameter estimates of the β_k arising from the least squares fit, and the correlation score can then be calculated.

In the context of associative potential estimation, defined by (2.3), suppose that each function arises from some finite basis, so that $g_1(x) = \sum_{1 \le k \le K_1} \beta_k h_{1k}(x)$ and $g_2(x) = \sum_{1 \le k \le K_2} \gamma_k h_{2k}(x)$. Then the solutions to the generalized correlation expression are equivalent to the solutions arising from minimizing

$$\sum_{i=1}^{n} \biggl[\biggl\{ \sum_{k=1}^{K_1} \beta_k h_{1k}\bigl(X_i^{(j_1)}\bigr) \biggr\} - \biggl\{ \sum_{k=1}^{K_2} \gamma_k h_{2k}\bigl(X_i^{(j_2)}\bigr) \biggr\} \biggr]^2 ,$$

subject to scale constraints to avoid redundancies when all parameters equal zero. Forcing the empirical variances of $g_1(X^{(j_1)})$ and $g_2(X^{(j_2)})$ to equal 1 will achieve this. This minimization does not have an analytic solution. However, conditional on the β_k the γ_k

can be estimated via least squares, and vice versa. This motivates an iterative algorithm where we alternate between the two sets of parameters until convergence is achieved. That yields a simple algorithm, similar in spirit to the backfitting algorithm for generalized additive models (Hastie and Tibshirani 1990, chap. 5) that is guaranteed to converge, generally very quickly. Thus, both predictive and associative generalized correlation may be efficiently estimated.

Since these scores can be estimated quickly, the estimation of generalized correlation matrices scales well with dimension. As an example, we created a synthetic dataset of "microarray size," with $n = 100$, $r = 1000$ and having its elements all standard normal with a true correlation of 0.5 for each pair of off-diagonal variables. We then estimated both types of generalized correlation matrices using our loosely optimized R code, with the class of nonlinear transformations being natural cubic splines with three interior knots (thus there were five terms in each basis expansion). On a typical desktop computer with an Intel Core 2 Duo E6550 processor the predictive matrix estimation took 6 and 70 minutes for the predictive and associative forms, respectively. This appears reasonable considering that almost a million optimizations are performed. In such situations visual representation is still possible, although the circular layout should be replaced by a more traditional network plot such as those in the article by Peng et al. (2009). Although such high-dimensional calculations are possible, in the present article we have focused on low- to moderate-dimensional situations, either by examining a problem that is genuinely low-dimensional, or by choosing a small subset of variables in higher dimensional situations. The main reason for doing this is that our use of generalized correlations aims to give a highly interpretable view of variable relationships, which becomes increasingly difficult in higher dimensions.

3. COMPARISONS

3.1 MATRIX PLOTS

The idea of matrix plots with varying intensities indicating strength has substantial precedent in the literature. The earliest examples date back to Bertin (1967). Hartigan (1972) introduced variable clustering. Chapter 15 of the book by Chen, Härdle, and Unwin (2008) contains numerous examples and extended discussion of these plot types, and many statistical packages are equipped to produce similar plots. Also of note are the plots in the article by Friendly (2002), which not only allow visualization of correlation matrices but also yield useful variations of the scatterplot matrix described in Section 2.4.

We have not found a precedent for the triangular splitting of matrix elements to display asymmetric relationships, although it does appear a natural choice for the circumstance.

3.2 NETWORK PLOTS

As with matrix plots, network graphs are well represented in the literature. Bertin (1999) provided some examples of network plots and historical references, while modern statistical software such as SAS and Himmeli (*www.finndiane.fi/software/himmeli/*) allows correlations to be represented as networks. Network representations have become particularly

relevant in genomics. See, for instance, the article by Zhu et al. (2005) and the review by Aittokallio and Schwikowski (2006).

A possible alternative to spacing variables out equally in a network plot is to use location to convey distance/closeness in the generalized correlation sense. Multidimensional scaling (MDS) is a popular way to achieve this, with visualizations dating back to the work of Hills (1969) and Kruskal and Wish (1978); the latter text is popular. As mentioned earlier, we have found that line thickness and color give good clarity, although MDS remains a valid alternative. Additionally, a modern perspective on automatic visualization can be found in the article of Wills and Wilkinson (2008), which provides a good contrast to the current work.

3.3 PARTIAL CORRELATION

An alternative approach to standard or generalized correlations for detecting variable relationships is using partial correlation. The partial correlation between two variables is the (standard) correlation between the two with the effect of a set of other variables removed. Applying partial correlation to the situation described in Section 2, the partial correlation between $X^{(j_1)}$ and $X^{(j_2)}$ is the correlation of the residuals of these variables when each is linearly regressed on $X^{-(j_1,j_2)}$, the set of variables excluding the j_1th and j_2th. This may be denoted by $\rho_{j_1,j_2\cdot-(j_1,j_2)}$. The estimation of partial correlation is closely related to that of estimating the inverse covariance matrix, and dates at least from the work of Dempster (1972). Recent work applying partial correlation to high-dimensional settings includes that of Meinshausen and Bühlmann (2006) and Peng et al. (2009).

The key benefit of using partial correlation is that causal relationships are extracted. For instance, if two variables are highly correlated due only to their dependence on a third controlling variable, then this correlation will be ignored when partial correlations are taken, leaving only the relationships with the controlling variable. This allows a network of variables to be generated with controlling variables clearly visible.

While partial correlation is a very useful tool in visualizing variable relationships, it does have some drawbacks, which prompts the use of alternative approaches such as generalized correlation in some contexts. The first relates to linearity. Partial correlation slightly outperforms generalized correlation in settings where the main relationships are linear (and where, therefore, standard correlation generally outperforms both methods). The explanation for this is clear—partial correlation is still, at its core, based on standard linear correlation, and does not "spend" information in the sample to estimate nonlinear behavior. Conversely, when relationships are nonlinear, then, as might be expected, generalized correlation performs much better than either standard or partial correlation. It would be possible to define a form of generalized partial correlation which accounted for nonlinear patterns; we leave this as an area for future research.

A second issue is that of errors in variables. Even in linear cases, partial correlation is rather seriously affected by errors in variables, as the following small theoretical example demonstrates. Suppose that we have three observed variables $X^{(1)}, X^{(2)}, X^{(3)}$ and that a "true" underlying variable $W^{(1)}$ exists such that:

- $X^{(1)}$ is an attempt to measure $W^{(1)}$, but may have some error in it. Thus let $\text{corr}(W^{(1)}, X^{(1)}) = \rho_1$.
- $W^{(1)}$ controls $X^{(2)}$ and $X^{(3)}$, that is, $X^{(j)} = W^{(1)} + \epsilon_j$ for $j = 2, 3$ and some error ϵ_j. For simplicity assume that $\text{corr}(W^{(1)}, X^{(2)}) = \text{corr}(W^{(1)}, X^{(3)}) = \rho_2$.

If there is no error-in-variable for $X^{(1)}$, then $\rho_1 = 1$ and the theoretical partial correlation between $X^{(2)}$ and $X^{(3)}$ with respect to $X^{(1)}$ is zero ($\rho_{23\cdot 1} = 0$). This is the desired result when the nonzero partial correlations relate each of $X^{(2)}$ and $X^{(3)}$ to $X^{(1)}$, but $X^{(2)}$ and $X^{(3)}$ are not themselves related. However, if $\rho_1 < 1$, then

$$\rho_{23\cdot 1} = 1 - \frac{1 - \rho_2^2}{1 - \rho_1^2 \rho_2^2},$$

which implies that a nonzero partial correlation between $X^{(2)}$ and $X^{(3)}$ is detected. Observe that this value does not disappear as $n \to \infty$; it is intrinsic to the setup.

A third issue relates to the existence of "clusters" of strongly related variables in the data. Suppose that there are d variables, all with pairwise (standard) correlation ρ. This is a situation where there is no controlling variable, either because it is hidden or because the variables are equally controlling. In this case it is possible to show that the pairwise partial correlation is $\rho/(1 + d\rho)$. This means that if d is larger than the range three to five, approximately, the relationships between these variables will be obscured when partial correlations are taken, and noise may prevent some of these genuine relationships from being seen.

4. NUMERICAL EXAMPLES

4.1 REAL-DATA EXAMPLES

First we study the Leukemia genetic dataset first analyzed by Golub et al. (1999). The aim was to use microarray evidence to distinguish between two types of acute leukemia, AML and ALL. There were $p = 7129$ variables and $n = 38$ observations in the training data (27 ALL and 11 AML), and 34 observations in a separate test dataset (20 ALL and 14 AML). Each variable corresponded to a particular gene, and the value corresponds to an observed expression level, analogous to the level of activity in that gene, compared to some baseline. The majority of approaches to microarray problems seek to first reduce the number of influential genes to a manageable set, and then use this set to draw conclusions and make predictions. See, for example, the works of Golub et al. (1999), Tibshirani et al. (2002), Fan and Lv (2008), and Hall and Miller (2009). It is useful to understand how the genes produced by the selection process relate to one another.

We took the top ten genes determined by Hall and Miller (2009); this was the smallest set of genes that still yielded strong predictive performance. Note that eight of these genes were also in Golub's original set of 50. Associative and predictive potential were calculated, with results plotted in Figures 1 and 2 above. We let the class of functions $\mathcal{G}$, in (2.1) and (2.2), be the set of all natural cubic splines with knots at data quartiles. This enables

our definition of "generalized correlation" to capture a simple wiggle or turning point in the middle of an otherwise monotone function. In this respect generalized correlation is more appropriate than classes of linear or quadratic functions.

The most interesting feature of Figure 1 is that the associations are generally very high. It can be seen quickly from the matrix diagram that the first variable is quite strongly associated with all the others, and, on the other hand, that the third variable is generally weakly related. Similar results are obtained in the case of prediction; see Figure 2. These results suggest that, for example, the first variable, X95735, might perform as well on its own as it does together with the variables that are associated with it, or which predict it. In particular, the variables that are good predictors of X95735 might be removed without predictive performance being appreciably affected. We shall explore this possibility four paragraphs below.

The arrow diagram in Figure 1 uses a threshold of 0.85 to exclude a reasonable number of pairs. In fact, the lowest association is 0.67 (the relationship between the genes M27783 and U50136), which is still strikingly high. This issue is discussed further below. The large number of close associations between X95735 and the others, and the small number of associations involving the M27783, are also reflected in the arrow diagram. However, that diagram displays relatively little information about strength, although it gives a clearer picture than does the matrix diagram of the pattern of linkages.

Figure 2 presents the asymmetrical predictive relationships $\rho_{j_1 j_2}$. Generally features are similar to the associative results. For example, there is clearly a relatively large number of predictive relationships involving X95735, and a relatively small number involving M27783. The threshold for the arrow diagram has been kept at 0.85 for direct comparison to Figure 1, although the diagram is noticeably more cluttered. Interesting features are those that display one-way predictiveness. For instance, the variable U50136 is strongly predictive for D88422 but not vice versa. The left panel of Figure 3 contains a scaled plot of the expression level for gene D88422 against that for U50136, and it is evident why this one-directional relationship exists; D88422 has increased expression only when U50136 has a significant expression. This means that, while U50136 is a good predictor for D88422, it is difficult to use D88422 to separate low to medium expression levels of U50136, as shown in the right panel. It can be observed that this relationship is quite strong, and would not be fully captured by traditional linear correlation measures.

As described in Section 2.4, it is possible to investigate, using regression models, the ability to predict variables from others. Figure 4 shows the scatterplot matrix for each pairwise combination of standardized variables. The fitted line is a natural cubic spline fit, as used for estimating generalized correlation. The scatterplots give further insight into previous results. For instance, the lighter third row seen in the matrix representations in Figures 1 and 2 is largely explained by an outlier in the third variable (M27891) that is not reflected in any of the other gene expression levels. Also note the high number of nonlinear fits; the activation pattern introduced above appears to be a relatively common feature.

As mentioned earlier, it is possible to use the idea of variable predictiveness to eliminate redundant variables. For example, suppose we use the arrow plot in Figure 2 to remove variables, by eliminating the variables whose dots are pointed to. If we decide to keep the

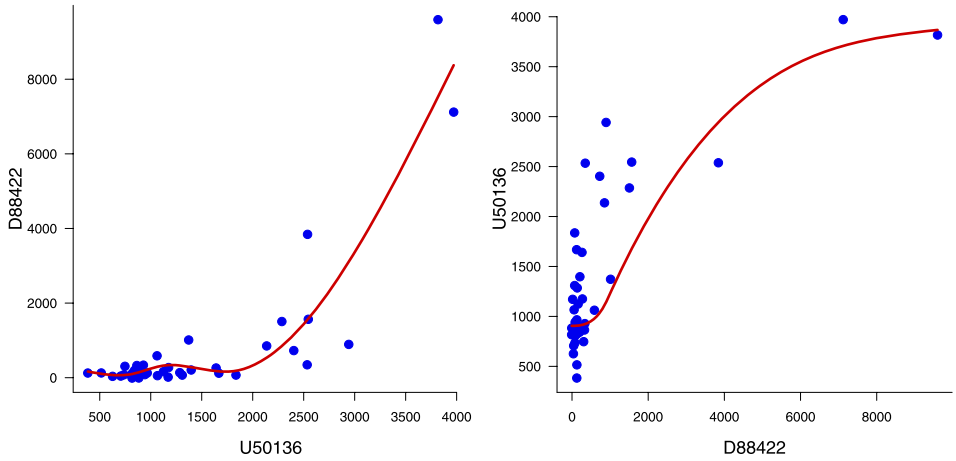

Figure 3. Scatterplots of U50136 against D88422 with natural cubic spline fits. The online version of this figure is in color.

gene X95735, judged the most important, but then remove D88422, M27891, M16083, and U50136, this would remove all arrows from the predictiveness diagram. A basic random forest model to predict leukemia type using all ten genes has one out-of-bag error (analogous to a cross-validated error) on the learn set and two errors on the test set. However, a random forest excluding the selected variables generates one out-of-bag error on the learn set and also one error on the test set. Although the improvement may partly be due to random noise in the data, it certainly suggests that performance of the six-gene model is at least as good as that of the ten-gene model. The analysis has successfully removed unnecessary variables in a very natural way.

Some further comment should be made regarding the high levels of correlation found in the AML/ALL example. Much of this is due to how the genes were originally selected; if two genes are powerful in separating the two cancer groups, then they are both likely to correlate closely with the response and hence with each other. This phenomenon is easily reproducible in theoretical examples. For example, suppose we have $p = 10{,}000$ uncorrelated standard normal random variables and a response variable which is also an independent standard normal variable. When we take $n = 30$ and choose the five variables that best correlate with the response, there is an average correlation (taking absolute values) of 0.41 between these variables, much higher than the 0.15 observed over all variables. This demonstrates that selected subsets tend to overstate the amount of correlation. However, this inflation effect is unlikely to account for all the correlation observed in real-data examples. There is also the alternative explanation that the gene responses are correlated for biological reasons; for example, the gene pathways may overlap. One possible extension of our methodology would be to net out the dependence on the response, which would deflate some of this heavy dependence.

Most other gene selections show similar high correlations between genes. For instance, consider the set of fifty genes selected by Golub et al. (1999). Twenty-five of these genes had high expression levels for AML, and for these the average predictive potential, $\rho_{j_1 j_2}$,

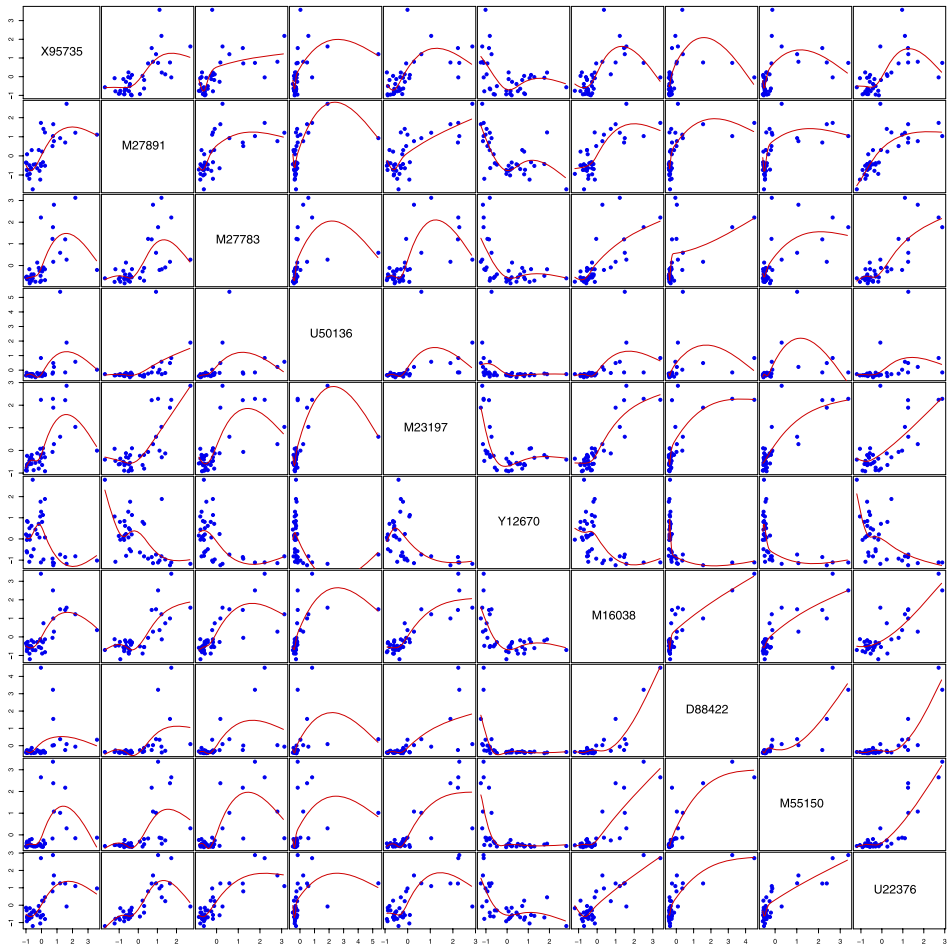

Figure 4. Scatterplot matrix of leukemia genes with local linear fit. The online version of this figure is in color.

was 0.60. The other twenty-five genes had high expressions for ALL and these had an average predictive potential of 0.58. The implications of these relationships for prediction are significant. For instance, a scheme where each gene "votes" for the Leukemia type has fewer effective votes than the number of genes due to the high correlations.

As foreshadowed earlier, we next examine the hepatitis survival dataset analyzed by Diaconis and Efron (1983) and Cestnik, Konenenko, and Bratko (1987). Their approaches used logistic regression on the 19 predictors to estimate the survival or non-survival of 155 hepatitis patients. The methodology was critiqued by Breiman (2001), who pointed to dependencies between variables as hindering the logistic models and reducing predictive accuracy. A semi-automatic means of detecting these dependencies, such as the methods introduced in this article, is therefore important in this situation.

The hepatitis data analysis is slightly more delicate compared to the previous one, due to the presence of missing values as well as the presence of two-level categorical variables (specifically, variables 2 to 13 as well as variable 19 were categorical). The first of these

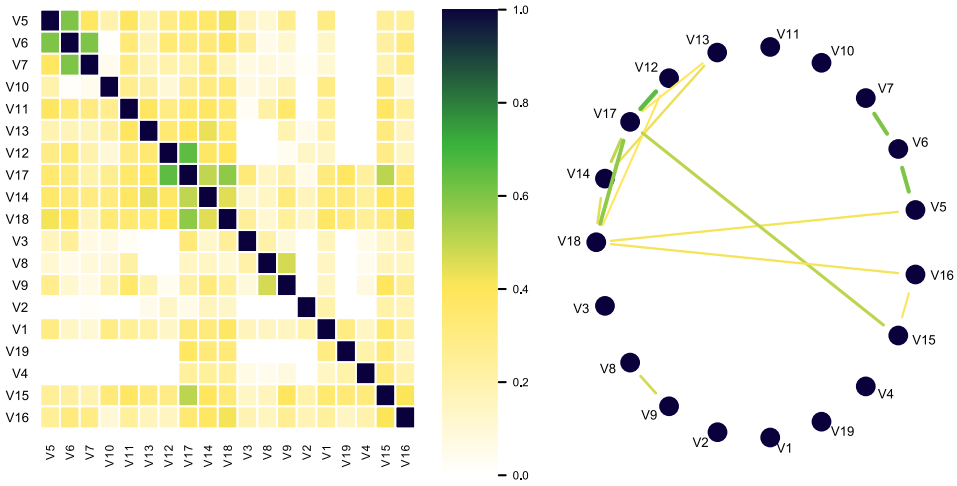

Figure 5. Association plots for hepatitis data. The online version of this figure is in color.

issues was addressed by considering only observations that contained no missing values for each pairwise choice of variables. To address the second, each categorical variable was coded as a 0–1 variable and was assessed numerically for correlations. With a slight abuse of notation, $\mathcal{G}$ was defined to contain only the identity function for the categorical variables, while the other variables contained the space formed by natural cubic splines with three interior knots. Results are presented in Figure 5. Aside from the strong associations of the sixth variable with both the fifth and the seventh, the dominant feature in the figure is the heavy relationship between the twelfth and seventeenth variables. This is exactly the relationship detected by Breiman (2001) and cited as the key problem with prior attempts at logistic regression. Variable 17 is a much stronger predictor of variable 12 than the reverse ($\rho_{1217} = 0.65$ while $\rho_{1712} = 0.55$), so a case could be made for excluding variable 12 from a final model. The asymmetry of this relationship is not surprising as variable 17 is continuous and variable 12 is binary, meaning that the former can carry more information.

4.2 THEORETICAL EXAMPLES BASED ON LATENT VARIABLE MODELS

Here we give examples of latent variable models, where each variable in X is derived from an underlying latent data vector of length s. For ease of exposition we use linear relationships, so that

$$X_{ij} = \sum_{k=1}^{s} \phi_k U_{ik}.$$

It makes sense to use the absolute value of standard correlations in this context, or equivalently to restrict $\mathcal{G}$ to consist of all linear transformations. This also implies that the predictive and associative scores coincide, so we shall refer simply to association. In the examples below, and in graphical depictions of the strength of association or prediction, each line linking two variables indicates that the points have latent variables in common. In a

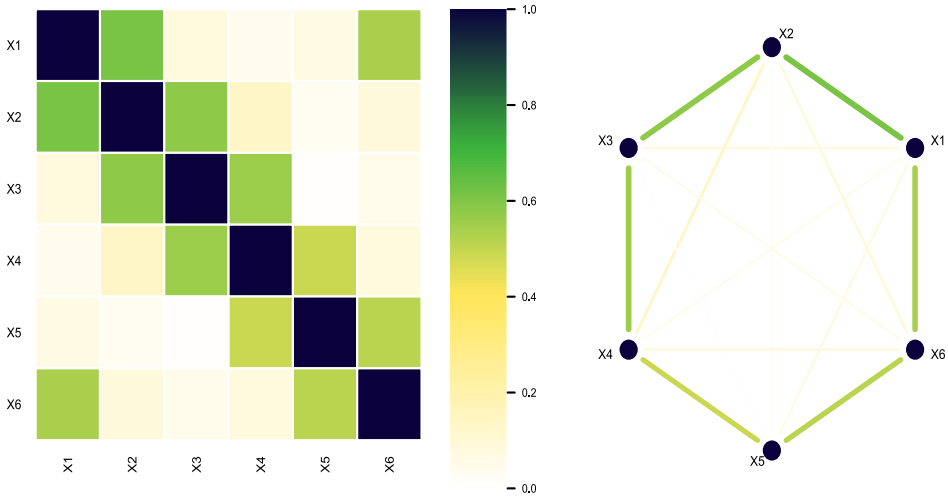

Figure 6. Association plots for periodic case $(r, t) = (6, 2)$. The online version of this figure is in color.

matrix depiction, the darkness that corresponds to a pair of variables is proportional to the number of latent variables that the variables share.

First we consider a "chain" structure, where each variable is associated with just two neighbors in a simple closed circuit. Let $X^{(j)} = U_j + U_{j+1}$ for $1 \le j \le r-1$, and $X^{(r)} = U_r + U_1$, where the variables U_j are independent and identically distributed with finite variance. Then $\mathrm{corr}(X^{(j)}, X^{(j+1)}) = \frac{1}{2}$ for $1 \le j \le r-1$, $\mathrm{corr}(X^{(1)}, X^{(r)}) = \frac{1}{2}$, and $\mathrm{corr}(X^{(j_1)}, X^{(j_2)}) = 0$ for all distinct pairs $\{j_1, j_2\}$ not covered by this scheme. More generally we could take

$$X^{(j)} = U_j + \cdots + U_{j+s} \tag{4.1}$$

for all $j \in [1, r-s]$, and either complete the sequence (for values $j = r-s+1, \ldots, r$) in a periodic fashion, as suggested in the previous, simpler example, or define the sequence nonperiodically, interpreting (4.1) as holding for all $j \in [1, r]$. Figures 6 and 7 show, in the case $(r, s) = (6, 2)$, graphical depictions in periodic and nonperiodic cases. Normal random variables with mean zero and standard deviation 1 were used for the U_j with $n = 100$. In the periodic case, the pairwise association is evidenced by both the matrix diagram, with dark shades on the diagonals closest to the main diagonal, as well as the arrow diagram, which has dark arrows between neighboring points. Here the arrow diagram includes all arrows, to give a sense of the relative associations. In the aperiodic case a threshold of 0.2 was used and only associations above this level were included in the diagram. Again the pairwise associations are clear.

As a slightly more complex example, the periodic case when $(r, s) = (8, 3)$ is presented, using the same assumptions for U_j and n. There the theoretical associations are $2/3$ for pairs (X_j, X_{j+1}) (interpreted cyclically), $1/3$ for pairs (X_j, X_{j+2}), and zero for other pairs. Figure 8 displays the results. The arrow diagram used a threshold of 0.2. The strong associations are clearly visible, while the weaker associations (those with a theoretical association of $1/3$) have been partially obscured by noise; not all of these associations

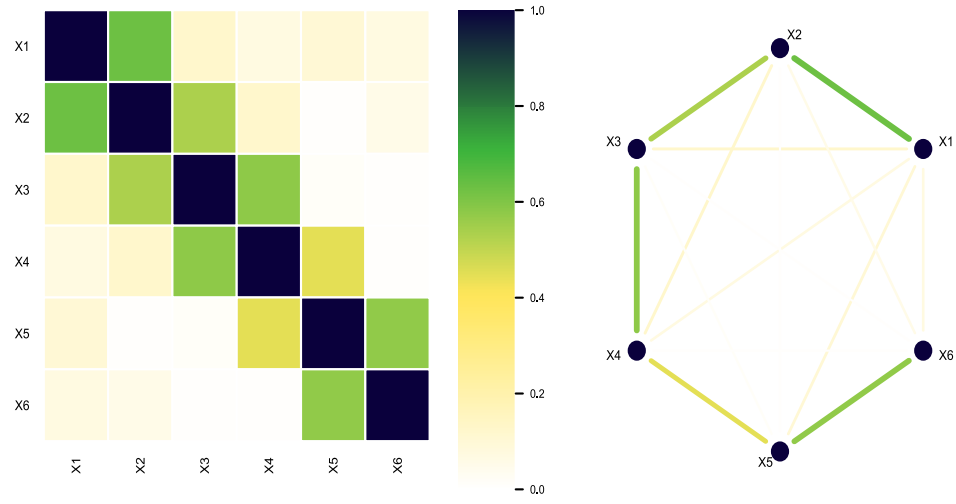

Figure 7. Association plots for aperiodic case $(r, t) = (6, 2)$. The online version of this figure is in color.

appear in the arrow diagram and some false associations appear in preference. While increasing n would better resolve the associations, the example demonstrates how noise can hide weaker associations. This effect aside, it should be clear that the graphical depictions are effective in highlighting variable relationships that point to latent effects.

4.3 COMPARISONS WITH PARTIAL CORRELATION

We give examples of some of the issues associated with partial correlation in Section 3.3, starting with the issue of errors in variables. In small samples, and even when the extent of errors is as small as 10% and the relationships are linear, the partial correlation deteriorates to the extent that it loses all of its potential advantages over generalized

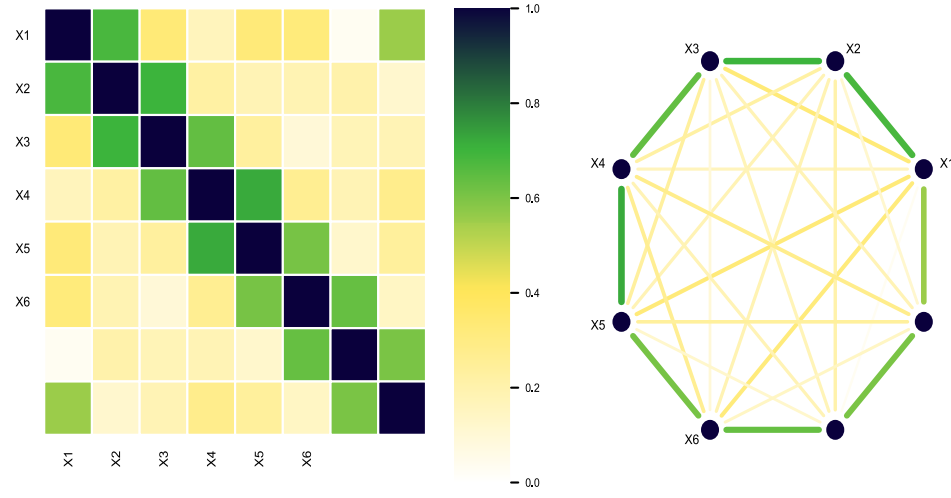

Figure 8. Association plots for periodic case $(r, s) = (8, 3)$. The online version of this figure is in color.

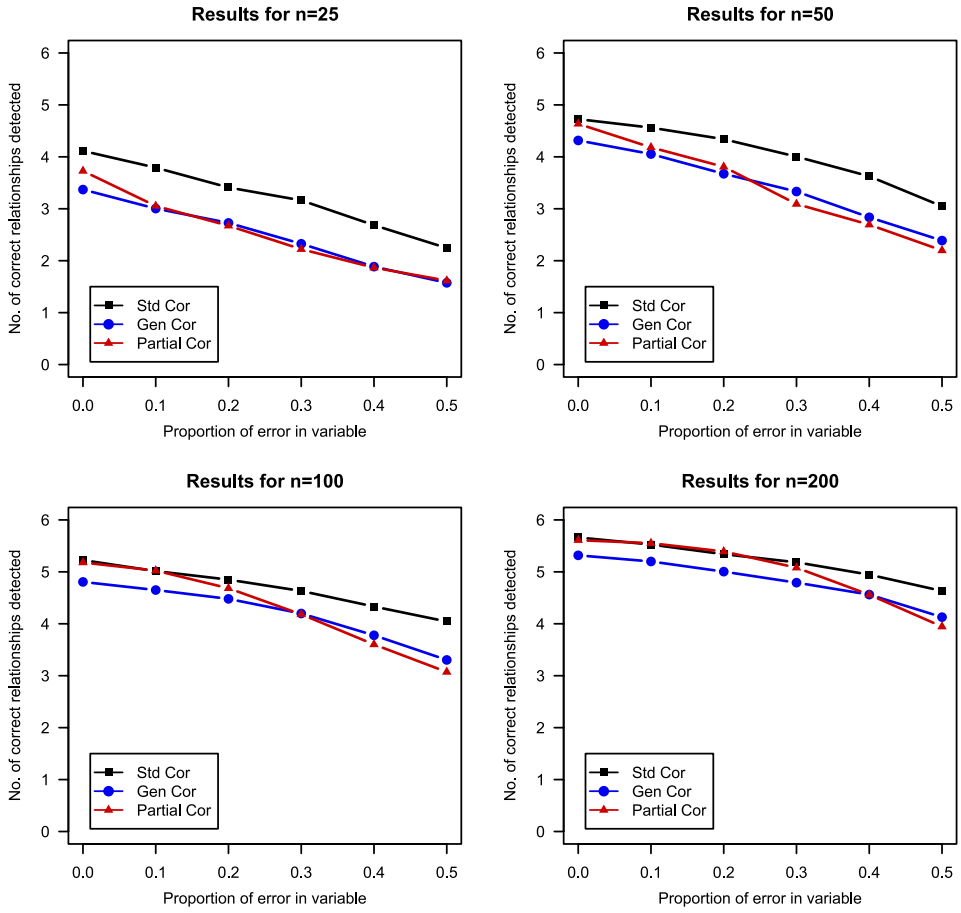

Figure 9. Comparison of relationship detection power for standard, generalized, and partial correlations in the presence of errors in variables. The online version of this figure is in color.

correlation. To show this, a simulated dataset containing ten standard normal variables was created. All (linear) correlations were set to zero, except for six relationships of respective strengths 0.9, 0.6, 0.6, 0.5, 0.3, and 0.2. A degree of Gaussian noise was also added to each observation. Figure 9 shows the average number of these six relationships detected by the three correlation methods for various sample sizes and errors in variables. Since the relationships are linear, generalized correlation would be expected to be the weakest method, but it is clear from the graphs that performance of partial correlation degrades quickly in the presence of errors.

Figure 10 demonstrates how the existence of larger clusters can degrade partial correlation performance; it shows, for $\rho = 0.8$ and various d, the proportion of genuine partial correlation relationships that can be detected above the average "noise" level. Sample size was 50, and variables were normally distributed. It is clear in this example that as the cluster size approaches 10, the size of a partial correlation is almost indistinguishable from random noise. This effect is detrimental, particularly when clusters of this type are of significance.

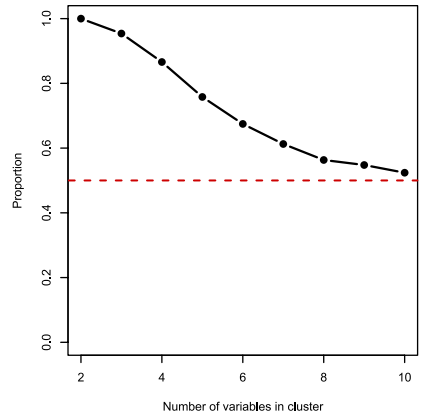

Figure 10. Proportion of partial correlations above average random noise level. The online version of this figure is in color.

5. DISCUSSION

We have shown how, after preliminary dimension reduction or variable ranking, graphical methods and generalized correlation can be combined to provide insight into relationships among variables. This leads to new insights into the ways in which genes can interact when influencing disease. The potential advantages of our approach, relative to one based on partial correlation, have also been pointed out. A marriage of general and partial correlation, combined with the graphical techniques developed in this article, also has potential and is a fruitful avenue for further research.

SUPPLEMENTARY MATERIALS

Dataset: Dataset containing a subset of the Leukemia dataset suitable for analysis. (Leuk.Rdata)

Plot functions: R-code functions to create plots of the type presented throughout this article. (plot_functions.r)

Manuscript plots: R-code for producing plots related to the Leukemia and Hepatitis datasets used in this article. (Manuscript_plots.r)

Java applet: A simple Java applet allowing the network plotting threshold to be varied for the Leukemia dataset. (JAVA applet.zip)

[Received October 2009. Revised February 2011.]

REFERENCES

Aittokallio, T., and Schwikowski, B. (2006), "Graph-Based Methods for Analysing Networks in Cell Biology," *Briefings in Bioinformatics*, 7, 243–255. [996]

Becker, R. A., Cleveland, W. S., and Shyu, M. J. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Graphical Statistics*, 5, 123–155. [994]

Bertin, J. (1967), *Sémiologie Graphique*, Paris: Editions Gauthier-Villars. English translation by William J. Berg as *Semiology of Graphics, Networks, Maps* (1983), University of Winsonsin Press. [995]

——— (1999), *Graphics and Graphic Information Processing*, San Francisco: Morgan Kaufmann. [995]

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–225. [1000,1001]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n," *The Annals of Statistics*, 35, 2313–2351. [991]

Cestnik, G., Konenenko, I., and Bratko, I. (1987), "Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users," in *Progress in Machine Learning*, eds. I. Bratko and N. Lavrac, Wilmslow, England: Sigma Press, pp. 31–45. [990,1000]

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth. [994]

Chen, C., Härdle, W., and Unwin, A. (2008), *Handbook of Data Visualization*, Berlin: Springer-Verlag. [995]

Chen, S. S., and Donoho, D. L. (1998), "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, 20, 33–61. [991]

Dempster, A. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [996]

Diaconis, P., and Efron, B. (1983), "Computer-Intensive Methods in Statistics," *Scientific American*, 248, 116–131. [990,1000]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space," *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911. [997]

Friendly, M. (2002), "Corrgrams," *The American Statistician*, 56, 316–324. [995]

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537. [990,997,999]

Grindea, S., and Postelnicu, V. (1977), "Some Measures of Association," in *Proceedings of the Fifth Conference on Probability Theory (Braşov, 1974)*, eds. B. Bereanu, M. Iosifescu, and G. Popescu, Bucharest: Editura Academia Republicii Socialiste România, pp. 197–203. [989]

Hall, P., and Miller, H. (2009), "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics*, 18, 533–550. [991,994,997]

Hall, P., Titterington, D. M., and Xue, J. (2009), "Tilting Methods for Assessing the Influence of Components in a Classifier," *Journal of the Royal Statistical Society, Ser. B*, 71, 783–803. [991]

Hartigan, J. A. (1972), "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association*, 67, 123–129. [995]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman & Hall. [995]

Hills, M. (1969), "On Looking at Large Correlation Matrices," *Biometrika*, 56, 249–253. [996]

Kruskal, J. B., and Wish, M. (1978), *Multidimensional Scaling*, Newbury Park, CA: Sage Publications. [996]

Meinshausen, N., and Bühlmann, P. (2006), "High Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [996]

Peng, J., Wang, P., Zhou, N. F., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Model," *Journal of the American Statistical Association*, 104, 735–746. [995,996]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [991]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences of the USA*, 99, 6567–6572. [997]

Wills, G., and Wilkinson, L. (2008), "AutoVis: Automatic Visualization," *Information Visualization*, 9, 47–69. [996]

Zhu, D., Hero, A. O., Cheng, H., Khanna, R., and Swaroop, A. (2005), "Network Constrained Clustering for Gene Microarray Data," *Bioinformatics*, 21, 4014–4020. [996]