

Statistical Computing and Graphics

Graphical Assessment of Dependence: Is a Picture Worth 100 Tests?

N. I. FISHER and P. SWITZER

Nonparametric tests of association have been around for a long time, and new ones continue to be proposed to cope with specific forms of association. However, graphs have the potential to assess a far richer class of bivariate dependence structures than any collection of tests. This article describes how chi-plots, used in conjunction with the usual scatterplot, provides a useful practical tool in this regard.

KEY WORDS: Chi-plot; Copula; Nonparametric association; Rank tests of independence.

1. INTRODUCTION

Nonparametric statistical tests for association have a long and honorable history, with their roots reaching back at least to the works of I. J. Bienamyé and O. Rodrigues in the 19th Century; see Heyde and Seneta (1977) and Fisher (1983), respectively. The still-used Spearman test for rank association appeared in 1904 and, even now, formal statistical tests are appearing in the literature (e.g., Kallenberg and Ledwina 1999). Graphical procedures relating to tests of association also have a long history (Fisher 1983), mainly because a graphical representation sometimes provided a way of calculating the test statistic. This role as a computational aid has long since been ceded to computers, and graphs are now used almost exclusively for revealing pattern. In this latter role, graphs can be a rich source of information about association, in contrast with formal tests, that can provide at best a single piece of information about a single form of association. The capabilities of the basic scatterplot are well known in this regard and, as a plot of the raw data (or a rank transform thereof), it will always remain a primary data-analytic tool. However, it does have at least one major deficiency.

When we look at a scatterplot to detect pattern, the “null” model, corresponding to independence, is a random scatter of points. Unfortunately, randomness is a difficult characteristic for the human eye to judge. So, it may be very desirable to have an auxiliary display in which independence is itself manifested in a characteristic fashion. The chi-plot (or χ -plot; Fisher and Switzer 1985) is designed to address this sort of problem. It

supplements a scatterplot of the data by providing a graph that has characteristic patterns depending on whether the variates (a) are independent, (b) have some degree of monotone relationship (i.e., nonzero grade correlation), or (c) have more complex dependence structure. The χ -plot depends on the data only through the values of their ranks. The purpose of this article is to illustrate some of the wide variety of forms of dependence that a single χ -plot can highlight.

Section 2 contains a brief description of the χ -plot algorithm and illustrates its behavior in the simplest cases of independence and monotone dependence. Section 3 provides a catalog of the typical behavior of the χ -plot in the presence of more complex forms of dependence. Section 4 contains some examples using real data, including an example with multivariate data.

2. THE CHI-PLOT

2.1 Definition

Let $(x_1, y_1), \dots, (x_n, y_n)$ be a random sample from H , the joint (continuous) distribution function for a pair of random variables (X, Y) , and let $I(A)$ be the indicator function of the event A . For each data point (x_i, y_i) , set

$$H_i = \sum_{j \neq i} I(x_j \leq x_i, y_j \leq y_i) / (n - 1),$$

$$F_i = \sum_{j \neq i} I(x_j \leq x_i) / (n - 1),$$

$$G_i = \sum_{j \neq i} I(y_j \leq y_i) / (n - 1),$$

and

$$S_i = \text{sign} \left\{ \left(F_i - \frac{1}{2} \right) \left(G_i - \frac{1}{2} \right) \right\}.$$

Now calculate

$$\chi_i = (H_i - F_i G_i) / \{ F_i (1 - F_i) G_i (1 - G_i) \}^{\frac{1}{2}}$$

and

$$\lambda_i = 4 S_i \max \left\{ \left(F_i - \frac{1}{2} \right)^2, \left(G_i - \frac{1}{2} \right)^2 \right\}$$

A χ -plot is a scatterplot plot of the pairs (λ_i, χ_i) , $|\lambda_i| < 4 \left\{ \frac{1}{n-1} - \frac{1}{2} \right\}^2$. The value λ_i is a measure of the distance of the data point (x_i, y_i) from the center of the dataset. A χ -plot

N. I. Fisher has an Honorary Position, School of Mathematics, University of Sydney. Mailing address: ValueMetrics Australia, P.O. Box 21, Roseville NSW 2069, Australia (E-mail: nif@valuemetrics.com.au). P. Switzer is Professor, Department of Statistics, 390 Serra Mall, Stanford University, Stanford, CA 94305.

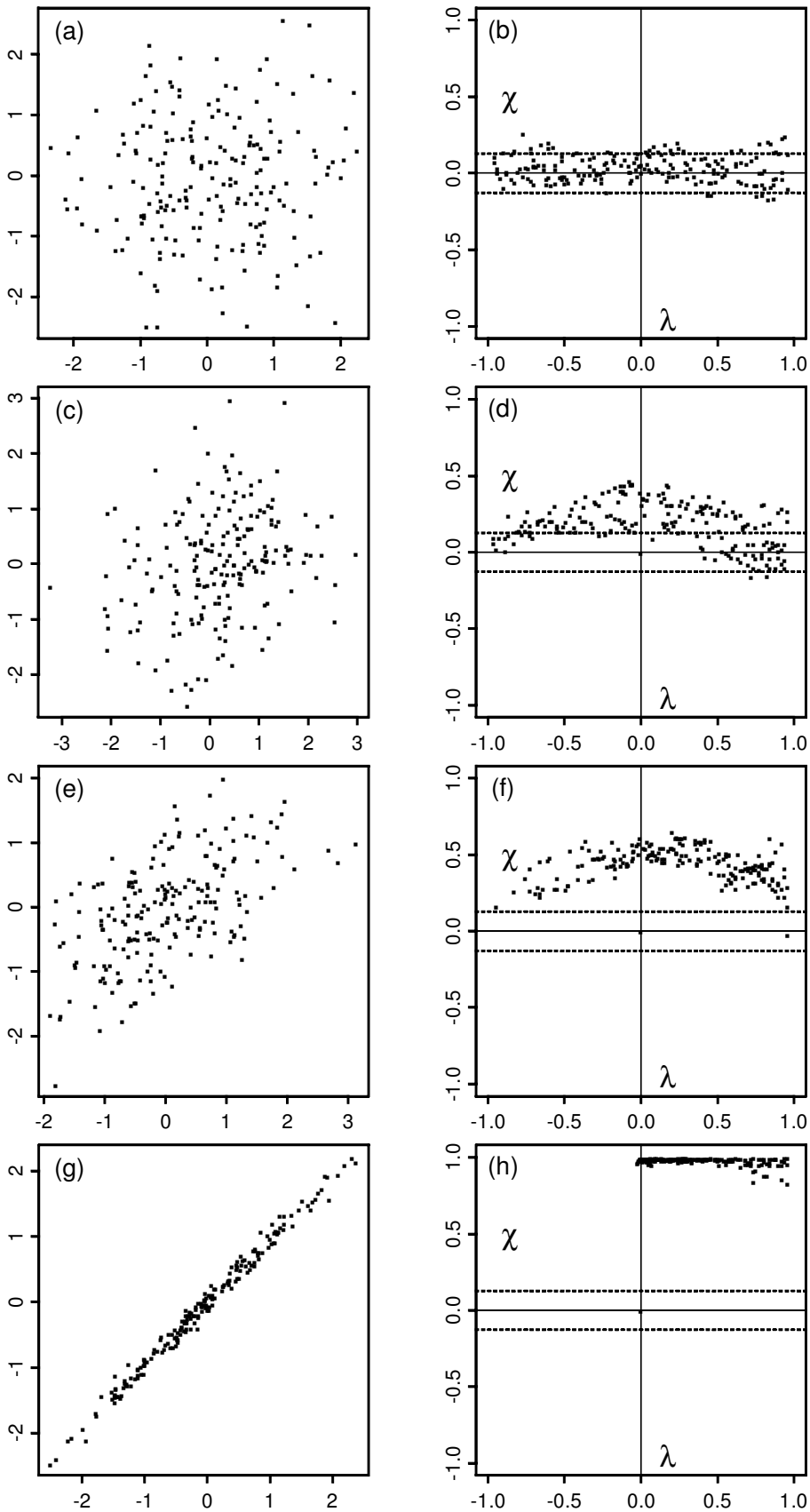


Figure 1. Behavior of the χ -plot in Simple Situations. The left column shows scatterplots and the right column their corresponding χ -plots, for simulated samples of size 100 from the bivariate normal distribution with respective correlation coefficients 0.0, 0.2, 0.5, and 0.95.

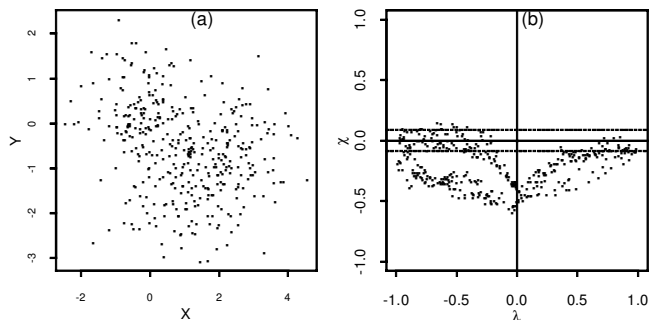


Figure 2. Behavior of the χ -plot for a mixture of two populations. (a). Scatterplot of data. In the first population, each variable tends to have lower values, and the variables are negatively associated; whereas in the second, the variables tend to have higher values and tend to be positively associated. (b) χ -plot, showing distinctive 'lobes'.

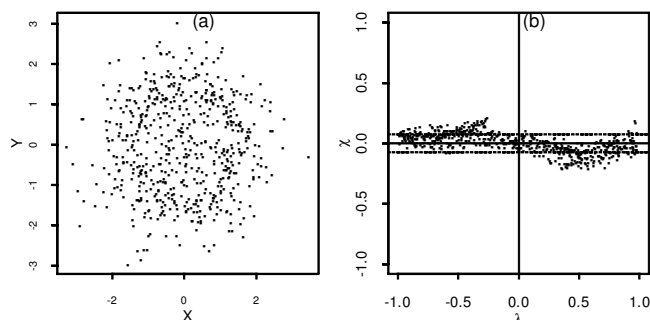


Figure 3. Behavior of the χ -plot data with a "hole." (a). Scatterplot of a sample of data simulated from a bivariate normal distribution with zero correlation, the sample having been depleted within a radius 1.5 of the mean. (b) χ -plot, showing the χ_i -values centered around 0, but having greater dispersion than would otherwise be expected.

is uniquely determined by the copula of the joint distribution (Fisher and Switzer 1985).

It is sometimes helpful to supplement the basic χ -plot with a pair of horizontal guidelines at $\chi = -c_p/n^{1/2}$ and $\chi = c_p/n^{1/2}$, where c_p is selected so that approximately $100p\%$ of the pairs (λ_i, χ_i) lie between the lines. The c_p -values 1.54, 1.78, and 2.18 correspond, respectively, to $p = 0.90, 0.95$, and 0.99 ; others are readily established by Monte Carlo methods. An S-Plus function implementing the χ -plot on a PC is available from <http://www.cmis.csiro.au/Nick.Fisher>.

2.2 Basic Properties

Figure 1 shows some prototypical examples of how a χ -plot performs when X and Y are independently distributed, and then in the presence of increasing monotone association. Each row contains a scatterplot and the corresponding χ -plot for a synthetic dataset of size $n = 100$, where X and Y have a joint Gaussian distribution with correlation $\rho = 0, 0.2, 0.5, 0.95$. If there is no relationship between X and Y , approximately 95% of the χ_i values should plot between the two control lines. λ_i is a measure of the distance of the point (x_i, y_i) from the center of the data as measured by (\tilde{x}, \tilde{y}) , where \tilde{x} and \tilde{y} are the median values. Thus, a positive value of λ_i means that both x_i and y_i are large relative to their respective medians, or both small; whereas, a negative value of λ_i corresponds to x_i and y_i being on opposite sides of their respective medians.

2.3 Extensions

The χ -plot can easily be adapted to situations other than the simple model of a random sample of bivariate data.

One variable is nonstochastic. A typical example would be one in which X is a design variable in a regression model and the corresponding Y values are the residuals from a fitted model where X is one of the explanatory variables. The control lines will remain approximately valid.

Serial correlation. Let X_1, X_2, \dots be a sequence of random variables hypothesized to be independently distributed with a common distribution F , and suppose that we wish to check for correlation at lag k . A χ -plot can be constructed for the data pairs $(x_1, x_{k+1}), \dots, (x_{n-k}, x_n)$. Again, the control lines will

remain approximately valid, even for k relatively large compared with the sample size n (e.g. $n = 1000, k = 500$).

3. PROTOTYPE CHI-PLOTS FOR COMPLEX DEPENDENCE

Change in form of association. In certain areas of application it is not uncommon to find that different parts of the dataset appear to have different association. Physical mechanisms that produce this effect include

1. data below a detection limit being recorded as being at the detection limit; and
2. measurements from different populations being inadvertently mixed.

Depending on how these occur, the χ -plot will manifest distinctive features. An example for a mixture of two populations is given in Figure 2. In the first population, each variable tends to have lower values, and the variables are negatively associated; whereas in the second, the variables tend to have higher values and tend to be positively associated. Let $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ denote the bivariate normal distribution with means μ_1, μ_2 , variances σ_1^2, σ_2^2 and correlation coefficient ρ . The simulated data comprise 200 measurements from each of $N(-2, -2, 1, 1, -0.4)$ and $N(2, 2, 1, 1, 0.4)$. As can be seen in Figure 2(b), the points for which $\lambda_i < 0$ are grouped into two "lobes."

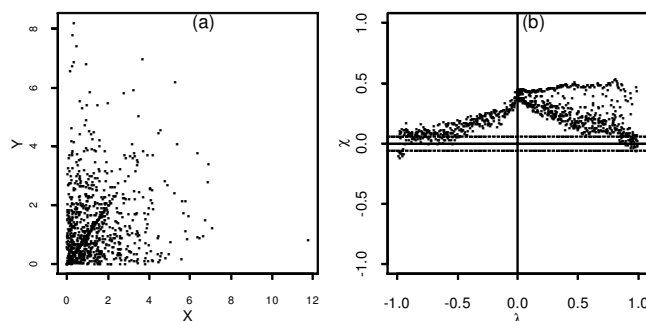


Figure 4. A scatterplot and its corresponding χ -plot for data simulated from a Marshall-Olkin distribution. In comparison with the prototype examples of monotone association in Figure 1, there is a distinctive feature that points to more complex association in the data.

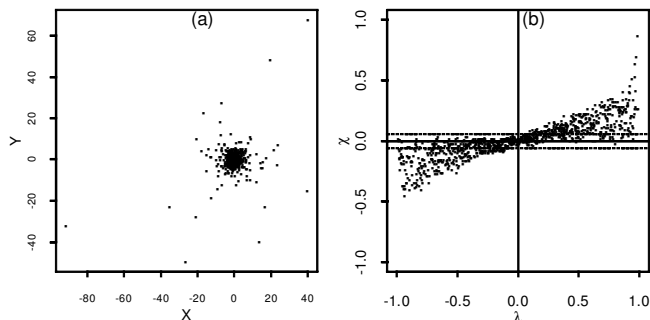


Figure 5. A scatterplot and its corresponding χ -plot for data simulated from a Pearson Type VII distribution. Comparison with the prototype examples of monotone association in Figure 1 shows that this form of association is manifested in a distinctly different way.

Distributions with holes. Sometimes, bivariate datasets have holes; that is, they are depleted in an interior region of the scatterplot. To illustrate how this effect can appear in a χ -plot, we use a sample of synthetic data drawn from a standard bivariate normal distribution with zero correlation, and subject to the restriction that a randomly selected 75% of the pairs $(x_i, y_i) : |x_i^2 + y_i^2| < 1.5$ have been omitted. The resulting sample of 627 pairs, depicted in Figure 3(a), has essentially no monotone association (rank correlation $\hat{\rho}_S = 0.026$, corresponding to a two-sided significance probability of $P = 0.52$). The corresponding χ -plot in Figure 3(b) is rather complex. Because $\hat{\rho}_S \approx 0$, the χ values tend to be centered at zero. However, the pattern clearly differs from that in Figure 1(b).

Special bivariate distributions. Kallenberg and Ledwina (1999) studied a class of tests of independence of particular application when the grade correlation may be very low yet more complex forms of association may be present. They illustrated their methods for several bivariate families. In Figures 4 and 5, two interesting cases from Kallenberg and Ledwina (hereafter denoted KL) are considered: *Alternative 2.2* (a Marshall–Olkin distribution) and *Alternative 2.4* (a Pearson Type VII distribution). Figures 4(a) and 4(b) contain, respectively, a scatterplot and its corresponding χ -plot for 1,000 values simulated from a Marshall–Olkin distribution with $\lambda_{12} = 0.25$; for this sample, $\hat{\rho}_S = 0.26$ ($P = 0.00$). Comparison with Figure 1(b) reveals

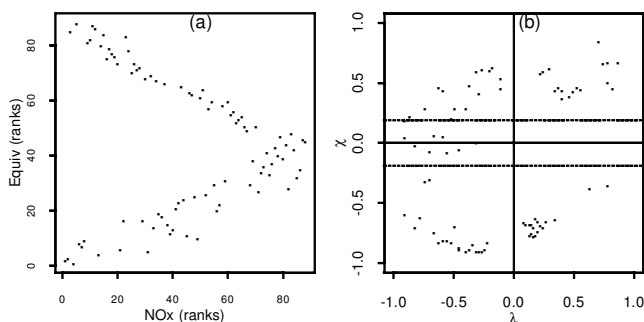


Figure 6. A rank scatterplot and its corresponding χ -plot for the automobile exhaust data (88 measurements of equivalence ratio and the corresponding air-ethanol) in Example 1. The complex form of association, coupled with lack of overall monotone association, lead to a χ -plot with χ -values centered around zero, but more dispersed than one would expect if the variates were independent.

distinct and possibly characteristic differences between the corresponding χ -plots. Figure 5 contains the corresponding plots for data simulated from a Pearson Type VII distribution with $\nu = 1.5$; here, $\hat{\rho}_S = 0.01$ ($P = 0.70$). Again, the contrast with Figure 1(b) is striking and would point to “significant” non-monotone association.

Other alternatives investigated in this way also produced χ -plots distinctively different from Figure 1(b).

4. EXAMPLES

Fisher and Switzer (1985) looked at a number of applications to real data. Here we look at three other examples, the first two from KL and the third an application to multivariate data.

Example 1. KL’s first example is a set of 88 measurements of the relationship between the Equivalence ratio (NOx, the concentration of nitric oxide NO and nitrogen dioxide NO₂ in engine exhaust, normalized by the work done by the engine), and a measure of the richness of the air/ethanol mix: see KL and Simonoff (1996, p. 137) for details. For these data, $\hat{\rho}_S = -0.14$ ($P = 0.19$). Figure 6 contains a *rank scatterplot* (a scatterplot of the ranks of the data) and the corresponding χ -plot. (We have chosen to use a rank scatterplot here for two reasons. First, in these examples we are interested in the gain in information provided by a χ -plot compared with a rank test. Second, the χ -plot seeks to distinguish independence from monotone association from more complex association. Because it depends on the data only through the values of their ranks, it cannot distinguish linear from more general monotone association.)

This example hardly calls for use of the χ -plot, but may be helpful in emphasizing some of its properties. The absence of monotone association means that the χ -values tend to be centered at zero. The structure in the data manifests itself as abnormal dispersion of the χ -values. A third factor may well explain the basic data pattern.

Example 2. KL’s second example is a set of 28 measurements of size of the annual spawning stock of salmon compared with corresponding production of new catchable-sized fish in the Skeena River. Figure 7(a) and 7(b) shows the rank scatterplot and χ -plot ($\hat{\rho}_S = 0.55$, $P = 0.75$). Comparison with the prototype plots in Figure 1 suggests that something more than simple monotone association is involved. Indeed, consider the 28 synthetic data in Figure 7(c), simulated from a bivariate normal distribution. The Spearman correlation is also about 0.55 for simple monotone association, but we see that the corresponding, and typical, χ -plot in Figure 7(d) is very different. One possible explanation is that two populations have been sampled.

Example 3. χ -plots can be particularly helpful when a large number of bivariate relationships have to be evaluated; for example, in a scatterplot matrix for a large number of variates. This situation arises commonly in the analysis of multielement geochemical data. Figure 8(a) shows a combination of a scatterplot matrix and a χ -plot matrix, a form of display that was used in initial screening of pairwise relationships by Griffin et al. (1999). (The data are used with the kind permission of W. L. Griffin.) The full dataset comprises 13,317 analyses of individual grains of mantle-derived peridotite garnet. One thousand measurements were selected at random, and a subset of 13 of

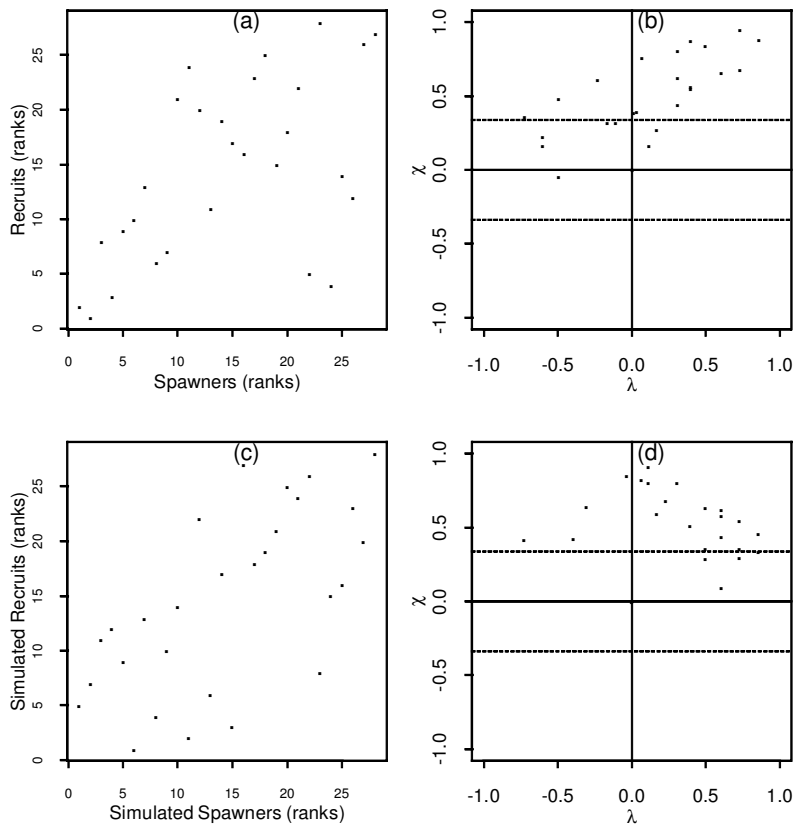


Figure 7. (a) and (b). A rank scatterplot and its corresponding χ -plot for the 28 measurements of size of the annual spawning stock of salmon and corresponding production of new catchable-sized fish in the Skeena River (Example 2). The pattern suggests significant monotone increasing association between the variates. However, the nature of the association is more complex than this. (c) and (d) show a scatterplot and the corresponding χ -plot for sample of 28 pairs simulated from a normal distribution, with the sample having the same value of grade correlation as the original data. The pattern in the χ -plot is consistent with data being sampled from a particular type of mixture of distributions.

the variables has been used in Figure 8(a). The lower triangular area shows scatterplots of the ranks of the data, with the corresponding χ plots shown above the diagonal. Here, plotting the ranks of the data has an added advantage, because each of the distributions is quite long-tailed. In such cases, the rank scatterplot is more effective than an ordinary scatterplot in revealing most sorts of association.

This form of display has been found to be very useful in initial screening of the data to detect potentially interesting associations that are not evident in the scatterplots. The set of χ -plots is rich with information, and needs careful study. To show how the analysis might be carried further, we investigate the subset of plots for the variables Cr_2O_3 , FeO, MnO, MgO, and CaO, and use as a reference set the prototypical examples in Figures 1–3.

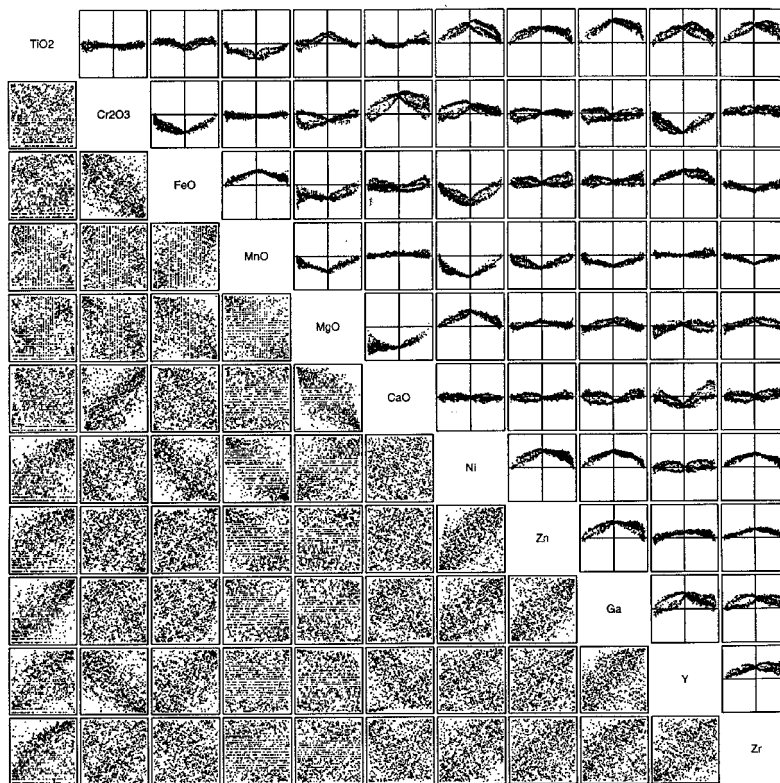
The pairs of variables that exhibit little or no association are (Cr_2O_3 , MnO) and (CaO, MnO); see Figure 1(b), which corresponds to data simulated from a pair of independent variates.

Next in terms of complexity are the pairs that suggest some degree of monotone association, either positive or negative. The principal form of association between FeO and MnO appears to be positive (monotone) correlation; see Figure 1(f), based on a sample from a bivariate normal distribution with correlation coefficient 0.5. Typical examples of negative monotone asso-

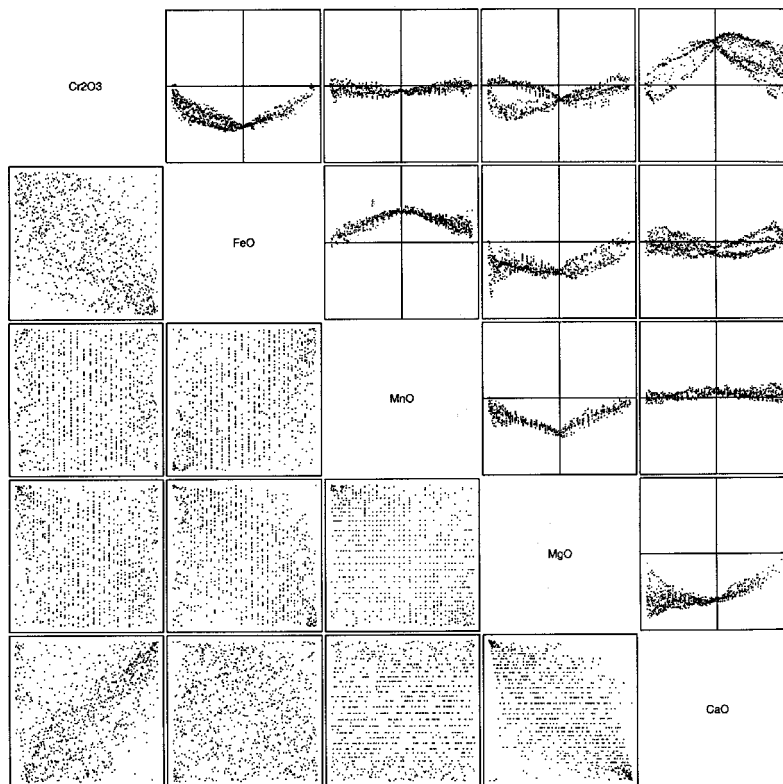
ciation are shown in the chi-plots for (Cr_2O_3 , FeO) and (MnO, MgO). By way of contrast, the corresponding scatterplots are rather harder to interpret; for example, (MnO, MgO).

The five remaining element pairs all exhibit more complex dependence structures:

- The χ -plot for (Cr_2O_3 , CaO) appears similar to the prototypical plot in Figure 2(b), which was generated from a mixture of two populations, with negative correlation for lower variate-values and positive correlation for higher variate values.
- The two pairs (Cr_2O_3 , MgO) and (FeO, CaO) appear not only to have been sampled from a mixture, but also to have elements of the dependence structure of Figure 3(b), which was derived from a bivariate normal distribution with zero correlation and with values near the origin under-represented.
- The remaining two pairs (Feo, MgO) and (MgO, CaO) each exhibit negative association. However, there is rather more: whereas with, say, correlated bivariate normal data, the association goes to zero at the extremes of the variate ranges, in these two plots the negative correlation persists at the low end.



(a)



(b)

Figure 10. (a). Rank scatterplots and corresponding χ -plots for 11 elements and minerals from a geochemical study. The χ -plots facilitate detection of potentially interesting and complex relationships between the variate pairs. (b) Rank scatterplots and corresponding χ -plots for a subset of 5 of the minerals shown in Figure 8(a).

5. CONCLUSION

We have demonstrated through simulated and real data examples a variety of patterns of bivariate association that are clearly exhibited by χ -plots and largely concealed in the corresponding scatterplots. This suggests that the χ -plot matrix is a useful tool for exploring multivariate data and, in particular, is far richer in information than a set of formal statistical tests such as those proposed by Kallenberg and Ledwina (1999). Simulations using bivariate normal data as well as non-normal bivariate distributions and mixtures of distributions can be used as a guide in interpreting χ -plots, aided by the close link between the χ -plot and the copula function of the bivariate distribution.

{Received February 2000. Revised April 2001.}

REFERENCES

- Fisher, N.I. (1983), "Graphical Methods in Nonparametric Statistics: A Review and Annotated Bibliography," *International Statistical Review*, 51, 25–58.
- Fisher, N. I., and Switzer, P. (1985), "Chi-plots for Assessing Dependence," *Biometrika*, 72, 253–265.
- Griffin, W. L., Fisher, N. I., Friedman, J. H., Ryan, C. G., and O'Reilly, S. (1999), "Cr-Pyrophe Garnets in Lithospheric Mantle. I. Compositional Systematics and Relations to Tectonic Setting," *Journal of Petrology*, 40, 679–704.
- Heyde, C. C., and Seneta, E. (1977), *I.J. Bienayme. Statistical Theory Anticipated*, New York: Springer-Verlag.
- Kallenberg, W. C. M., and Ledwina, T. (1999), "Data-Driven Rank Tests for Independence," *Journal of the American Statistical Association*, 94, 285–301.
- Simonoff, J. (1990), *Smoothing Methods in Statistics*, New York: Springer-Verlag.