Computers & Geosciences 37 (2011) 1677–1684

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Plotting and checking the bivariate distributions of multiple Gaussian data ☆

Jared L. Deutsch *, Clayton V. Deutsch

Centre for Computational Geostatistics (CCG), Department of Civil and Environmental Engineering, University of Alberta, 3-133 NREF Building, Edmonton, AB, Canada T6G 2W2

ARTICLE INFO

Article history:
Received 21 July 2010
Received in revised form
18 January 2011
Accepted 21 January 2011
Available online 21 March 2011

Keywords:
Geostatistics
Multinormal distribution
Multivariate statistics
Outlier detection

ABSTRACT

The geostatistical modeling of continuous variables relies heavily on the multivariate Gaussian distribution. It is remarkably tractable. The multivariate Gaussian distribution is adopted for K multiple variables (often K is between 2 and 10) and for N multiple locations (often N is in the tens of millions). Our focus is on the relationship between the K variables. Each variable is transformed to be univariate Gaussian, but the multivariate nature of the data is not necessarily Gaussian after univariate transformation. If multiple data variables are deemed non-Gaussian, then additional steps need to be taken such as linearization by alternating conditional expectation (ACE) or multivariate transformation by the stepwise conditional transformation (SCT). Although all L-variate distributions ($1 < L \leq K$) should be checked, the bivariate distributions are practically important; there are relatively few data in practice to investigate higher order distributions. A quantitative measure of departure from the bivariate Gaussian distribution is established based on quadrants and the distribution of differences from the theoretically expected distribution. Although approximate, the measure of departure is useful for comparing different distributions and guiding the geostatistician to look closer at some data variables. A `scatnscores` program is shown that will plot all $K(K-1)/2$ bivariate cross plots associated with K variables. The correlation coefficients, number of data, degree of departure from the bivariate Gaussian distribution, and bivariate Gaussian probability contours associated with specified cumulative probabilities are shown. The data ID numbers can also be shown to help identify outlier or problematic data.

1. Introduction

Geostatisticians are increasingly faced with multiple regionalized variables including multiple secondary data sources and multiple correlated variables to predict. Variables in the earth sciences are almost always related in some manner. Quantifying these relationships is relatively simple if the multivariate distribution is Gaussian after univariate transformation of each variable. Often, important relationships are bivariate in nature and require only the assumption that the variables conform to a bivariate Gaussian distribution. In geostatistical models, this conformance to a bivariate Gaussian distribution is often assumed and not checked because of the difficulty in constructing a true statistical test. Moreover, there may be too few data to test for conformance to a multivariate Gaussian distribution with any significance. This paper addresses this problem with the introduction of a statistical measure of departure from the bivariate Gaussian distribution. A program `scatnscores` to make bivariate plots with the probability contours for the bivariate normal distribution is also introduced to aid in plotting, finding outliers, and detecting bivariate relationships that warrant closer attention.

Consider two random variables X and Y that are univariate standard normal with a known correlation coefficient, ρ. It can be convenient to assume that the distribution between X and Y is bivariate Gaussian but this is impossible to conclude solely based on the univariate normality of X and Y. The three most important sources of non-Gaussian behavior are illustrated in Fig. 1: nonlinearity, heteroscedasticity, and constraints (Leuangthong and Deutsch, 2003). When these deviations become significant, practitioners become uncomfortable assuming conformance to a bivariate Gaussian distribution for further calculations. Additional transformations would be necessary such as linearization by alternating conditional expectations (ACE) (Breiman and Friedman, 1985) or a multivariate transformation such as the stepwise conditional transformation (SCT) (Leuangthong and Deutsch, 2003). Alternatively, subsetting the data into different geological populations may lead to bivariate Gaussian distributions within the subsets.

There are a large number of statistical tests for the multivariate Gaussian distribution. Summaries of these tests and comparisons are available (Gnanadesikan, 1996; Thode, 2002) so only a brief survey of some of the major tests is included here. Most tests for the multivariate Gaussian distribution can be classified as a graphical method, skewness and kurtosis method, or related to the W test.

One of the principal graphical methods for testing for the multivariate normal distribution is the construction and inspection of a chi-squared plot of the Mahalanobis distances

☆ Code available from server at http://www.iamg.org/CGEditor/index.htm.

* Corresponding author. Tel.: +1 780 492 9916; fax: +1 780 492 0249.
E-mail address: jdeutsch@ualberta.ca (J.L. Deutsch).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2011.01.014

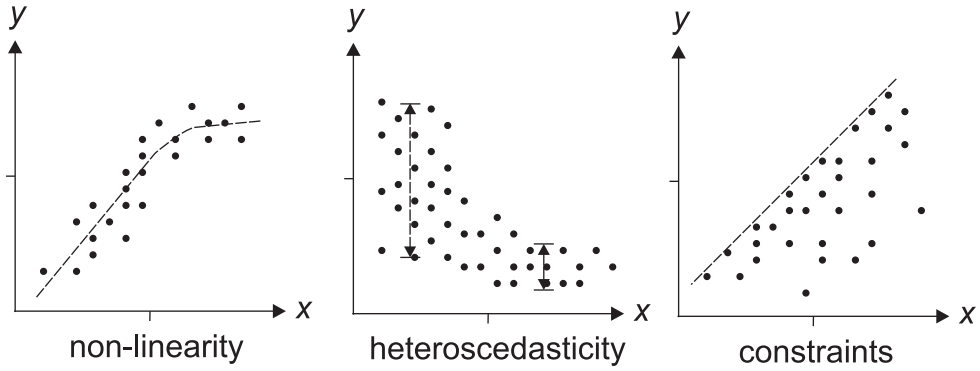


Fig. 1. Three important sources of non-Gaussian behavior: nonlinearity, heteroscedasticity, and constraints.

(Gnanadesikan, 1996; Johnson and Wichern, 2002; Wilks, 2006). The deviation from multivariate Gaussian behavior is measured using a correlation coefficient or similar measure. This is a powerful test for the bivariate Gaussian distribution but requires individual treatment of each plot to determine whether the deviation is significant or not, making it unwieldy for a large number of variables.

Another class of multivariate Gaussian tests includes the skewness and kurtosis methods (Mardia, 1970, 1974, 1975). These techniques have been applied to ore body analysis (Baxter and Gale, 1998) and are useful for detecting departures from the multivariate Gaussian distribution but are not as powerful as the *W* tests.

Tests using the *W* statistic comprise a powerful class of testing techniques for conformance to the univariate Gaussian distribution (Shapiro and Wilk, 1965; Shapiro et al., 1968). These are generally sensitive to all the major departures from normal behavior including both skewed and symmetric distributions. The *W* statistic method has been applied to multivariate Gaussian distribution testing (Royston, 1983) but did not show the same power as the test for the univariate Gaussian distribution since it must be applied to the marginal distributions of the multivariate.

All the above techniques generally involve a large amount of investigative work into each of the marginal distributions of the multivariate by the statistician and so do not lend themselves to a geoscientific application, which can involve a large number of variables.

A specialized check for the bivariate Gaussian distribution applied to indicator variables has been developed (Deutsch and Journel, 1998). This check calculates the cumulative probability for each quadrant in the bivariate Gaussian distribution and compares this with the observed proportion. This technique is suitable for assessing the viability of indicator techniques, but is not as suitable when applied to many different continuous variables. Deutsch and Journel note that a statistical check differs from a formal statistical test in that it is not suitable for rejecting a hypothesis. It can be, however, a useful approximation when a formal statistical test is unwieldy or unavailable. As with the technique proposed by Deutsch and Journel, the method proposed in this paper is a statistical check and not a formal statistical test.

2. Plotting

The probability density function of two random variables *X* and *Y* with a bivariate standard normal distribution (BVSN) is parameterized by the correlation coefficient, ρ, and describes a bell-shaped surface:

$$P(x,y)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right] \quad (1)$$

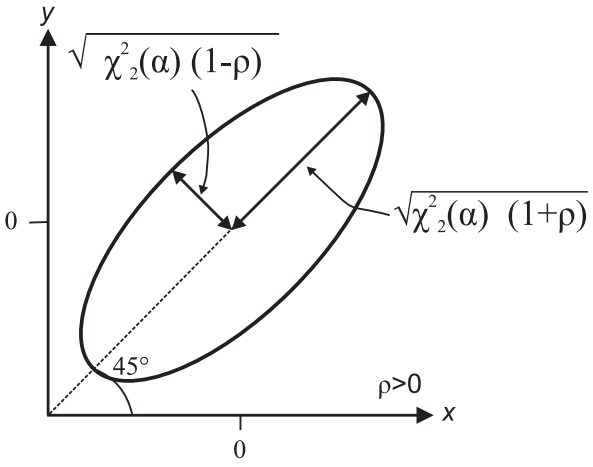


Fig. 2. Constant density ellipse ranges depend on the chi-squared value and the correlation coefficient.

While the probability density function of a bivariate distribution is three-dimensional, it is possible to trace constant density contours from the distribution onto the scatter plot of *X* and *Y*. A contour of this distribution on the scatter plot of *X* and *Y* has the general form of $c^2=x^2-2\rho xy+y^2$.

Choosing c^2 equal to the chi-squared probability with two degrees of freedom for a given probability α, $\chi^2_2(\alpha)$, gives a cumulative probability inside the ellipse equal to α. The contour of this ellipse is then given by $\chi^2_2(\alpha)=x^2-2\rho xy+y^2$. These constant density ellipses will always be oriented at $\pm 45^\circ$ relative to the *x*-axis depending on the sign of ρ. The ranges of these ellipses are dependent on the chi-squared value and correlation coefficient. This ellipse configuration is illustrated in Fig. 2.

To facilitate visual inspection of bivariate normal scatter plots, the `scatnscores` program plots the constant probability density contours for 25%, 50%, and 95%. It also calculates the correlation coefficient and the measure of deviation from the bivariate Gaussian distribution described in Section 5. An example scatter plot using the 2DWellData.dat data (Deutsch, 2006) is shown in Fig. 3. With the aid of these contours, it is reasonable to visually inspect individual scatter plots and determine if the variables are approximately bivariate Gaussian. However, as the number of variables increases this soon loses appeal due to the number of scatter plots to inspect (equal to $K(K-1)/2$). The contours are also useful in detecting outliers such as well 554, which falls far outside the 95% contour, as shown in Fig. 3.

For this reason we propose a quantitative check for the bivariate Gaussian distribution, which will identify bivariate relations that cannot be reliably considered Gaussian.

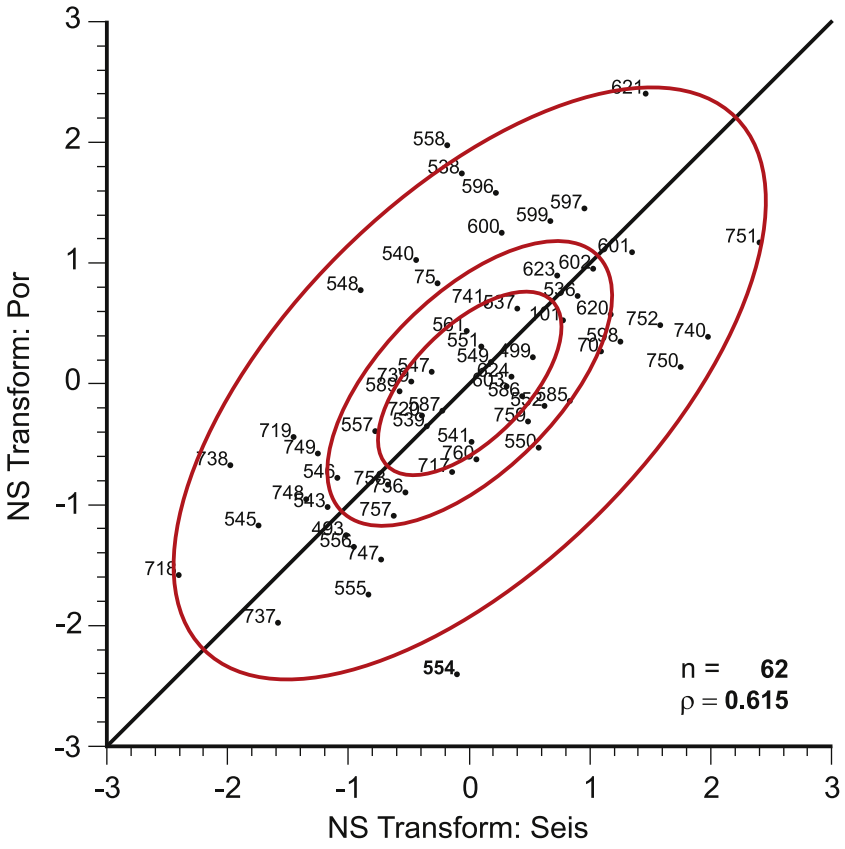


Fig. 3. Example cross plot of the familiar 2DWellData.dat. Note that well 554 has an unusually low porosity for the seismic value and should probably be checked.

3. Check for the bivariate Gaussian distribution

The proposed check for the bivariate Gaussian distribution focuses on the properties that a sample drawn from a bivariate Gaussian distribution should satisfy. The check is less powerful when fewer data pairs are used and more powerful as the number of data pairs is increased. This is considered in the final measure of how far the bivariate distribution departs from bivariate Gaussian.

The first step counts the fraction of points falling in each of the 25%, 50%, and 95% contours and compares them with the expected fraction. To check for this, the Mahalanobis distance, D_i^2, is calculated for each data pair (x_i, y_i):

$$D_i^2 = \frac{x_i^2 - 2\rho x_i y_i + y_i^2}{1-\rho^2} \tag{2}$$

The fraction of D^2 values that are less than each of the $\chi^2_2(\alpha)$ values should be equal to α. Calculated χ^2_2 values for α equal to 0.25, 0.50, and 0.95 are 0.5753, 1.3863, and 5.9915, respectively. It is expected that limited data will result in deviations from the expected fractions.

The second step compares the fraction of points falling within each of the four quadrants of the constant density ellipses and compares this with the expected value of 25% per quadrant. To do this, the data points are first transformed so that they are along the principal directions of the ellipse and then the fraction of points in each quadrant is determined. The rotation matrix, corrected for the sign of the correlation coefficient, is

$$\begin{bmatrix} x_R \\ y_R \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{\rho}{\sqrt{2}|\rho|} \\ \frac{-\rho}{\sqrt{2}|\rho|} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{3}$$

The signs of x_R and y_R are correlated to a quadrant using the scheme depicted in Fig. 4. The maximum allowed deviations are optimized so that they reflect a departure from the Gaussian behavior and not limited data.

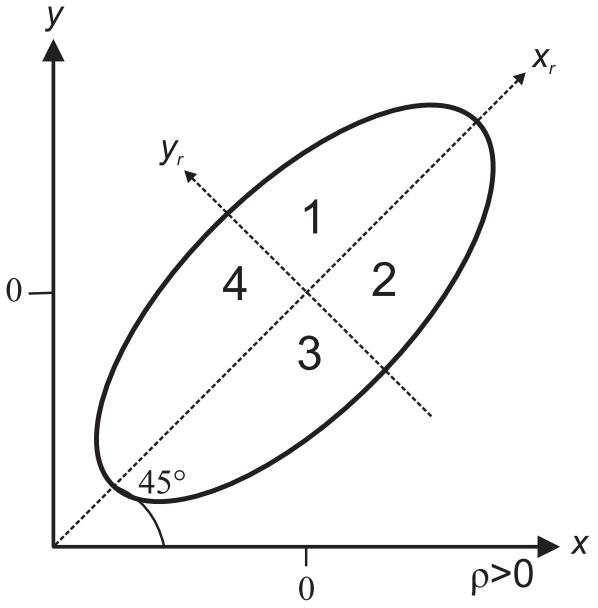


Fig. 4. Quadrants of the constant density ellipse relative to the rotated axes x_r and y_r.

4. Measure of deviation

The sum of small deviations is used to combine the two checks into a single measure of deviation from bivariate normality. The deviations from the expected number in each quadrant for

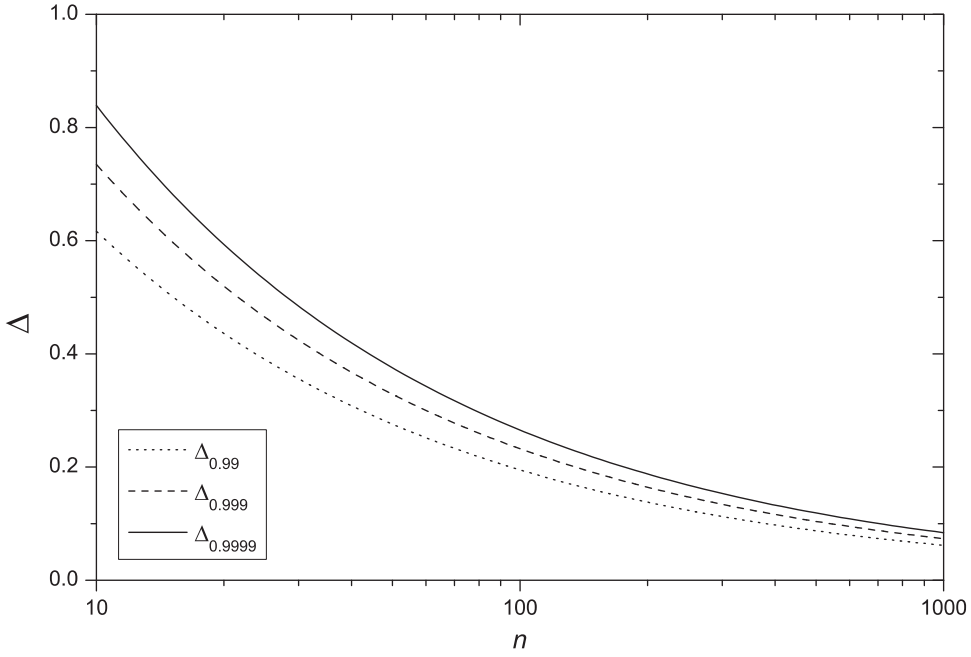


Fig. 5. Simulation of Δ values from a BVSN for various n. The upper curve is the 0.9999 quantile, the middle is the 0.999 quantile, and the lower curve is the 0.99 quantile.

each of 25%, 50%, and 95% contours are summed and averaged. This value, Δ, is an approximate measure of deviation from perfect bivariate normal behavior with infinite data. For n paired data points, this is

$$\Delta = \frac{1}{12}\sum_{i=1}^{3}\sum_{j=1}^{4}\left|\alpha_i - \frac{\text{Number} < \chi^2_2(\alpha_i) \text{ in quadrant } j}{n/4}\right| \tag{4}$$

A Monte Carlo simulation study was undertaken to determine the largest expected Δ for a given number of data points. 100,000 realizations each for n values from 10 to 1000 in increments of 10 were drawn from a bivariate standard normal distribution. The 0.99 quantile value of Δ for each value of n was calculated and fit using a power model: $\Delta_{0.99} = 1.94917n^{-0.49749}$. The simulating quantiles are plotted in Fig. 5 as well as the fit.

It can be seen that the power of n is very close to -0.5, which is expected given the $1/n$ decrease in variance and the linear relation between the quantiles and the standard deviation. For a Gaussian distribution, the standard deviation, which is normalized by $n^{-0.5}$, is a measure directly linked to quantiles. This relationship occurs in the standardization of a univariate Gaussian random variable X with mean μ and standard deviation σ to a standard normal variable Z, $Z=(X-\mu)/\sigma$, where the standard deviation is proportional to $n^{-0.5}$. To approximate the values for $\Delta_{0.999}$ and $\Delta_{0.9999}$, 10 million realizations were generated for low and high n values and the corresponding quantiles calculated. The power model fit of $\Delta_{0.99}$ was updated to reflect the Gaussian behavior of the distributions and is normalized by $n^{-0.5}$: $\Delta_{0.99} = 1.9492/\sqrt{n}$, $\Delta_{0.999} = 2.3234/\sqrt{n}$, and $\Delta_{0.9999} = 2.6528/\sqrt{n}$.

These empirically determined quantiles are provided because the distribution of Δ values is significantly skewed for larger sample sizes. A histogram of simulated Δ values for $n=100$ is provided in Fig. 6. For this reason, the power of this check is only for those quantiles (0.99, 0.999, and 0.9999). This level is suitable for most geostatistical applications as it can highlight bivariate distributions that are worth investigating using one or more of the formal approaches described earlier.

If a Δ value above $\Delta_{0.99}$ for a given number of data points is calculated then it is unlikely that the assumption of bivariate

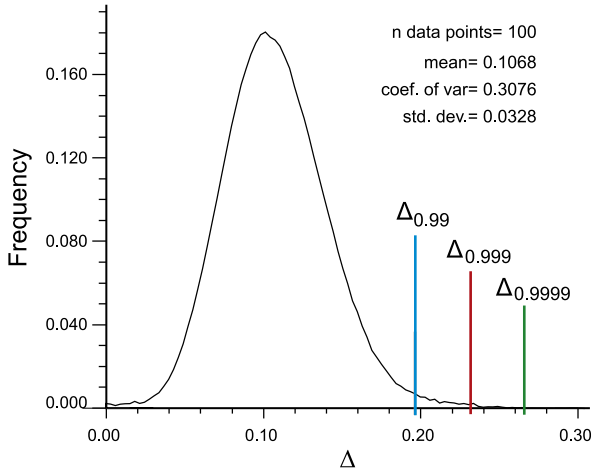


Fig. 6. Distribution of Δ values for various n. Histogram corresponds to $n=100$ exhibiting a slightly skewed shape as Δ is bounded on the left by 0. Histograms for $n=10$ and 1000 are very similar.

Gaussian is met and linearization by ACE or multivariate transformation by the stepwise conditional transformation may be necessary. We recommend a check scheme that uses a standardized value δ:

$$\delta = \frac{\Delta}{\Delta_{0.99}} = \frac{\Delta\sqrt{n}}{1.94917} \tag{5}$$

The calculated value of δ is checked to see what range it falls in and the bivariate distribution is classified as follows:

Degree of departure	Range	Classification
0	$\delta < 1$	Not enough evidence to assume non-Gaussian
1, $>99\%$ non-BVSN	$1 \le \delta < 1.192$	Bivariate is very likely non-Gaussian, check

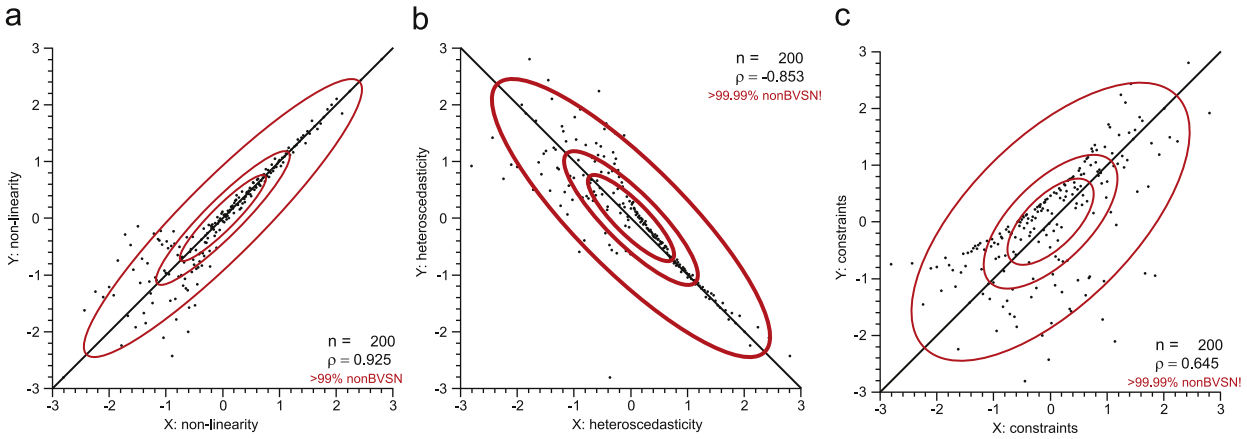
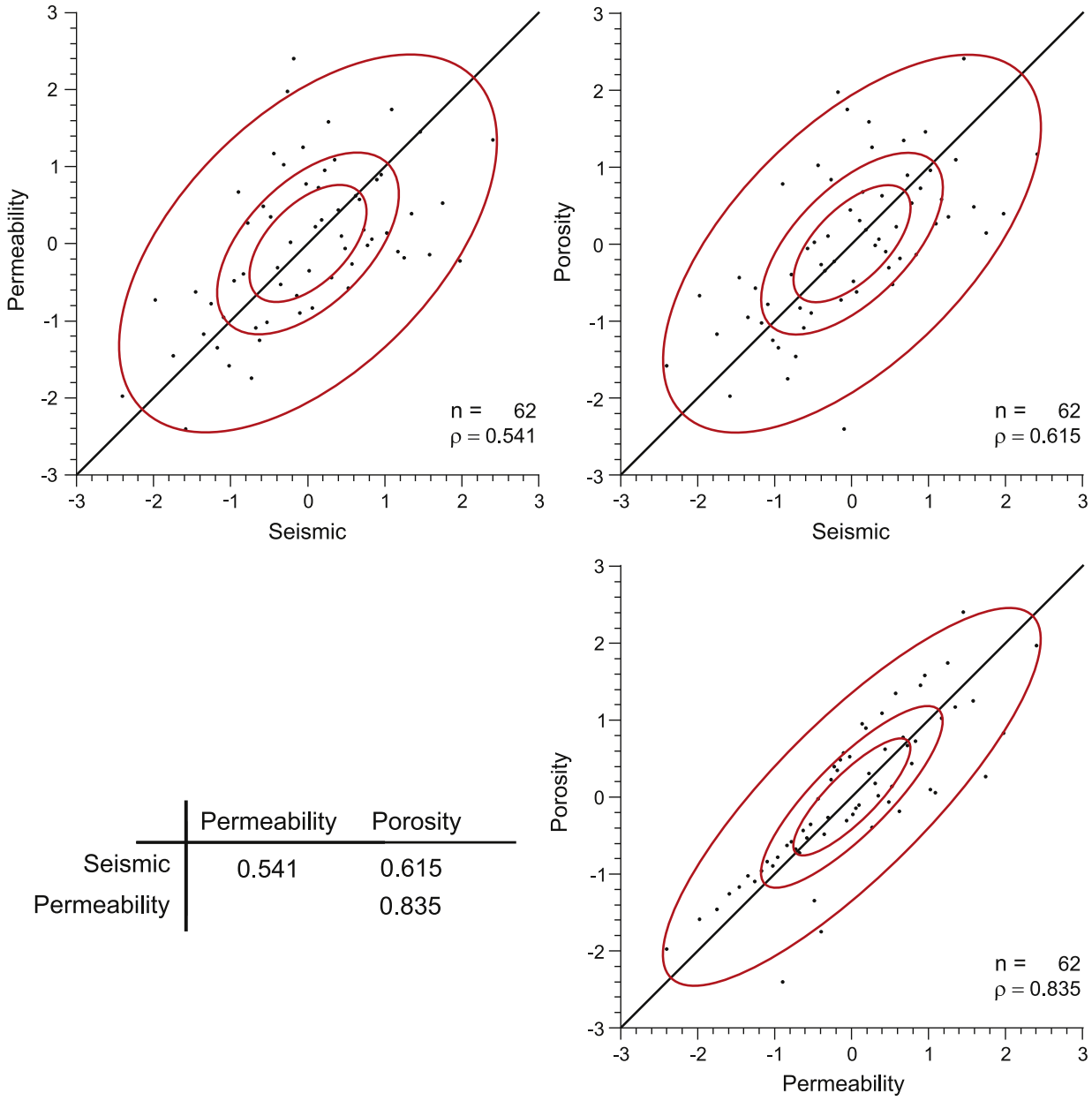


Fig. 7. Three fabricated data sets illustrating each of the principle phenomena responsible for non-Gaussian behavior: (a) nonlinearity, (b) heteroscedasticity, and (c) constraints.



	Permeability	Porosity
Seismic	0.541	0.615
Permeability		0.835

Fig. 8. Cross plots of training data from Deutsch (2006). A matrix of the correlation coefficients is given in the lower left hand corner.

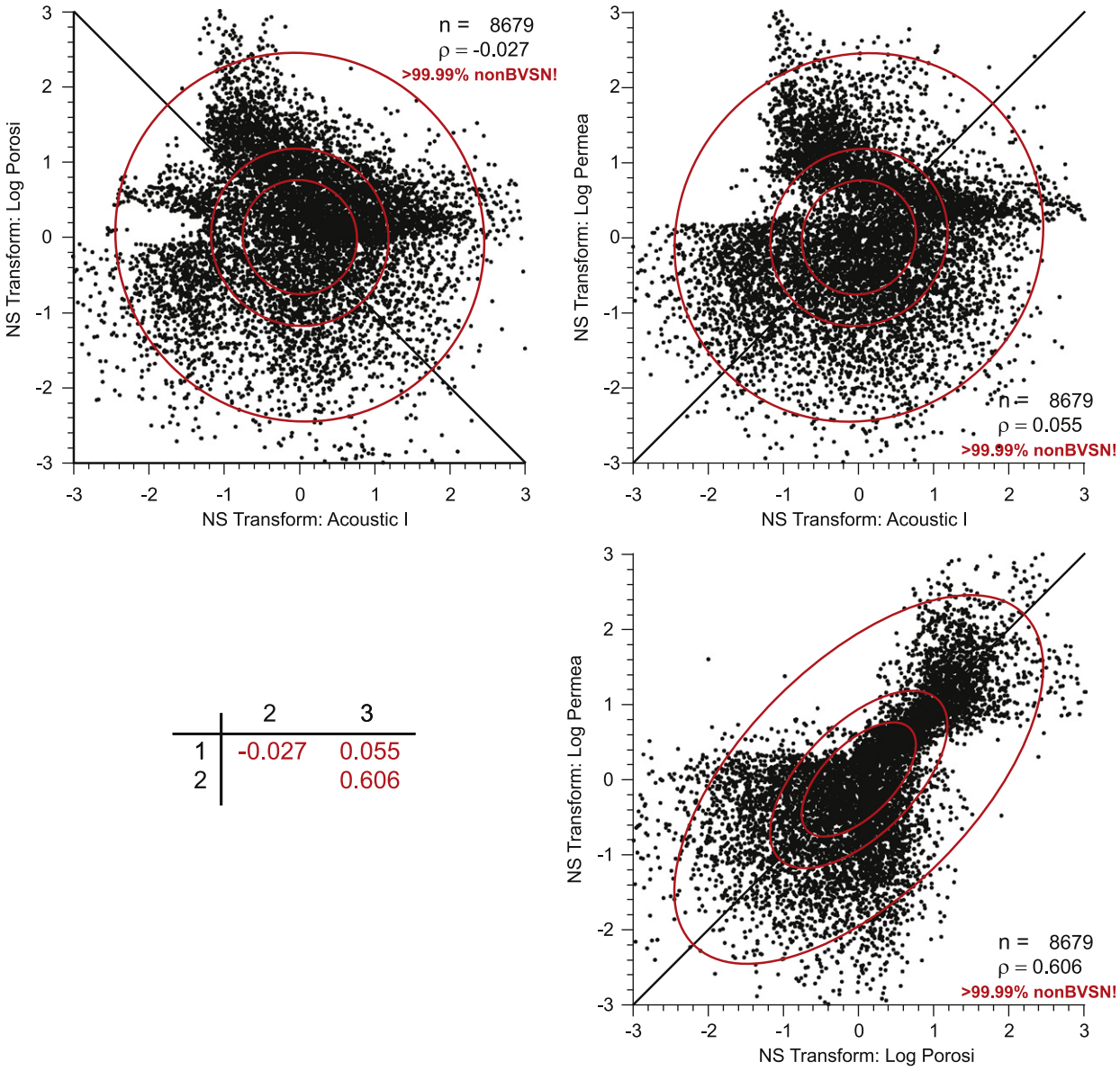


Fig. 9. Cross plots of DV well data. All bivariates exhibit significant non-Gaussian behavior, likely stemming from one or more three principle phenomena responsible for the non-Gaussian behavior (Fig. 7).

2, >99.9% non-BVSN	$1.192 \leq \delta < 1.361$	Bivariate is extremely likely non-Gaussian, check
3, >99.99% non-BVSN	$1.361 \leq \delta$	Bivariate is non-Gaussian and should be transformed

With fewer than 10 data points, it is unreasonable to use any quantitative check for the bivariate Gaussian distribution. With few data points only extreme departures from bivariate normality can be detected with reasonable assurance (Johnson and Wichern, 2002; Shapiro et al., 1968). The ease at which departures from the bivariate Gaussian distribution can be detected increases as the number of data points increases.

Given a large amount of data, the idea of comparing the proportion of data in the sector to the expected proportion can easily be extended to many sectors. For example, instead of checking quadrants, octants could be checked instead. Consider the limiting case with n sectors, each with a different radius. If each of the sectors contains the correct proportion of data for a bivariate Gaussian distribution, then the data must be bivariate Gaussian. However, as the number of sectors increases, the number of data points required greatly increases making it unrealistic to use many sectors for most geostatistical applications. For this reason, quadrant checking is implemented. The number of contours checked could be increased from 25%, 50%, and 95%. Increasing the number of contours and sectors would be warranted with many data. There is no specific threshold for the number of data required to warrant checking more than quadrants and three cumulative probability intervals; however, the authors speculate that more than five values outside the high probability contour would lead to stable proportions. Checking more than quadrants could be interesting with more than, say, 400 data.

5. Implementation

The program `scatnscores` implements the plotting and checks discussed above. Standard `GSLIB` conventions are used for the data file and parameter file. For details on the specific

implementation and FORTRAN90 code; see the accompanying electronic documentation. The parameters for the program:

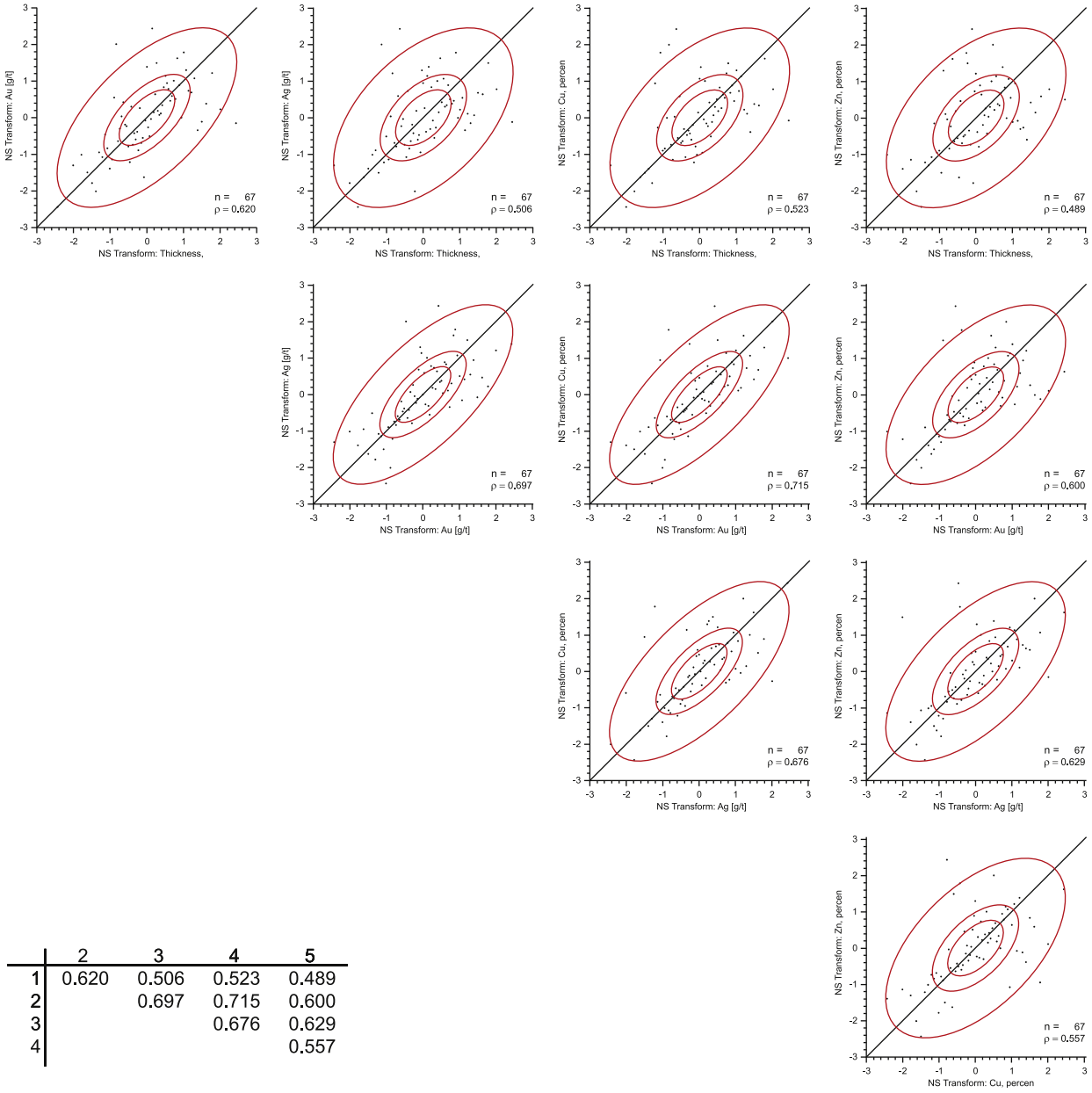
```
Line  START OF PARAMETERS:
1     nscores.dat      file with data
2     4                number of variables
3     12 14 15 16      columns for variables
4     0                column for data ID
5     scatnscores.ps   file for Postscript output
6     0.5              bullet size: 0.1(sml)-1(reg)-
                       10(big)
```

A number of data sets (Fig. 7) illustrating each of the principal behaviors responsible for non-Gaussian distributions are plotted using scatnscores and the Δ values calculated. Each set is composed of two random variables that have been normal score transformed with a calculated correlation coefficient. For simplicity, each of the data sets has 200 data pairs.

The first data set exhibits significant nonlinearity which results in a 1st degree departure from the bivariate Gaussian distribution. The second and third data sets, illustrating heteroscedasticity and constraints respectively, are also flagged with 3rd degree departures from bivariate Gaussian.

5.1. Case study 1: training data

The first data set studied is the 2DWellData (Deutsch, 2006). This data set was simulated from a multivariate Gaussian distribution for geostatistical training purposes so should not be flagged as non-Gaussian. The matrix of cross plots generated by scatnscores is included (Fig. 8) and there are indeed no cross plots flagged using the conditions laid out in Section 4.



	2	3	4	5
1	0.620	0.506	0.523	0.489
2		0.697	0.715	0.600
3			0.676	0.629
4				0.557

Fig. 10. Cross plots of the red.dat data. No bivariate exhibits significant non-Gaussian behavior; however, outliers are easily discernable using the 95% ellipse.

5.2. Case study 2: DV well data

The second data set checked was the `DV_Well.dat` set which is a data set with acoustic impedance, log porosity, and log permeability data. The three bivariates, shown in Fig. 9, all show significant deviations from Gaussian behavior. All three principal phenomena responsible for non-Gaussian behavior can be seen. The upper two cross plots of log porosity vs. acoustic impedance and log permeability vs. acoustic impedance both show signs of constraints and possible nonlinearity. The lower plot of log permeability vs. log porosity shows significant heteroscedasticity. The assumption of the bivariate Gaussian distribution is likely invalid and the data should be transformed if geostatistical techniques relying on this assumption are to be used.

5.3. Case study 3: red data

The final data set checked was data from a small polymetallic vein deposit, `red.dat`. This data set included thickness, gold, silver, copper, and zinc assays. The cross plots are shown in Fig. 10. There was no reason to reject any of the bivariate distributions as non-Gaussian, possibly due to the limited number of data (62). For this data set the multivariate Gaussian distribution could likely be assumed. There are, however, visible outliers in the cross plots which should be checked. With the large number of bivariates for this data set, checking for the bivariate Gaussian distribution using traditional techniques would be a time-consuming process.

6. Conclusion

The assumption of multivariate normality is widely employed for the geostatistical analysis of multiple variables. Deviations from multivariate normal behavior may have to be dealt with separately to ensure that the results of Gaussian simulation are acceptable. The check proposed in this note could be used whenever the set of bivariate distributions of a number of variables is suspect. This check is more powerful when large amounts of data pairs are available and is an efficient method for checking the assumption of the bivariate Gaussian distribution. In addition to checking the assumption of bivariate normality, the contour ellipses should be used to aid in tracking down data that deviate significantly from the expected distribution and could be problematic.

References

Baxter, M.J., Gale, N.H., 1998. Testing for multivariate normality via univariate tests: a case study using lead isotope ratio data. Journal of Applied Statistics 25, 671–683.

Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association 80 (391), 580–598.

Deutsch, C.V., 2006. What in the reservoir is geostatistics good for? Journal of Canadian Petroleum Technology 45 (4), 14–20.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide, 2nd ed. Oxford University Press, New York, NY (369 pp.).

Gnanadesikan, R., 1996. Methods for Statistical Data Analysis of Multivariate Observations, 2nd ed. John Wiley and Sons, New York, NY.

Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Statistical Analysis, 5th ed. Prentice-Hall, NJ (767 pp.).

Leuangthong, O., Deutsch, C.V., 2003. Stepwise conditional transformation for simulation of multiple variables. Mathematical Geology 35 (2), 155–173.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530.

Mardia, K.V., 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya: The Indian Journal of Statistics, Series B 36, 115–128.

Mardia, K.V., 1975. Assessment of multinormality and the robustness of Hotelling's T2 test. Applied Statistics 24, 163–171.

Royston, J.P., 1983. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. Applied Statistics 32, 121–133.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611.

Shapiro, S.S., Wilk, M.B., Chen, H.J., 1968. A comparative study of various tests for normality. Journal of the American Statistical Association 63, 1343–1372.

Thode Jr., H.C., 2002. Testing for Normality, 1st ed. Marcel Dekker, New York, NY (368 pp.).

Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences, 2nd ed. Elsevier Academic Press (648 pp.).