Multiple Bivariate Gaussian Plotting and Checking

Jared L. Deutsch and Clayton V. Deutsch

The geostatistical modeling of continuous variables relies heavily on the multivariate Gaussian distribution. It is remarkably tractable. If data are deemed nonGaussian, then additional steps need to be taken such as linearization by ACE or multivariate transformation by the stepwise conditional transformation. 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 to compare different distributions and guide 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 bivariate Gaussianity and bivariate Gaussian probability contours associated to specified cumulative probabilities are shown. The data IDs can also be shown to help track down outlier or problematic data.*

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 of bivariate Gaussianity. In geostatistical models, bivariate Gaussianity is often assumed and not checked due to lengthy procedures required, especially with a large number of variables. This short note addresses this problem with the introduction of a statistical measure of departure from bivariate Gaussianity.

Background

Consider two random variables X and Y that are univariate standard normal with a known correlation coefficient, ρ. It can be convenient to assume the distribution is bivariate Gaussian but this is impossible to conclude solely based on the univariate Gaussianity of X and Y. Recall the three most important sources of non-Gaussian behavior, illustrated in Figure 1: non linearity, heteroscedasticity and constraints. When these deviations become significant, we would not assume bivariate Gaussian behavior. Further transformations or subdivisions would be necessary.

There are a large number of statistical tests for multivariate Gaussianity. Comprehensive summaries of these tests are given by Gnanadesikan (1996) and Thode (2002) so only a brief survey of some of the major tests is included here. Most tests for multivariate Gaussianity can be classified as either a graphical method, a skewness and kurtosis method, or related to the *W*-test.

One of the principal graphical methods for testing for multivariate Gaussianity is the construction and inspection of a chi-squared plot of the Mahalanobis distances (Gnanadesikan, R., 1996, Johnson and Wichern, 2002 and Wilks, 2006). The deviation from multivariate Gaussian behavior is measured using a correlation coefficient or similar measure. This is a powerful test for bivariate Gaussianity but requires individual treatment of each plot by the statistician to determine whether the deviation is significant or not, which makes it unwieldy for a large number of variables.

Another class of multivariate Gaussianity tests are the skewness and kurtosis methods, pioneered by Mardia (1970, 1974 and 1975). These techniques have been applied to ore body analysis, (Baxter and Gale, 1998) and are useful for detecting departures from multinormality but were found by Baxter and Gale to not be as powerful as the *W*-tests.

Tests using the *W*-statistic, introduced by Shapiro and Wilk (1965) comprise a powerful class of univariate normality testing techniques as illustrated by the comprehensive study conducted by Shapiro *et al.* (1968). These are generally sensitive to all the major departures from normal behavior including both skewed distributions and symmetric distributions. Application of the *W*-statistic for multivariate

normality testing was detailed by Royston (1983) but did not show the same power as the test for univariate normality since it must be applied to the marginal distributions of the multivariate.

All of the above detailed 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 geostatistical application which can involve a large number of variables.

A specialized check for bivariate Gaussianity applied to indicator variables was discussed by 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 check and is not a formal statistical test.

Plotting

The probability density function of two random variables *X* and *Y* with a bivariate standard normal distribution (BVSN) is parametrized by the correlation coefficient, ρ (Equation 1) and describes a bell shaped surface. 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 scatterplot of *X* and *Y*.

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

A contour of this distribution on the scatterplot of *X* and *Y* has the general form of Equation 2. Choosing c^2 equal to the chi-squared probability with two degrees of freedom for a given probability α, $\chi^2_2(\alpha)$, means that the cumulative probability inside the ellipse is equal to α.

$$c^2 = x^2 - 2\rho xy + y^2 \tag{2}$$

For a bivariate standard normal distribution, the constant density ellipses will always be oriented at ±45° 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 Figure 2.

To facilitate visual inspection of bivariate normal scatterplots, 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 bivariate Gaussianity described below. An example scatterplot using the familiar 2DWellData.dat from Deutsch (2006) is shown in Figure 3. With the aid of these contours, it is reasonable to visually inspect individual scatterplots 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 scatterplots to inspect (equal to *K(K-1)/2*). The contours are also useful in detecting outliers such as well 554 which falls significantly outside the 95% contour as shown in Figure 3.

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

Check for bivariate Gaussianity

The proposed check for bivariate Gaussianity 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 Gaussianity.

The first step counts the fraction of points falling in each of the 25%, 50% and 95% contours and compares them to the expected fraction. To check for this, the Mahalanobis distance, D_i^2, is calculated for each data pair (x_i, y_i) using Equation 3. 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 for α values of 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 check is optimized so a large deviation from bivariate normal behavior is flagged while small deviations possibly stemming from limited data are ignored.

$$D_i^2 = \frac{x_i^2 - 2\rho x_i y_i + y_i^2}{1-\rho^2} \tag{3}$$

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 given below (Equation 4). The sign of x_R and y_R are correlated to a quadrant using the scheme depicted in Figure 4.

$$\begin{bmatrix} x_R \\ y_R \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{\rho}{\sqrt{2}|\rho|} \\ \frac{-\rho}{\sqrt{2}|\rho|} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{4}$$

The maximum allowed deviation from 25% per quadrant is again optimized so that it reflects a departure from Gaussian behavior and not limited data available.

Measure of Deviation

The sum of small deviations is used to combine the two checks into a single measure of deviation from perfect bivariate normality. The deviations from the expected number in each quadrant for 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. Mathematically, this is shown in Equation 5 where *n* is the number of paired data points.

$$\Delta = \frac{1}{12}\sum_{i=1}^{3}\sum_{j=1}^{4}\left|\alpha_i - \frac{\text{Number} < \chi_2^2(\alpha_i) \text{ in quadrant } j}{n/4}\right| \tag{5}$$

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 BVSN. The 0.99 quantile value of Δ for each value of *n* was calculated and fit using a power model. The resulting equation and plot of simulated delta quantiles as a function of *n* are given in Equation 6 and Figure 5 respectively.

$$\Delta_{0.99} = 1.94917 n^{-0.49749} \tag{6}$$

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 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. To approximate 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. Equation 6 was updated to reflect the Gaussian behavior of the distributions and is normalized by $n^{-0.5}$ (Equation 7). The functions fit are provided in Equation 8.

$$\Delta_{0.99} = \frac{1.9492}{\sqrt{n}} \tag{7}$$

$$\Delta_{0.999} = \frac{2.3234}{\sqrt{n}}$$

$$\Delta_{0.9999} = \frac{2.6528}{\sqrt{n}} \tag{8}$$

These empirically determined quantiles are provided because the distribution of Δ values is significantly skewed for larger sample sizes. Histograms of simulated Δ values for *n* values of 10, 100 and 1000 are provided in Figure 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 s 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 Gaussianity is met and linearization by ACE or multivariate transformation by the stepwise conditional transformation may be necessary. We recommend a check scheme which uses a standardized value δ (Equation 9). The calculated value of δ is checked to see what range it falls in (Table 1) and the bivariate distribution is classified.

$$\delta = \frac{\Delta}{\Delta_{0.99}} = \frac{\Delta\sqrt{n}}{1.94917} \tag{9}$$

Table 1: Classification of δ based on evidence of nonGaussian behavior

Degree of Departure from BVSN	Range	Classification
0	δ < 1	Not enough evidence to assume nonGaussian
1, >99% nonBVSN	1 ≤ δ < 1.192	Bivariate is very likely nonGaussian and should be checked
2, >99.9% nonBVSN	1.192 ≤ δ < 1.361	Bivariate is extremely likely nonGaussian and should be checked
3, >99.99% nonBVSN	1.361 ≤ δ	Bivariate is nonGaussian and should be transformed

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

Implementation

The program scatnscores implements the plotting and checks discussed above. Standard GSLIB convention is used for the data file and parameter file. For details on the specific implementation and FORTRAN90 code, an Appendix is included.

A number of fabricated data sets (Figure 7) illustrating each of the principle behaviours responsible for nonGaussian distributions are plotted 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 non-linearity which results in a 1st degree departure from bivariate Gaussianity. The second and third data sets, illustrating hereoscedasticity and constraints respectively, are also flagged with 3rd degree departures from bivariate Gaussianity.

Case Study 1: Training Data

The first set of data studied is the familiar 2DWellData from Deutsch (2006). This data set was simulated from a multivariate Gaussian distribution for geostatistical training purposes so should not be flagged as nonGaussian. The matrix of cross plots generated by scatnscores is included (Figure 8) and there are indeed no cross plots flagged using the conditions above (recall Table 1).

Case Study 2: DV Well Data

The second data set checked was the DV_Well.dat set which is a real world data set with acoustic impedance, log porosity and log permeability data. The three bivariates, shown in Figure 9, all show significant deviations from Gaussian behavior. All three principle phenomena responsible for nonGaussian 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 non-linearity. The lower plot of log permeability vs log porosity shows significant heteroscedasticity.

Case Study 3: Red Data

The final set of data checked was the Red Dog data, red.dat. This data set included thickness, gold silver, copper and zinc contents. The cross plots are shown in Figure 10. There was no reason to reject any of the bivariate distributions as nonGaussian, possibly due to the limited number of data (62). For this data set the multivariate Gaussian distribution could likely be assumed, however, visible outliers in the cross plots should be checked.

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 a large amount of data pairs are available and is an efficient method for checking the assumption of bivariate Gaussianity. 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. and Gale, N.H., 1998, Testing for multivariate normality via univariate tests: A case study using lead isotope ratio data, *Journal of Applied Statistics*, 25, pp 671-683.

Deutsch, C.V., 2006, What in the Reservoir is Geostatistics Good For?, *Journal of Canadian Petroleum Technology*, 45, no 4, pp 14-20.

Deutsch, C.V. and Journel, A.G., 1998, *GSLIB: Geostatistical Software Library and User's Guide,* Oxford University Press, New York, 2nd Ed., pp 142-143.

Johnson, R.A. and Wichern, D.W., 2002, *Applied Multivariate Statistical Analysis,* Prentice-Hall, New Jersey, 5th Ed., pp 183-190.

Gnanadesikan, R., 1996, *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley and Sons, New York, 2nd Ed., pp 187-226.

Mardia, K.V., 1970, Measures of Multivariate Skewness and Kurtosis with Applications, *Biometrika*, 57, pp 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, pp 115-128.

Mardia, K.V., 1975, Assessment of Multinormality and the Robustness of Hotelling's T^2 Test, *Applied Statistics*, 24, pp 163-171.

Royston, J.P., 1983, Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W, *Applied Statistics*, 32, pp 121-133.

Thode, H.C. Jr., 2002, *Testing for Normality*, Marcel Dekker, New York, 1st Ed., pp 181-224.

Shapiro, S.S., Wilk, M.B., 1965, An Analysis of Variance Test for Normality (Complete Samples), *Biometrika*, 52, pp 591-611.

Shapiro, S.S., Wilk, M.B. and Chen, H.J., 1968, A Comparative Study of Various Tests for Normality, *Journal of the American Statistical Association*, 63, pp 1343-1372.

Wilks, D.S., 2006, *Statistical Methods in the Atmospheric Sciences*, Elsevier Academic Press, 2nd Ed, pp 440-443.

Appendix

Two programs written in Fortran90 code are detailed here. The first program, `nscore_MV`, normal score transforms a multivariate data set simultaneously. This program was pulled together by John Manchuk. It keeps a transfomation table so the data sets can be back-transformed after analysis. The second program, `scatnscores`, implements the plotting and checks discussed above. It plots the full matrix of cross plots and correlation coefficient matrix for a multivariate data set where each of the univariate distributions are Gaussian.

The program *nscore_MV* normal score transforms a specified set of variables to be univariate normal. Standard GSLIB convention is used for the data file and parameter file. An example parameter file is shown below.

```
1                Parameters for NSCORE_MV
2                ************************
3
4  START OF PARAMETERS:
5  nscores.dat                  -File with data
6  3                            -  number of variables to transform
7  5 6 7 8                      -  columns for variables
8  0 0 0                        -  columns for weights
9  -998.0 1.0e21                -  trimming limits
10 nscores_MV.dat               -File for output
11 nscore.trn                   -File for output transformation table
12 0  0  0  0                   -Transform according to ref. dist., column numbers
13 histsmth.out                 -  file with reference dist.
14 0  0  0  0                   -Transform according to ref. dist., column numbers
15 histsmth.out                 -  file with reference dist.
```

The parameter file for *nscore_MV* is similar to that of *nscores* detailed by Deutsch and Journel (1998). The data file is specified (line 5) and the number of variables and variable columns (lines 6, 7) are specified. If desired, the data can be weighted or trimmed (lines 8, 9). The transformed data and transformation table are output to the specified files (lines 10, 11). Source code for *nscore_MV* is available in the CCG software catalogue.

The program *scatnscores* implements the plotting and checks discussed in this paper. As with *nscore_MV*, standard GSLIB convention is used. An example parameter file is given below. Line 6 specifies the number of variables to cross plot. Line 7 specifies the columns of data containing each of the variables and if desired, line 8 the data ID numbers. If you do not wish to have the points labeled with the data ID, line 8 is left as 0. The image output file and bullet size (lines 9, 10) can also be specified.

```
1                    Parameters for SCATNSCORES
2                    **************************
3
4  START OF PARAMETERS:
5  nscores.dat                        -file with data
6  4                                  -  number of variables
7  12 14 15 16                        -  columns for variables
8  0                                  -  column  for data ID
9  scatnscores.ps                     -file for Postscript output
10 0.5                                -bullet size: 0.1(sml)-1(reg)-10(big)
```

Figure A2: An example parameter file for CCG program *scatnscores* which implements the described plots and checks

The program will plot the full matrix of cross plots automatically. This is a (*K*-1)x(*K*-1) matrix since the cross plot of a variable against itself is not useful. For each cross plot, the check statistic, δ, is calculated and flagged if it is greater than or equal to 1.0 (see Table 1). The correlation coefficient matrix is calculated and displayed. If a δ value is calculated which exceeds 1, then the degree of nonGaussian behavior (recall Table 1) is indicated on the cross plot.

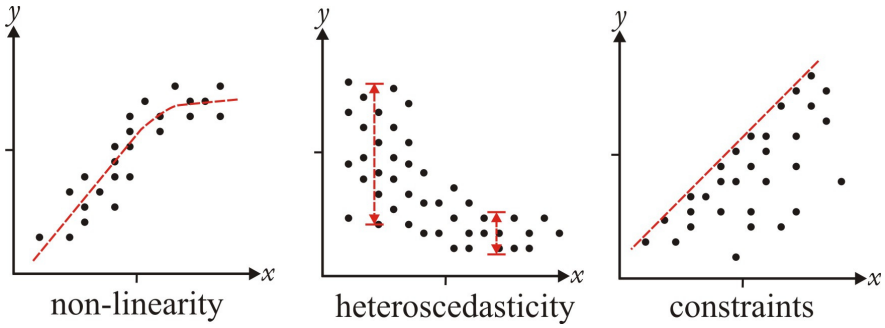


Figure 1: The three most important sources of nonGaussian behavior: non-linearity, heteroscedasticity and constraints

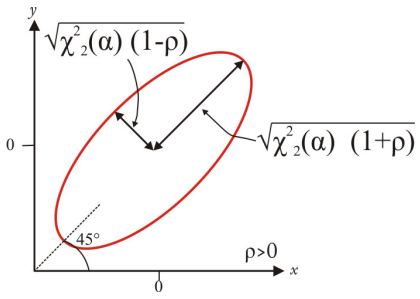


Figure 2: constant density ellipse ranges depend on the chi-squared value and the correlation coefficient

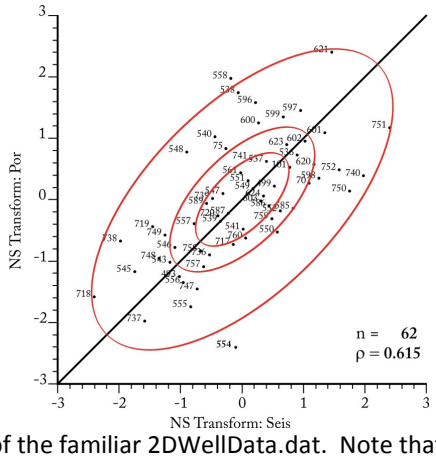


Figure 3: An example cross plot of the familiar 2DWellData.dat. Note that well 554 has an unusually low porosity for the seismic value and should probably be checked

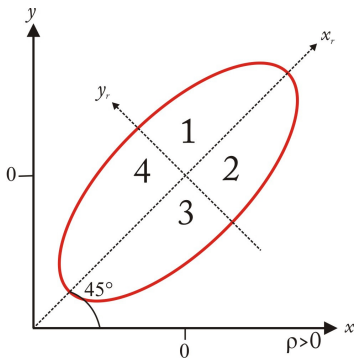


Figure 4: Quadrants of the constant density ellipse relative to the rotated axes; x_r and y_r

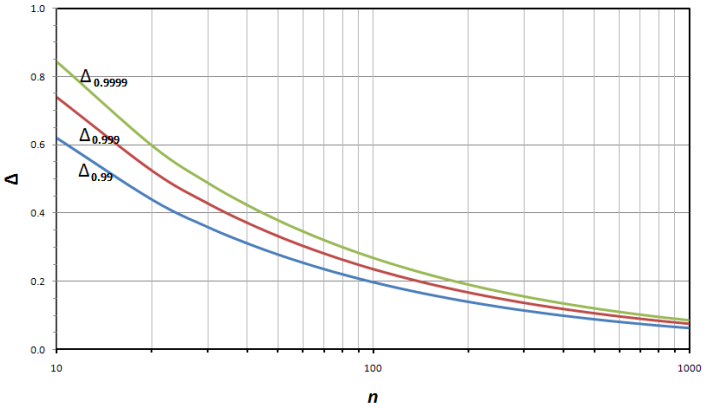


Figure 5: Simulation of Δ values from a BVSN for various *n*. The upper curve is the 0.9999 quantile, the middle is the 0.999 quantile and the lower curve is the 0.99 quantile.

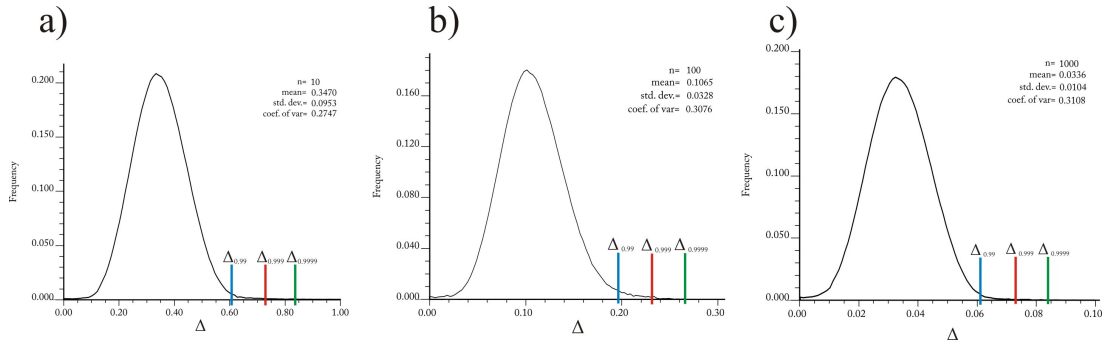


Figure 6: Distribution of Δ values for various *n*. Histogram a) corresponds to *n*=10, b) *n*=100 and c) *n*=1000. Each histogram has a slightly skewed shape as Δ is bounded on the left by 0.

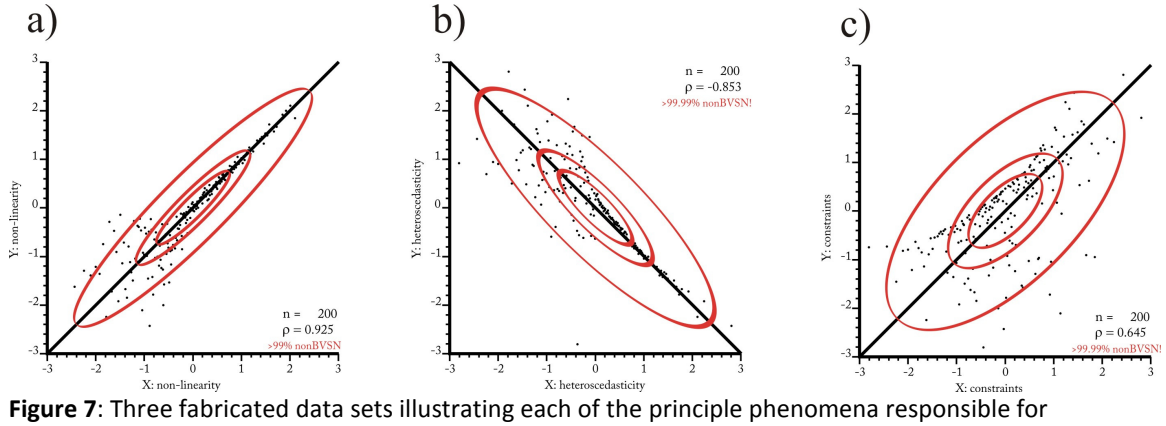
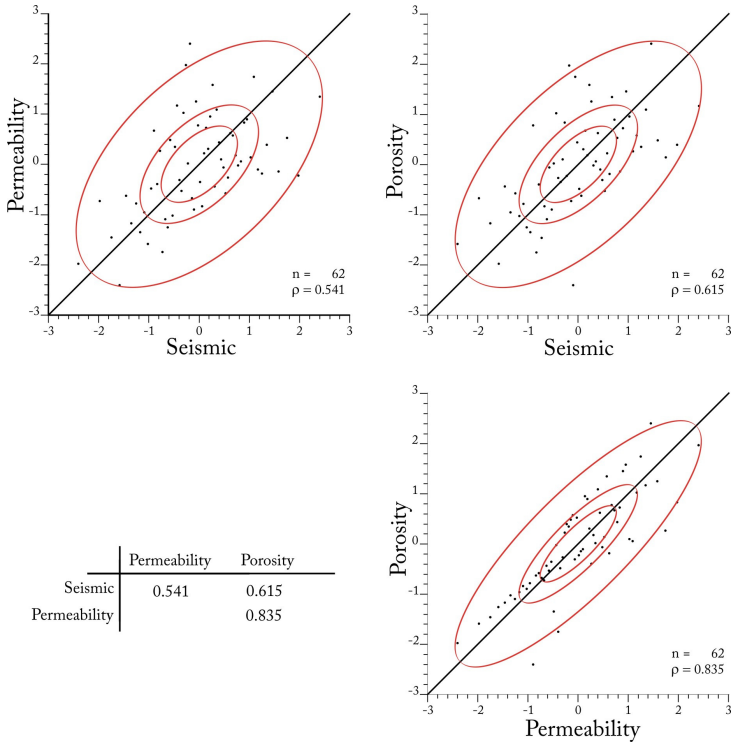
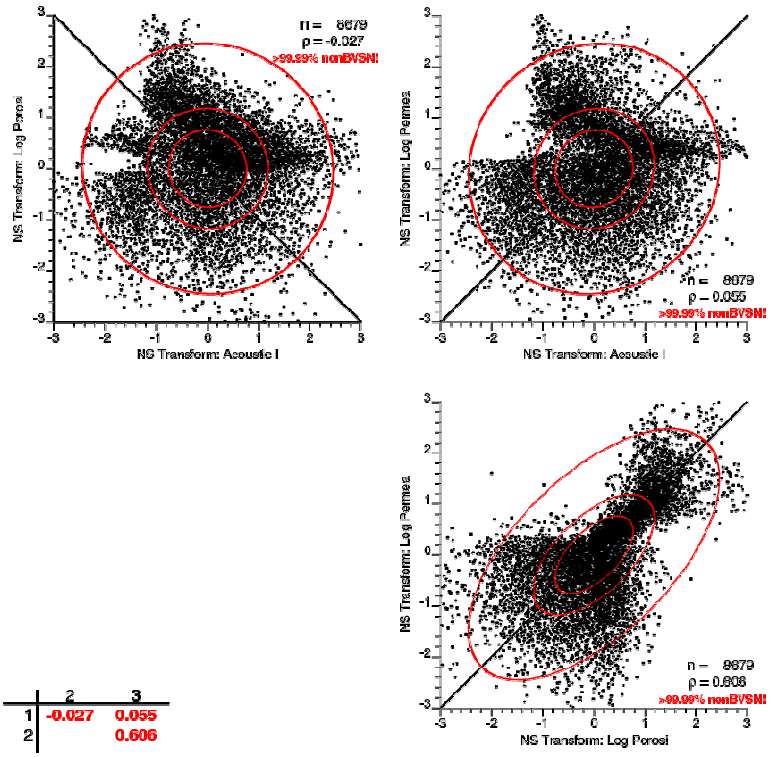


Figure 7: Three fabricated data sets illustrating each of the principle phenomena responsible for nonGaussian behavior: a) non-linearity, b) heteroscedasticity, c) constraints



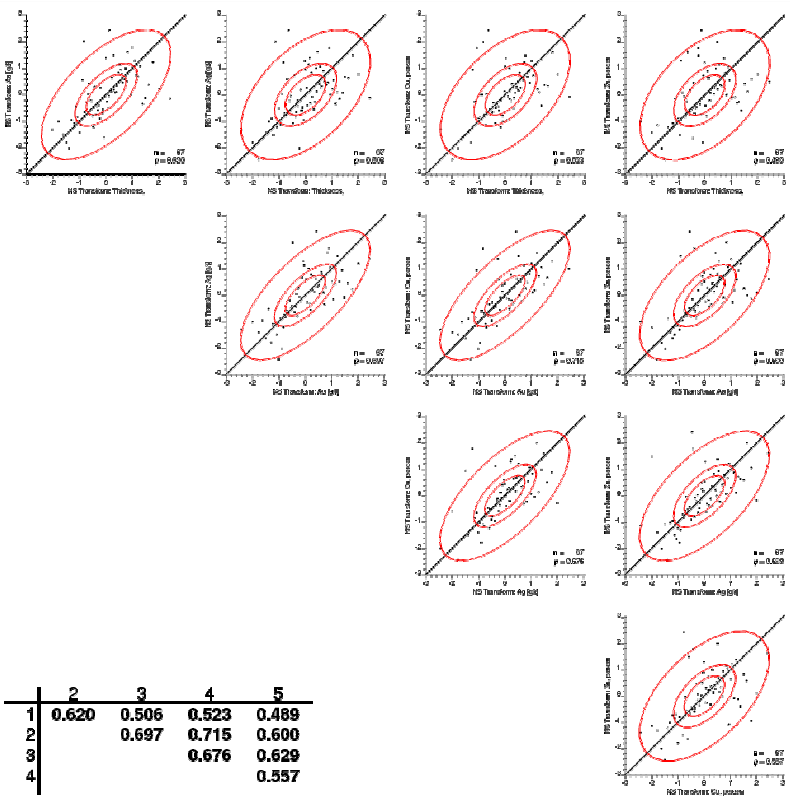
	Permeability	Porosity
Seismic	0.541	0.615
Permeability		0.835

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



	2	3
1	-0.027	0.055
2		0.606

Figure 9: Cross plots of DV well data. All bivariates exhibit significant nonGaussianity, likely stemming from one or more of the three principle phenomena responsible for nonGaussianity (Figure 7)



	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

Figure 10: Cross plots of the red.dat data. No bivariate exhibits significant nonGaussian behavior, however outliers are easily discernable using the 95% ellipse.