SHORT NOTE

MultiGaussian Model for Uncertainty in a Non-Spatial Context

C. V. Deutsch
University of Alberta, Edmonton, Alberta, CANADA (cdeutsch@civil.ualberta.ca)

A. S. Cullick
Landmark Graphics Corp., Austin, Texas USA (scullick@lgc.com)

A common problem is uncertainty estimation and calculation of a best estimate in presence of related secondary data. The multiGaussian distribution provides a simple and powerful model to predict uncertainty in this context. This short note provides documentation and implementation details.

Setting of the Problem

Consider N variables that are related to each other: Z^i, $i=1,\ldots,N$, where a superscript i is used because these are N different variables and not N different observations of the same variable. These variables could be anything including petrophysical properties, economic quantities, fluid characteristics, or the same variable at different spatial locations. Multiple measurements of the different variables must be available to inform the N histograms. Either data or expert knowledge must also be available to determine the correlation between each pair of variables. The problem is to predict the uncertainty in 1 (or more) variables values given knowledge of some subset of the other N-1 variables.

$$Z^{i*} \quad \text{and} \quad F_i(z^i|\text{available data})$$

The multiGaussian probability distribution and the well established normal equations provide a convenient model for prediction in this context. The steps required are to (1) transform each variable one-at-a-time to a Gaussian histogram, (2) establish the correlation coefficients that define the multivariate Gaussian distribution, (3) use the normal equations for prediction of the mean and variance of the Gaussian transforms, and (4) back transformation to original variable units and calculation of statistics such as the mean.

Normal Score Transform

The normal scores transformation takes data following any arbitrary probability distribution and transforms them to have the standard normal distribution. The standard normal probability distribution is the well known bell shaped distribution following:

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{y^2}{2}} \tag{1}$$

Figure 1 illustrates the normal scores (a.k.a. *quantile* or *graphical*) transformation procedure of a set of core porosity data to the standard normal distribution. Note that although the cumulative

distribution for the standard normal distribution has no closed form analytical solution, there are excellent numerical approximations. The `nscore` program in GSLIB implements normal scores transformation. In mathematical notation the transform of the Z^i, $i=1,\ldots,N$, variables is written:

$$y_j^i = G^{-1}\left(F_i\left(z_j^i\right)\right), \; i=1,\ldots,N, \;\; j=1,\ldots,n_i \tag{2}$$

where y_j^i is the normal score transform of the j^{th} observation of the i^{th} variable, $G^{-1}(\bullet)$ is the inverse of the cumulative Gaussian distribution, $F_i(\bullet)$ is the cumulative distribution of the i^{th} variable, z_j^i is the j^{th} observation of the i^{th} variable, N is the number of variables, and n_i is the number of observations of the i^{th} variable

Each variable is independently transformed to a normal distribution. The transformation can be reversed at any time by matching quantiles (see the bottom of Figure 1 and the `backtr` program in GSLIB). The back transformation of any particular normal score value y_j^i is accomplished by:

$$z_j^i = F_i^{-1}\left(G\left(y_j^i\right)\right) \tag{3}$$

This transformation / back transformation is straightforward and could be performed by any number of different procedures. The normal score transformation is general and widely used except when there are very few data. In presence of few data, the distributions F_i, $i=1,\ldots,N$ could be fit by an appropriate parametric distribution model.

Cross Correlation Coefficients

A multivariate Gaussian probability distribution can be adopted for the Y^i, $i=1,\ldots,N$, variables that are each univariate Gaussian or normal. There are a number of tests for multiGaussianity. These are recommended; however, rarely used in practice since there is no simple alternative if the multiGaussian model is rejected. These tests will be documented in another short note. Let's proceed under the multiGaussian model.

An important feature of the multiGaussian distribution model for N variables is that the full distribution is defined by the N by N matrix of correlation coefficients:

$$\begin{bmatrix} 1 & \cdots & \rho_{1,j} & \cdots & \rho_{1,N} \\ \vdots & \ddots & & & \vdots \\ \rho_{j,1} & & 1 & & \rho_{j,N} \\ \vdots & & & \ddots & \vdots \\ \rho_{N,1} & \cdots & \rho_{N,j} & \cdots & 1 \end{bmatrix} \tag{4}$$

This matrix of correlation coefficients is symmetric, that is, $\rho_{ij}=\rho_{ji}$. The diagonal elements are $\rho_{ii}=1$ for standard Gaussian variables. Thus, there are $N(N-1)/2$ required correlation coefficients. These values can be calculated directly from available data or inferred from an understanding of the variables (for example, porosity and permeability are expected to have a positive correlation of, say, 0.7).

The $N(N-1)/2$ correlation coefficients cannot be set independently; they must be jointly positive definite. Mathematically, the determinant of matrix (4) must be non-zero. Positive definiteness is not restrictive; it is quite reasonable. Consider a three variable example that would *not* be positive definite: $\rho_{1,2}=0.85$, $\rho_{2,3}=0.90$, and $\rho_{3,1}=-0.5$. Variables 3 and 1 simply cannot be negatively correlated if 1 and 2 are positively correlated and 2 and 3 are positively correlated?

The N by N matrix (4) must be established by calculations from data (see Figure 2 for two example normal score cross plots) or from expert judgment. Then, the matrix of coefficients

must be checked for positive definiteness. An iterative procedure is normally considered to modify the correlation coefficients to ensure positive definiteness if the determinant does not come out to be positive.

Prediction of Uncertainty

Recall the original problem: predict the uncertainty in 1 (or more) variables values given knowledge of some subset of the other *N*-1 variables. The prediction of uncertainty under the multiGaussian model is exact. For the sake of simple notation, let's make the variable number we are predicting 0 and order the data variables 1 through *n*, where $n < N$. The estimate of the normal scores mean is given by a weighted linear estimator:

$$y^* = \sum_{i=1}^{n} \lambda_i \cdot y_i \tag{5}$$

The weights λ_j, $j=1,\ldots,n$ are given by the well-known normal equations:

$$\sum_{j=1}^{n} \lambda_j \cdot \rho_{i,j} = \rho_{i,0} \quad i = 1,\ldots,n \tag{6}$$

The *n*-by-n system of linear equations must be solved to determine the weights, which are used to calculate the estimate (4) and the estimation variance:

$$\sigma_E^2 = 1 - \sum_{i=1}^{n} \lambda_i \cdot \rho_{i,0} \tag{7}$$

The uncertainty in the estimate y^* is completely defined in "Gaussian space;" the uncertainty follows a normal distribution with mean of and variance of σ^2_E. This distribution of uncertainty must be transformed back to real *Z* units.

The back transformation of the non-standard normal distribution $N(y^*,\sigma^2_E)$ is accomplished using the back transformation equation (3) to transform a large number of quantiles. Note that we cannot back transform the mean y^* by equation 3; the back transformed *z* value would not be a biased estimate of the mean of *Z* since the transformation is non-linear. Quantiles can be back transformed with no bias; therefore, the following procedure is used to establish the distribution of uncertainty in *z*-units.

1. Choose *L* equally spaced quantiles for back transformation, for example, the L=99 equally spaced percentiles p_1=0.01, p_2=0.02,…,p_{99}=0.99.
2. Calculate the Gaussian or normal deviate for each probability value:

$$y_l = y^* + G^{-1}(p_l) \cdot \sigma_E \qquad l = 1,\ldots,L$$

3. Back transform each normal deviate using equation 4, that is:

$$z_l = F_i^{-1}(G(y_l)) \qquad l = 1,\ldots,L$$

4. Assemble the distribution of uncertainty and expected value of the *Z* values, the variance of the *Z* values and any other desired statistic.

$$z^* = \frac{1}{L}\sum_{l=1}^{L} z_l \qquad\qquad \sigma_z^2 = \frac{1}{L}\sum_{l=1}^{L} \left[z_l - z^*\right]^2$$

The L back transformed values are equally probably because the probability values $p_l, l=1,\ldots,L$ are equally spaced. The software to perform these calculations is straightforward. The building blocks exist in the GSLIB programs `nscore` and `backtr`.

It is important to realize that a multiGaussian probability distribution is assumed after normal score or Gaussian transformation; however, that does not mean that the resulting distributions of uncertainty are Gaussian. Figure 3 attempts to illustrate this point. The distributions are non-standard Gaussian, $N(y^*,\sigma^2_E)$, in transformed Gaussian space, but account for the correct histogram in original data units.

An Example

Consider four variables v_1, v_2, v_3, and v_4 with arbitrary non-Gaussian distributions of uncertainty; see Figure 4. The matrix of correlation coefficients in these variables after normal score transform is given below:

$$\begin{bmatrix} 1 & 0.8 & 0.7 & -0.5 \\ & 1 & 0.6 & -0.6 \\ & & 1 & -0.3 \\ & & & 1 \end{bmatrix} \tag{4}$$

There are four situations where we could be estimating one of the variables from the other three (no missing data of the other three). The weights assigned to the other three data are given below. These weights are calculated by solving equation (6) with three data.

Estimating v_1		Estimating v_2		Estimating v_3		Estimating v_4	
v_2	0.552	v_1	0.590	v_1	0.617	v_1	-0.143
v_3	0.350	v_3	0.105	v_2	0.173	v_2	-0.572
v_4	-0.064	v_4	-0.274	v_4	0.113	v_3	0.143

The combinatorial of all possible subsets will not be considered. The equations for any specific subset can be solved as needed.

Consider the task of estimating variable 1 with data on variables 2, 3, and 4 of 3.00, 1.50, and 6.00. The following steps are followed to get the distribution of uncertainty, mean, and variance for variable 1:

1. Determine the normal score transforms of the three data values: (3.00 $\rightarrow$ 1.745, 1.5 $\rightarrow$ 0.648, and 6.0 $\rightarrow$ -2.108).
2. Calculate the mean and variance in normal space:

$$y^* = 0.552 \bullet 1.745 + 0.350 \bullet 0.648 - 0.064 \bullet -2.108 = 1.325$$

$$\sigma^2_E = 1 - 0.552 \bullet 0.8 + 0.350 \bullet 0.7 - 0.064 \bullet -0.5 = 0.281$$

3. Back transform quantiles of the non-standard normal distribution (1.325/0.281) to get the uncertainty in variable 1 (see Figure 6) and calculate the mean (7.0), standard deviation (0.8), and other needed statistics.

This procedure is easily automated.

Conclusion

Given *N* correlated variables, a procedure has been shown to calculate the uncertainty in estimating one or more of them from a subset of the remaining known variables. This classic Gaussian model has wide applicability in simplified Monte Carlo analysis where there are inadequate data to inform the probability law more completely.

References

- Deutsch C.V., Journel A.G. (1998) *GSLIB: Geostatistical Software Library: and User's Guide* Oxford University Press, New York, 2nd Ed.
- Any classic statistics book that covers regression and probability distributions.

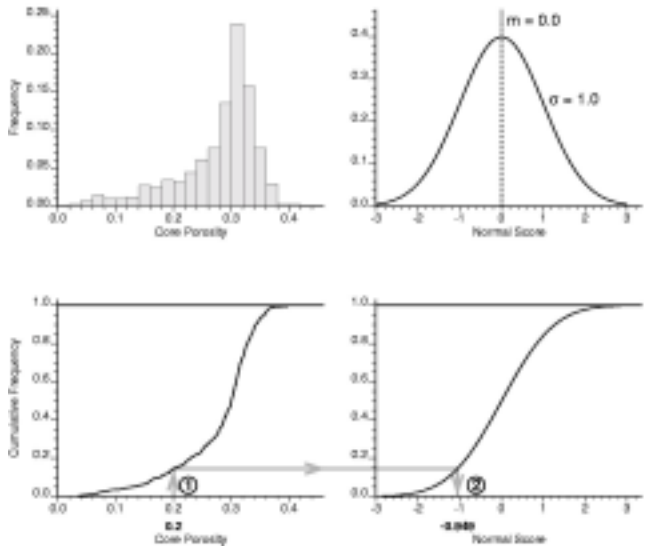

Figure 1: the normal score transform. The upper figures show an arbitrary histogram of core porosity values and the target normal distribution. The lower figures show the corresponding cumulative distributions and the transformation procedure, which consists of matching quantiles (values with the same cumulative probability.

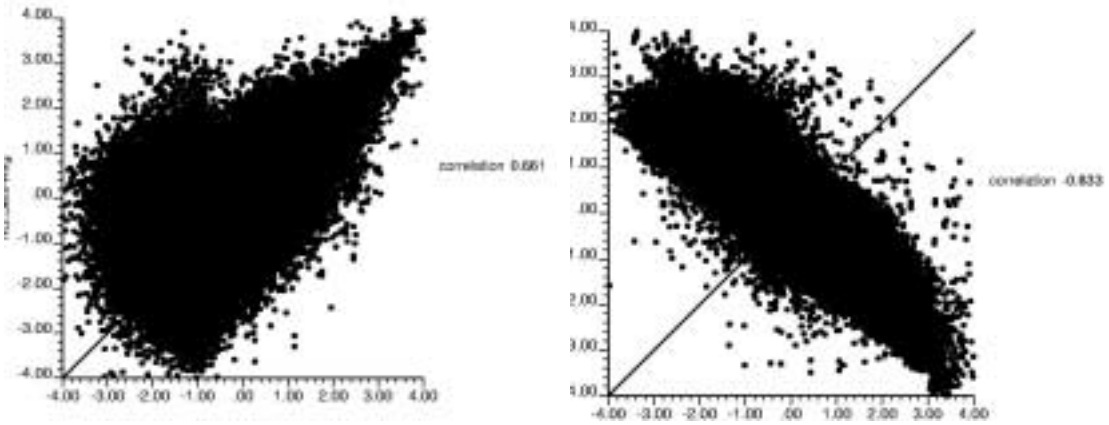

Figure 2: two cross plots of normal score transformed variables with correlation coefficients of 0.661 and –0.883. There are an unusually large number of points on these cross plots.

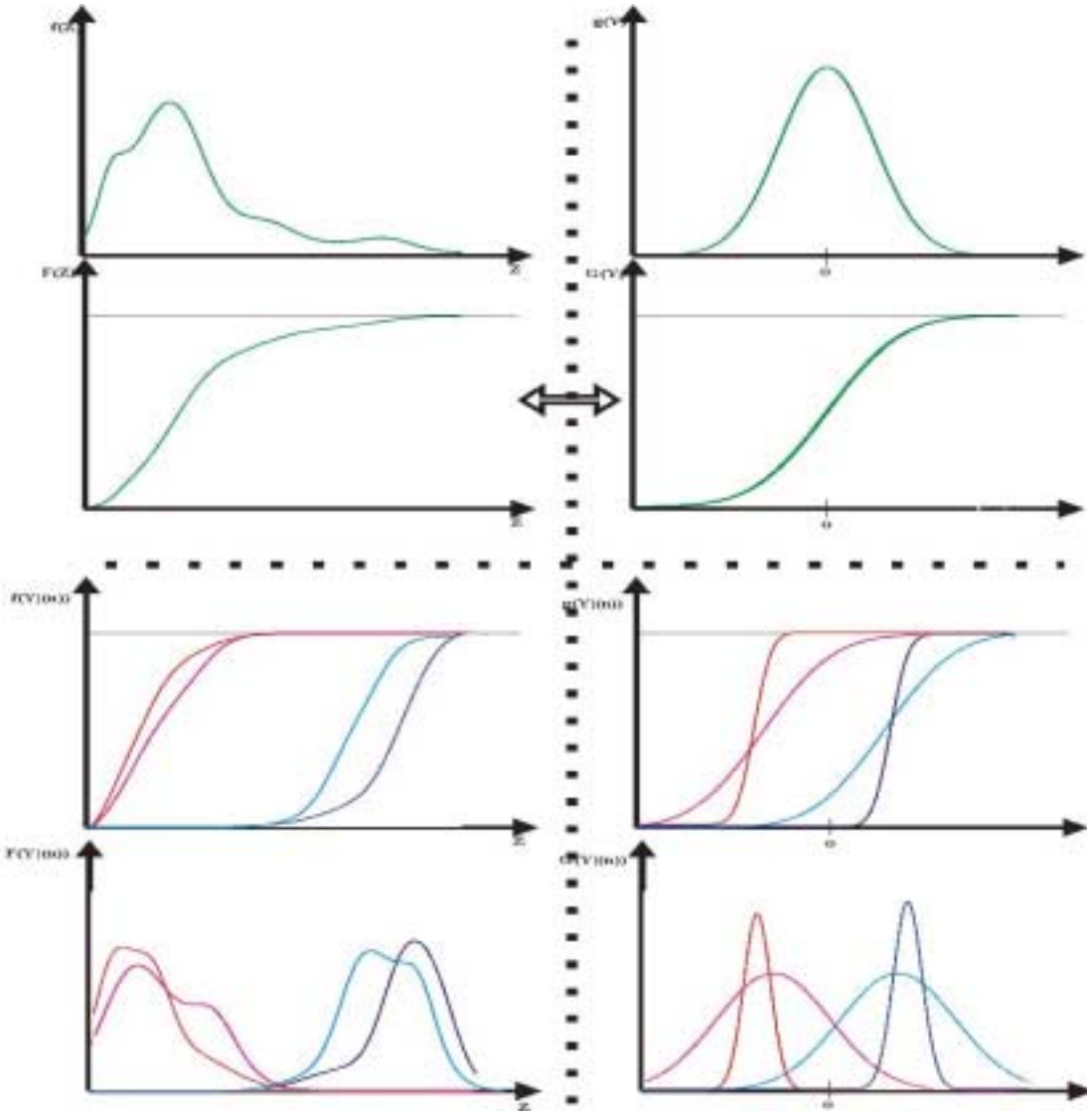

Figure 3: illustration of uncertainty in normal space (right side) and in real data units (left side). The top row shows (in green) an original z-histogram *f(z)* and corresponding normal distribution *g(y)*. The cumulative distributions *F(z)* and *G(y)* are shown (in green) on the second row from the top. The two-sided arrow from *z*-space to *y*-space illustrates that this is the link between real data units and Gaussian data units. The four conditional cumulative distributions (multicolored) on the third row from the top are different distributions with different mean and variance values. The shapes of these distributions are Gaussian on the right and some other shape, determined by the back transformation, on the right. The bottom row shows the distributions as histograms of probability density functions.

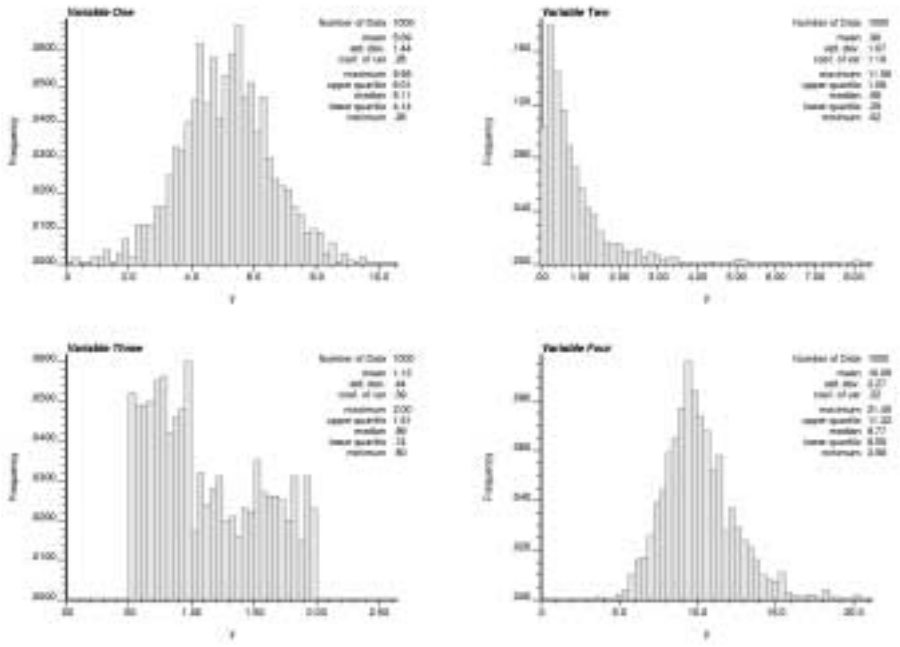

Figure 4: histograms of four variables for example. Aside from the first variable, none of the histograms are Gaussian.

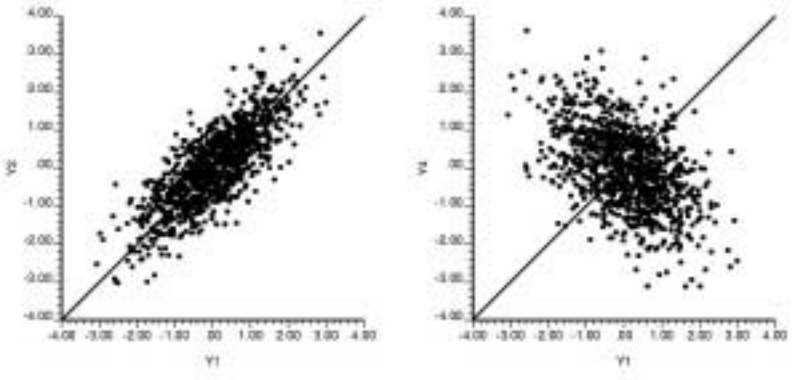

Figure 5: scatterplots between the Gaussian transforms of the first and second and first and fourth variables.

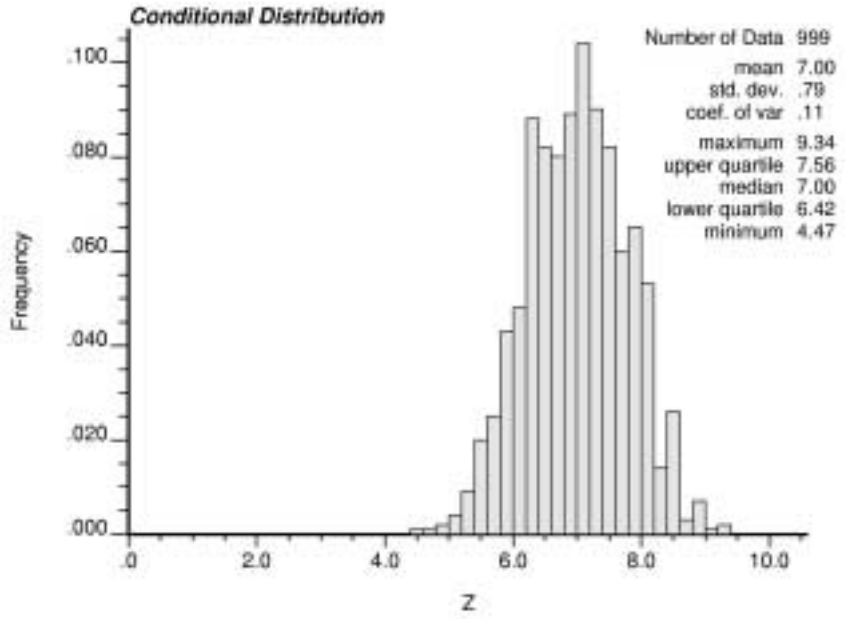

Figure 6: conditional distribution of variable 1. The distribution in Gaussian space has a conditional mean of 1.325 and a conditional variance of 0.281. This distribution is determined by back transformation.