Order Relations Correction and Tail Extrapolation for Stepwise Conditional Transformation

Clayton V. Deutsch

Centre for Computational Geostatistics (CCG)
Department of Civil and Environmental Engineering
University of Alberta

The need for many data hinders reliable implementation of stepwise conditional transformation for multivariate geostatistical simulation. Enforcing order relations consistency in the multivariate distribution and permitting flexible tail extrapolation for the conditional distributions improves the reliability of the transformation and back transformation. Updated transformation and back transformation software is presented.

Introduction

In many cases, covariance-based methods like full cokriging, collocated cokriging or Bayesian updating work well to account for multivariate relationships; however, stepwise conditional transformation (SCT) is particularly useful in cases where multivariate constraints, non-linear behavior or heteroscedasticity are important features of the multivariate relationship. Constraints may be important, for example, when accounting for a trend or when accounting for mineralogical data. Non-linear behavior can be important, for example, with flow-related variables – some porosity-permeability relationships show a characteristic non-linear relationship where the permeability flattens off at high porosity values. Heteroscedasticity is important, for example, with some remotely sensed variables – the relationship between porosity and acoustic impedance may be more variable with low impedance values.

The stepwise conditional transformation technique was first introduced by Rosenblatt (1952). Leuangthong (2003) worked significantly with this transform in her Ph.D. thesis and popularized its use in modern geostatistics. The technique is identical to the normal score transform in the univariate case. The normal scores transformation of the second variable is conditional to the probability class of the first variable. The transform for the third variable is conditional to the first two and so on:

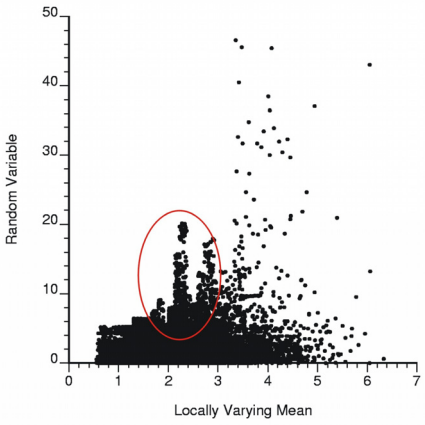
$$y_1 = G^{-1}(F_1(z_1))$$

$$y_2 = G^{-1}(F_{2|1}(z_2|z_1))$$

$$y_3 = G^{-1}(F_{3|1,2}(z_3|z_1,z_2))$$

The resulting Y variables have univariate Gaussian distributions and the collocated values are independent. This greatly facilitates cosimulation. The variables are independently simulated and the values are back transformed in reverse order. The multivariate relationships of collocated values are reproduced in the back transformation. The increasing use of SCT has revealed some important implementation details. Correcting the distributions to be monotonically increasing or decreasing and permitting more flexible tail extrapolation is important.

Order Relations

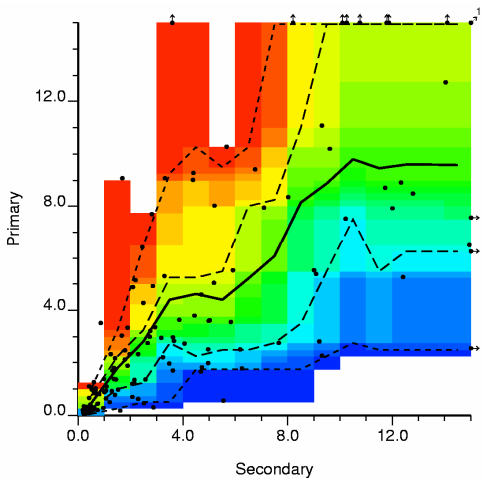


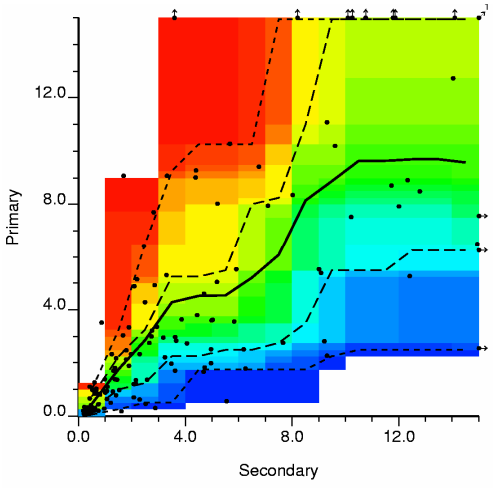
The problem considered here is not order relation deviations in the conventional sense. Too few data and outlier high values cause the conditional distributions to be erratic. The example at the right is the back transformed results for an example of using stepwise for trend modeling. There were nearly 3000 data in the original calibration and only 20 classes were used. Nevertheless, there are artifacts, particularly in the tails of the distribution. The importance of these artifacts is not well understood. They certainly have a visual impact on the cross plot and undermine the credibility of the model. In many cases, I believe they will have no affect whatsoever on decisions made with the final model. Nevertheless, the problem of artifacts and nonphysical relationships should be corrected. This problem becomes much worse for the third and subsequent variables because conditioning to two or more variables significantly decreases the available data for any particular class.

The direction of the increasing or decreasing correction is based on the sign of the covariance between the data. The discussion below refers to increasing distributions; however, the implementation also works with decreasing distributions.

The key idea for correcting the distributions is to use a procedure similar to that used in GSLIB for correcting distributions from indicator kriging: (1) correcting the distributions in an upward direction, (2) correcting them in a downward direction, and then (3) averaging the result. The average of two non-decreasing functions is another non-decreasing function. The average of non-increasing functions is also non-increasing.

The example shown below is from the 140 `cluster.dat` data. The cross plot on the left shows the conditional distributions of the data. The conditional mean values are connected by a solid black line in the middle. The 0.25 and 0.75 quantiles are shown by long dashed black lines. The 0.05 and 0.95 quantiles are shown by the dashed lines. The cross plot on the right shows the results after quantiles are enforced to be increasing.





The methodology is easily understood visually. Here are more details. The set of conditional distributions used for the second variable are denoted:

$$F_{Z_2|Z_1=z_{1,k}}(z_2),\ k=1,...,K$$

There are K conditional distributions. The thresholds defining the Z_1 classes are often chosen so that there is the same number of data in each class. To facilitate the correction, all conditional distributions will be represented by a fixed number of quantiles. Experience shows that n_q=200 evenly spaced quantiles work well. The variance of a skewed conditional distribution may be underestimated with fewer quantiles. The probability values:

$$p_i = \frac{i}{n_q+1},\ i=1,...,n_q$$

The $K \bullet n_q$ quantiles derived from the available data at the specified probability values are:

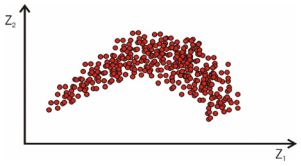
$$z_{2,i,k},\ i=1,...,n_q,\ \ k=1,...,K$$

These quantiles will also account for the tail options discussed in the next section. The correction proceeds in three steps:

$$\left.\begin{aligned} z_{2,i,k}^{(upward)} &= \max\left(z_{2,i,k-1}^{(upward)}, z_{2,i,k}\right), \text{proceed upward}: k=2,...,K \\ z_{2,i,k}^{(downward)} &= \min\left(z_{2,i,k+1}^{(downward)}, z_{2,i,k}\right), \text{proceed downward}: k=K-1,...,1 \\ z_{2,i,k}^{(corrected)} &= \frac{z_{2,i,k}^{(upward)} + z_{2,i,k}^{(downward)}}{2},\ \ k=1,...,K \end{aligned}\right\},\ i=1,...,n_q$$

The "max" and "min" are switched if the correction is for a decreasing bivariate distribution.

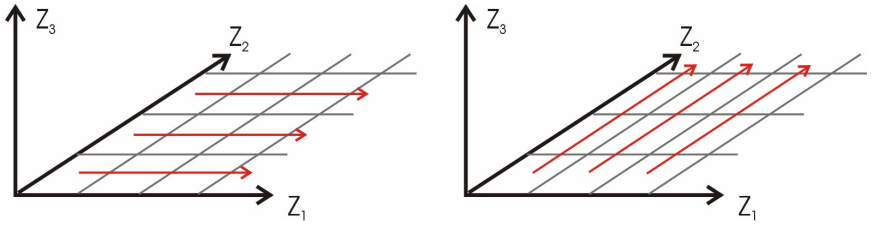
There are bivariate distributions that are not monotonically increasing or decreasing, see the sketch to the right. There may be value in smoothing the quantiles to reduce unwarranted noise; however, this correction will have to be modified for use in cases where the distributions are not monotonic.



The methodology for the third variable is essentially the same. The set of conditional distributions used for the third variable are denoted:

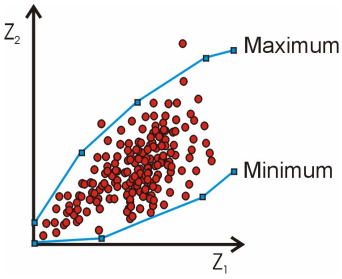
$$F_{Z_3|Z_1=z_{1,k},Z_2=z_{2,k}}(z_3),\ k=1,...,K$$

The upward/downward correction described above proceeds along all Z_1 values and then Z_2 values. The covariance between Z_3 and Z_1 and Z_2 and Z_1 is calculated to judge whether the correction should be upward or downward.



Tail Extrapolation

There are two main reasons for setting minimum and maximum limits for the conditional distributions: (1) there may be spurious outlier data that should be clipped, and (2) more reliable back transformation when we are simulating many more locations than we have original data. The back transformation according to conditional distributions requires a full specification of all conditional distributions for p values from 0 to 1. Some of the p values may be very close to 0 or 1 and fall outside the range available from the data. An extrapolation is required.



The indicator programs IK3D and SISIM in GSLIB permit flexible tail extrapolation because there may be less than 10 thresholds. The power law, hyperbolic and tabulated quantile options are not suitable for stepwise transformation: (1) there are not enough data to infer the additional parameters, and (2) they are somewhat unstable permitting extreme values in the back transformation. A straightforward linear interpolation to minimum and maximum values has been implemented.

The minimum/maximum values that specify the lower and upper tails are simply 2 numbers for the first variable Z_1, they are curves for the bivariate case of Z_2 (as shown in the sketch above right) and they are surfaces for the trivaraite case of Z_3. A piecewise linear interpolation is used for the bivariate curves and an inverse distance interpolation is used to estimate the minimum/maximum surfaces in the trivariate case.

The interpolation of the minimum/maximum surfaces in the trivariate (and higher order) cases must account for the units of the data. The distance for the inverse distance weighting uses the minimum and maximum values for each variable to specify the anisotropy:

$$d_i = \sqrt{\left(\frac{z_1 - z_{1,i}}{z_{1,\max} - z_{1,\min}}\right)^2 + \left(\frac{z_2 - z_{2,i}}{z_{2,\max} - z_{2,\min}}\right)^2}$$

This distance is calculated for all i pairs specified by the user and a conventional inverse squared distance estimation is used. The squared distance is used to ensure that the local control points are given a large weight.

For simplicity in the user input, the minimum and maximum must be specified for all control values, that is, the current software does not allow different Z1 control values when the Z2 maximum values are specified. The sketch above to the right shows five control points for the maximum and four for the minimum, which is not valid input. There are three tables of input values as illustrated below in three tables:

Z_1 Tails	
Lower	Upper

Z_2 Tails		
Z_1 value	Lower	Upper

Z_3 Tails			
Z_1 value	Z_2 value	Lower	Upper

Outlier data and the minimum and maximum values are fixed before the conditional distributions are established and the transform established.

Modified Computer Code

Modified programs are available to implement the stepwise conditional transformation and back transformation: `sctrans` and `scback`. The version number has been incremented to 2.000. There is an indicator flag for the correction of order relations and tabulated minimum/maximum values. The new parameters:

```
1                              -apply order relations corrections (1=yes)
0.0    1.2                     -min/max values for first variable
2                              -number of min/max values for second variable
0.0  0.00 0.005                -   second tail: Z1, Z2min, Z2max
1.2  0.01 0.200                -   second tail: Z1, Z2min, Z2max
1                              -number of min/max values for third variable
0.5  0.2  0.01  0.40           -   third tail: Z1, Z2, Z3min, Z3max
```

In practice, there would be more than just one or two control points for the second and third variable tails; this is just an example.

Examples

The first example is from a set of molybdenum and copper data. Figure 1 shows the original data and the results of back transforming 100,000 simulated Gaussian values. There were 1505 original data and 20 classes were considered which gives about 75 data per class. The results on Figure 1 are based on the conventional approach with no order relations correction and no limiting by tails. Figure 2 shows the results with the order relations turned on – note that the conditional distributions systematically increase. There are fewer points to the far right of the plot (high Cu values), therefore fewer high Mo values. Figure 3 shows the results when upper and lower tail values were imposed on the distribution. A summary of the statistics reproduction is shown in the table below. The minor difference in the mean values for the classical stepwise conditional transform values is due to a slight departure from standard normality and the class discretization. The correlation increases somewhat when order relations and the tails are corrected.

		Results of Stepwise Back Transformation		
	Original Data	Conventional	Order Relations	Tails/Order
Cu mean	0.365	0.362	0.362	0.362
Mo mean	0.027	0.026	0.029	0.025
ρ	0.239	0.256	0.214	0.405
ρ - rank	0.507	0.515	0.520	0.536

A second example with log K and porosity data from a North Sea reservoir was considered. Figure 4 shows the original data and the results of back transforming 100,000 simulated Gaussian values. There were 3725 original data and 20 classes. Figure 5 shows the results with the order relations turned on. The statistics are reproduced very closely in all cases, see below.

		Results of Stepwise Back Transformation	
	Original Data	Conventional	Order Relations
ϕ mean	7.656	7.636	7.636
Log K mean	0.421	0.416	0.416
ρ	0.689	0.686	0.689
ρ - rank	0.714	0.712	0.713

A number of other examples were considered. The order relations correction led to visually improved results in all cases. The tails extrapolation limits are particularly useful with very erratic data.

Conclusions

Implementation details are important. This research note discusses some important aspects for stepwise conditional transformation (SCT) in presence of relatively sparse data. The *order relations* correction amounts to smooth the conditional distributions and permits more reliable (at least stable) inference of conditional distributions when there are limited calibration data. The tail extrapolation options to user specified minimum and maximum values also permits more stable inference of the conditional distributions and leads to more reliable back transformation when there are many simulated values.

References

- Leuangthong, O., *Stepwise Conditional Transformation*, Ph.D. Thesis, University of Alberta, 2003.
- Leuangthong, O., and Deutsch, C.V., "Transformation of Residuals to Avoid Artifacts in Geostatistical Modelling with a Trend," *Mathematical Geology,* accepted September 2003.
- Leuangthong, O., and Deutsch, C.V., "Stepwise Conditional Transformation for Simulation of Multiple Variables," *Mathematical Geology,* Vol.35, No. 2, pp. 155-173.

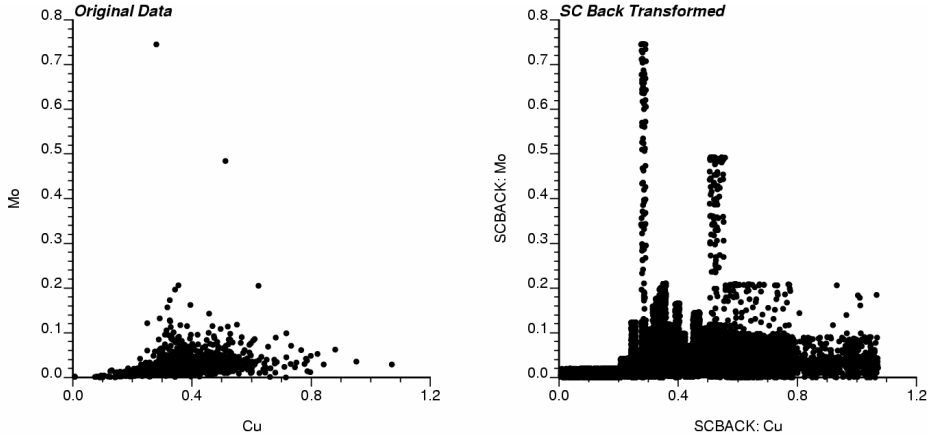


Figure 1: Cross plot of 1505 original molybdenum (Mo) and copper (Cu) data (on the left) and the result of simulating 100,000 locations (on the right).

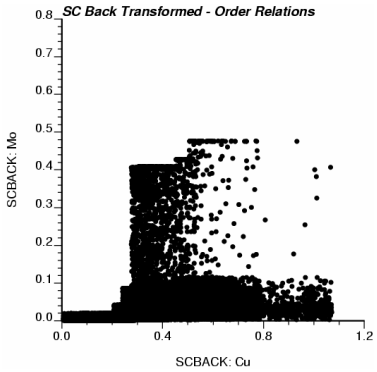


Figure 2: Cross plot of backtransformed 100,000 values when the original conditional distributions have order relations corrected.

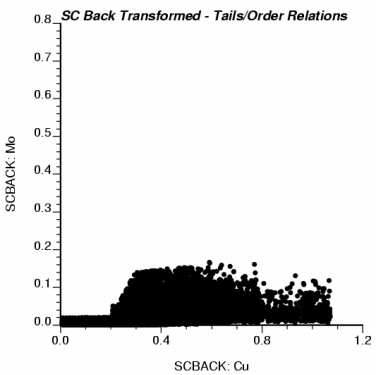


Figure 3: Cross plot of backtransformed 100,000 values when the original conditional distributions had the tails imposed and order relations corrected.

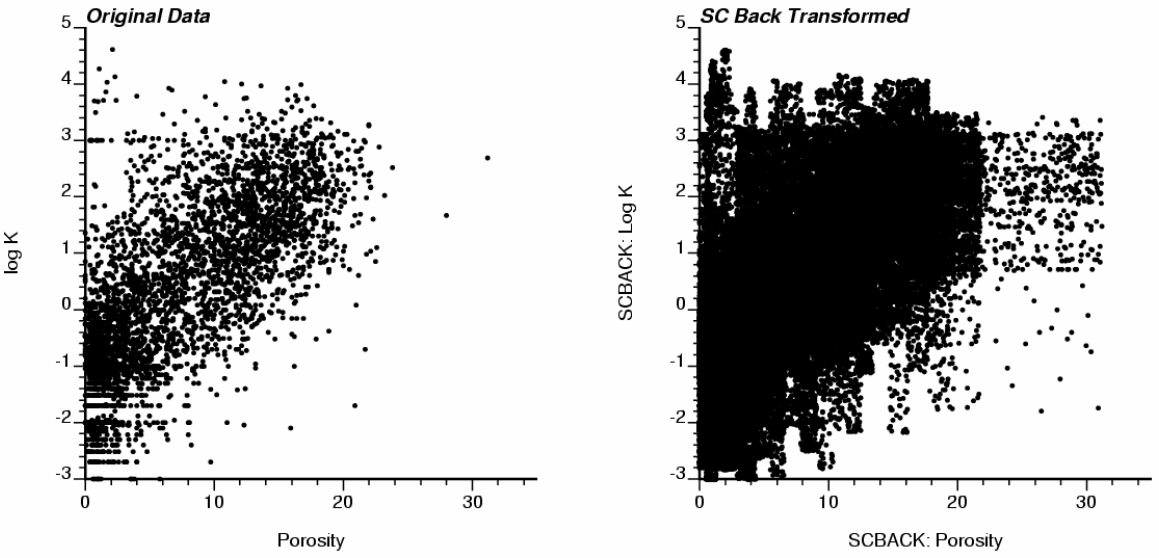


Figure 4: Cross plot of 3725 original log-permeability and porosity data (on the left) and the result of simulating 100,000 locations (on the right).

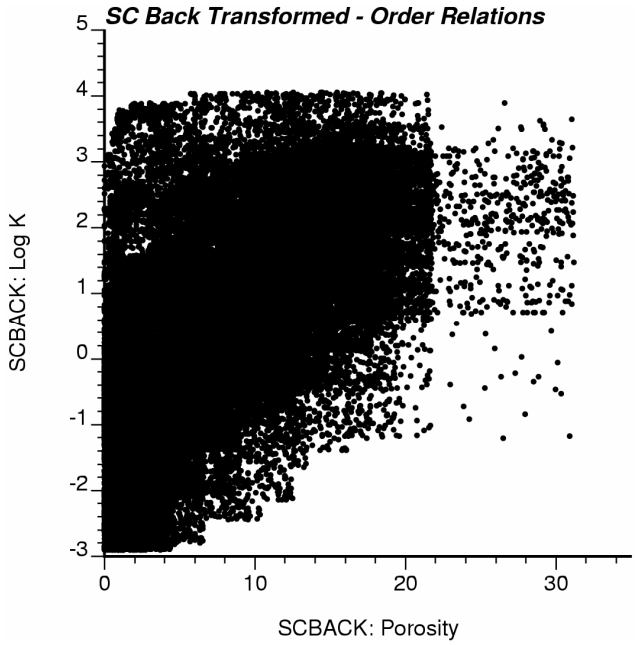


Figure 5: Cross plot of backtransformed 100,000 values when the original conditional distributions have order relations corrected.