A Gaussian Framework for Multivariate Multiscale Data Integration

Talal Alahaidib and Clayton V. Deutsch

Uncertainty quantification and data integration are two longstanding challenges addressed by geostatistics. The Gaussian framework is used extensively, but perhaps not to the fullest extent possible. This short note presents a comprehensive Gaussian framework for the problem of integrating multivariate multiscale data in the prediction at unsampled locations. Data are considered to be data events and the multivariate distribution of the Gaussian transform of the data events and the event being predicted is assumed to be multivariate Gaussian. The parameterization of the multivariate Gaussian distribution is done by means and covariances making the required assumptions to infer the scale-dependent covariance between different variables and locations.

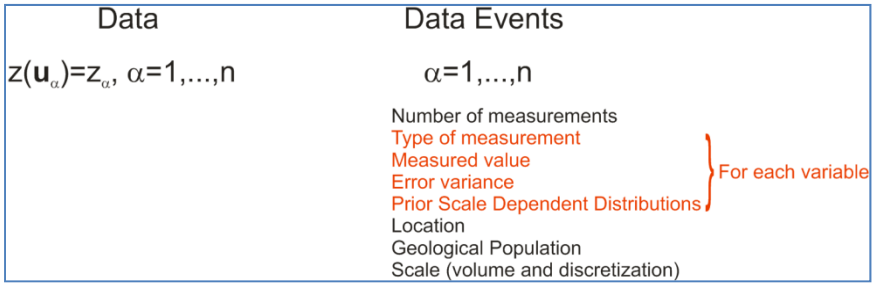
1. Introduction

Multivariate multiscale data are common in all fields of geostatistical modeling. It is increasingly common to deal with multiple variables. The context here is the common multivariate situation with three to ten variables. The massively multivariate setting with more than 25 variables would require separate processing and consideration of dimension reduction and other multivariate techniques. The input data are often at different scale and there are different scale grid blocks being assigned. At times it is reasonable to simulate points and scale up, but the direct consideration of scale is important when the data are at different scales.

Assuming that multiple variables follow a multivariate Gaussian distribution within reasonably defined geological subsets has proven useful. Once the point scale values are assumed Gaussian, then the larger scale values will also be Gaussian in most cases because most averaging is linear and this averaging will reinforce the multivariate Gaussian distribution through the central limit theorem. The limitation to take full advantage of the Gaussian distribution is primarily software and parameter inference. This research aims to fill those gaps and demonstrate the full potential of multivariate multiscale data integration.

2. Data Events

The notion of data will go beyond $z(\mathbf{u})$. Modern geostatistics is almost always related to multiple variables at different scale and with different error content. The following shows schematically the change in outlook: data values must be replaced by data events with the required number of attributes.



Each data is an event that has a number of attributes. The list of attributes in the graphic is not complete, but indicative of the proposed approach. In particular, it would be reasonable to have QA/QC information saved with the data (such as the processing date…). Some of the data attributes:

1. Measurements – there could be more than one. The measurements relate to the same location, scale and geological population (see below). For every measurement, there is a type (specified by a unique description), a measured value – the $z_k(\mathbf{u}_\alpha)$ values we are familiar with, an anticipated error variance associated to the measurement technique, and prior scale dependent distributions. These distributions are expanded on below. Some of the attributes such as data error may be unknown, but an approximate estimate is better than assuming the data are without error.
2. Centroid or anchor location for the data event. The location may be in multiple coordinate systems such as original and stratigraphic coordinates.

3. The scale of the data will be identified by an absolute volume, a shape and a discretization template that would be specified outside of the actual data event.
4. The geological population that the measured location corresponds to such as a zone, facies, geological unit or estimation domain.

The scale dependent prior distribution is quite important as it captures declustering/debiasing and non-stationarity. The distribution may be global or local (using weights along the lines of work by McLennan and Machuca-Mory). The distribution would be different at different scales.

The raw data will start with only some of the information available. Then, some data processing programs will fill in attributes in an iterative fashion. One processing program will assign the volume scale of the data and a discretization template. The absolute volume of the data measurement is particularly important for the nugget effect since that aspect of the variance changes inversely proportional to the volume. The discretization is required for numerical evaluation of average covariance values. The discretization should be sufficient (more than 10 points), but not too excessive because the computer effort for the numerical calculation of average covariance values will be a square of these values.

The scale dependent prior distribution is important because this will be used for transformation and back transformation. Another processing program will assign these distributions depending on the decision of stationarity. If enough data are available at the same scale, the distribution could easily be assembled by proportions perhaps with declustering weights. In general, it will not be required to downscale distributions. The data-scale distributions are primarily dependent on the choice of geological populations and trend model. The scale dependence will be influenced by the pattern of variability (fitted point scale model of coregionalization). The scale dependent distributions may be filled in after the model of coregionalization is established, see below.

3. Inference

The statistics needed to parameterize the *n* multivariate Gaussian distribution is a 1x*n* vector of mean values and an *n*x*n* variance-covariance matrix. A prior distribution is required for every data event to permit transformation to a Gaussian distribution.

Some attributes of the data events are not completely filled in at the start. The prior distributions at the data scale may be known from the previous step, but the scale dependence aspect of the distributions depends on the model of coregionalization fitted to the data. A point scale model of coregionalization will be required between the variables. The cross variogram values cannot be calculated for non collocated data, but cross covariances can be calculated (see Paper 412 in this report) and scaled to appear like cross variograms for fitting. Large scale variograms/covariances would have to be down scaled with established techniques (see Kupfersberger et al, 1998). Fitting a linear model of coregionalization is not possible manually, but there are automatic techniques such as the `varfit_lmc` program. The software tools and methodology to establish a point scale model of coregionalization given a set of multivariate multiscale data is non-trivial, but the basic approach is clear

The scaling laws for each variable to establish the scale dependent distributions will primarily be taken from the Gaussian approach used in the discrete Gaussian model (DGM). This model is straightforward to apply and different mineralization styles could be used if required to introduce flexibility in the change of shape.

4. Processing Data Events

The basic engine is simple cokriging. Non-stationarity is captured in the prior distributions. The data types must be considered in a model of coregionalization and the variances must be scaled.

Once all of the data are in Gaussian units, the calculation of the conditional mean and variance are with the well known normal equations.

$$\bar{y} = \sum_{\alpha=1}^{n} \lambda_\alpha y_\alpha \qquad \sigma^2 = C_{0,0} - \sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha,0} \qquad \sum_{\beta=1}^{n} \lambda_\beta C_{\alpha,\beta} = C_{\alpha,0}\,, \qquad \alpha = 1, \ldots, n$$

These may be applied for each variable being predicted at an unsampled location. The full kriging variance/covariance matrix will be retained for simulation. The covariances may be direct or cross covariances – depending on the data types involves. The covariances may also be averaged using the discretization template if the data are not at a point scale. The variances (the $C_{0,0}$ value in the equation above) would also depend on the scale and would have to be consistent with the Gaussian transformation.

The conditional distribution of the unsampled event is determined in Gaussian units as a mean and variance. This must be back transformed – either a single quantile (in simulation mode) or multiple quantiles (in estimation mode). In the case of estimation, the conditional mean and variance are sufficient to back transform multiple quantiles; however, multiple correlated values would have to be drawn from the variance/covariance matrix in simulation mode.

5. Software Considerations

There are a number of software considerations. The standard GSLIB data file format is somewhat limiting. There are many different attributes for each data. Some variables are simple, but others have a reasonably high dimension, for example, the scale dependent prior distributions. A database of distributions would be created and referred to by an index. Discretization templates would also be stored in a lookup table and referred to by index.

One program would be a data screening program to check the integrity of a database and to report on required steps before the data could be used in processing for unsampled locations. Text and visual reports on the data would be provided.

The declustering and debiasing programs would be adapted to work with multiple data events and determining stationary distributions. Trend modeling would come after the determination of stationary distributions. The local distributions would be modified to capture local changes in the mean, variance and shape of the distributions.

The fitting of the point scale model of coregionalizaiton would be achieved in a number of steps. Experimental direct and cross covariances at an arbitrary scale would be calculated in one step. Large scale covariances would be downscaled in another step. Finally, all of the direct and cross variograms would be assembled and fit in a final step. A consistent format would be used to save the fitted models (see the paper on the new Ultimate SGSIM in this volume).

The cokriging program for data integration is relatively straightforward. Care will be required to ensure the correct scale-dependent variance and average covariance values are used. Simulation will be performed in a sequential fashion.

The post processing program will be a much the same as the PostMG approach, but for multiple variables and non-stationary scale dependent transforms.

6. Conclusions

Traditional geostatistical programs mostly treat all of the data and locations being predicted as if they were the same scale. Some programs discretize the blocks being estimated and we often post process simulated point-scale realizations to scale them up to blocks. The proposal here is to explicitly treat the data scale of the data events and the unsampled locations being predicted. Large scale values are almost certainly Gaussian once we accept that the point scale values are multivariate Gaussian; averaging invokes the central limit theorem and reinforces the Gaussian distribution.

The software will be relatively slow because of the need to compute average (cross) covariances between multiple scale locations; nevertheless, the speed is not considered to be a major problem.

7. References

Chiles J.-P. and Delfiner P., 1999, Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, 641 pages

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.

Deutsch, C. V., 2002, *Geostatistical reservoir modeling*, Oxford University Press, New York.

Journel, A. G., 2002, Combining Knowledge from Diverse Sources: An Alternative to Traditional Data Independence Hypotheses, *Mathematical Geology*, Vol. 34, No. 5.

Kupfersberger, H., C. V. Deutsch, A. G. Journel, 1998, Deriving constraints on small-scale variograms due to variograms of large-scale data, *Mathematical Geology*, Vol. 30, No. 7.

Oz, B. and Deutsch C.V., 2000, Size scaling of cross-correlation between multiple variables. Centre for Computational Geostatistics, Annual Report 3, 1-25

Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York

Scott, D. W., 1992, *Multivariate Density Estimation*, John Wiley and Sons, Inc., New York.

Verly, G. W., 1984, Estimation of Spatial Point and Block Distributions:The Multi{G}aussian Model, Ph.D. Thesis, Stanford University