Joint Uncertainty Assessment with a Combined Bayesian Updating/LU/P-Field Approach

Clayton V. Deutsch, Weishan Ren and Oy Leuangthong

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

Reservoir characterization and, in general, site characterization in any spatial setting requires consideration of multiple correlated variables. There are often 10 or more correlated variables to be mapped. Some variables are geological trends or remotely sensed variables that are used to constrain the primary variables of direct interest. Trend maps and prior maps are used to understand each variable independently. Correlation matrices and likelihood maps are used to understand the correlation between variables and to show the predictive information contained in secondary structural data. The Bayesian updating approach is used to integrate the information from prior and secondary maps; the result is a local model of uncertainty for each variable.

Joint uncertainty between multiple variables and multiple locations is not directly calculated from such local uncertainty. For example, the calculation of recoverable reserves is a function of net pay thickness, porosity and oil saturation. Simulation methods are necessary to add the multivariate and spatial correlations into joint uncertainty. The LU method is ideally suited to multivariate uncertainty characterization between multiple variables at the same location. The P-field method is ideally suited for sampling joint uncertainty between multiple locations. These techniques are combined for assessing joint uncertainty in a practical setting. The theoretical framework will be developed and practical examples shown in the presentation. The combined Bayesian Updating/LU/P-field approach is remarkably simple to implement because of assumptions that permit the decomposition of different data sources and the decomposition of multivariate and spatial correlation.

Introduction

Conventional geostatistical techniques have been designed to create models of heterogeneity and uncertainty in static rock properties. This is appropriate for input to process evaluation. There are times, however, when the goal is uncertainty assessment and detailed realizations are not necessarily required. Moreover, there are times when we have many different variables: measured variables, large-scale remotely sensed variables, interpreted trend-like variables, and other response variables. These data often cover different areas, provide data at different scales, and are variably correlated together. Statistical techniques like principal components, factor analysis, ACE, and cluster analysis could be used to summarize the relationships between the variables, but they do not account for spatial correlation. Conventional geostatistical techniques incorporate the spatial structure but these techniques are cumbersome in the presence of many secondary variables. We propose that all secondary data be merged statistically by a multivariate Gaussian approach into a single variable that contains all of the secondary variable information; this provides a likelihood distribution. The spatial distribution of each variable by itself is mapped independently of the secondary variable information; this provides a prior distribution.

The likelihoods and priors are merged to provide updated posterior distributions. Further processing is needed to calculate joint uncertainty from the local uncertainty in each variable

The LU and p-field simulation methods are combined to assess joint uncertainty between multiple variables and multiple locations. The former was used primarily for simulation of multiple variables, while the latter was used for the spatial component of simulation.

Context for Methodology

There are a number of geostatistical techniques designed to work with multiple variables. These techniques account for the spatial relationships between the variables and provide a measure of uncertainty at every estimated location. The main technique is cokriging that can be applied in a multivariate Gaussian or an indicator framework. There are simplifying assumptions such as collocated cokriging and the Markov-Bayes approach. A concern with all these techniques is the inference of the direct and cross variogram measures of correlation, which requires a large number of data. For K variables, they require a total of $(K+1)K/2$ variogram models, which is difficult in practice. Automatic fitting algorithms have helped; however, the problem of inference remains when we have K=10 or more variables with relatively few data.

Collocated cokriging, in the Gaussian or Bayesian form, simplifies the process to consider only the collocated secondary variables. This also removes the need to model the large number of variograms mentioned above. There are implementation problems associated with this simplification such as variance inflation, but the method has proved very practical. These geostatistical methods for considering multiple variables really only consider 1 to 3 secondary variables; there is no simple way to consider 10 to 30 secondary variables simultaneously. We must tailor the multivariate statistical and geostatistical tools to the problem of a large number of variables and relatively few data.

The proposed methodology is Gaussian, that is, all data variables must be transformed to univariate Gaussian distributions prior to analysis and results must be back transformed. A parametric distribution model for each variable or a non-parametric normal-scores transformation could be used. Care should be taken to decluster/debias the original data histograms. The variables are assumed to be multivariate Gaussian after univariate transformation of each variable. There are some clear indications of non-Gaussian behavior: non-linear relationships, proportional effect (dependency of the variance on the mean), constraints due to mineralogical constraints, or a constant sum constraint. A special transformation may need to be considered for these non-Gaussian cases. Log-ratios and the stepwise conditional transformation are two alternatives.

There are two key ideas of the proposed methodology (1) prediction of the conditional distribution of uncertainty in all variables at all locations using a Bayesian updating formalism, and (2) assessment of joint uncertainty with a combined LU/p-field approach.

Bayesian Updating

It is common to have two types of data variables. *Primary* data variables are to be predicted with uncertainty and are available at relatively few locations. Consider N_P ($y_{P,i}$, $i=1,\ldots,N_P$) primary variables. *Secondary* data variables are not to be predicted and are often available more extensively. Examples of secondary variables include geophysical measurements and geologic

trend maps. Consider N_S secondary data variables ($y_{S,i}$, $i=1,\ldots,N_S$). The location is often denoted as $\mathbf{u}_\alpha$ where α is the data index.

All secondary variables are merged into a single *likelihood* distribution for each primary variable at each location. Of course, the number of secondary variables available at each location could vary; the notation $N_s(\mathbf{u})$ denotes the number of secondary data available at location $\mathbf{u}$. The mean and variance of the likelihood distribution are calculated as:

$$\left.\begin{aligned} \overline{y}_{L,p}(\mathbf{u}) &= \sum_{i=1}^{N_S(\mathbf{u})} \nu_{i,p}(\mathbf{u})\, y_{S,i}(\mathbf{u}) \\ \sigma^2_{L,p}(\mathbf{u}) &= 1 - \sum_{i=1}^{N_S(\mathbf{u})} \nu_{i,p}(\mathbf{u})\, \rho_{S,i,p} \\ \sum_{j=1}^{N_S(\mathbf{u})} \nu_{i,p}(\mathbf{u})\, \rho_{S,i,S,j} &= \rho_{S,i,p},\ i = 1,\ldots,N_S(\mathbf{u}) \end{aligned}\right\} p = 1,\ldots,N_P, \forall\, \mathbf{u} \in A \tag{1}$$

These equations are the well understood normal equations or simple kriging equations. The last row shows how to calculate the weights $\nu_{i,p}(\mathbf{u})$. Correlation coefficients between all pairs of secondary data and all secondary and primary data are required. The likelihood distributions are summarized by a set of mean and variance values for all locations and primary variables:

$$\overline{y}_{L,p}(\mathbf{u}), \sigma^2_{L,p}(\mathbf{u});\ p = 1,\ldots,N_P, \forall\, \mathbf{u} \in A \tag{2}$$

These distributions are a "collapsed" version of all available secondary variables at location **u**. The final likelihood distributions account for the relationships between the secondary variables and will be used to help inform the primary estimate. Spatial information is not accounted for; these values summarize all of the information available in the secondary data related to the primary variables of interest. The spatial information comes in through the prior distributions.

The distribution of uncertainty in each primary variable is predicted from surrounding data (in a spatial sense) using simple kriging. These estimates are called prior distributions and are denoted with a *P*; the context will clarify the distinction between prior and primary variable. Similar to the likelihood distribution, the parameters of the prior distribution are obtained as:

$$\left.\begin{aligned} \overline{y}_{P,p}(\mathbf{u}) &= \sum_{i=1}^{N_p(\mathbf{u})} \lambda_{i,p}(\mathbf{u})\, y_P(\mathbf{u}_i) \\ \sigma^2_{P,p}(\mathbf{u}) &= 1 - \sum_{i=1}^{N_p(\mathbf{u})} \lambda_{i,p}(\mathbf{u})\, \rho_{p,\mathbf{u}_i,\mathbf{u}} \\ \sum_{j=1}^{N_p(\mathbf{u})} \lambda_{i,p}(\mathbf{u})\, \rho_{p,\mathbf{u}_i,\mathbf{u}_j} &= \rho_{p,\mathbf{u}_i,\mathbf{u}},\ i = 1,\ldots,N_p(\mathbf{u}) \end{aligned}\right\} p = 1,\ldots,N_P, \forall\, \mathbf{u} \in A \tag{3}$$

The correlation coefficients between all primary data at different locations come directly from a semivariogram or correlogram model. The prior distributions are summarized by a set of mean and variance values for all locations and primary variables:

$$\overline{y}_{P,p}(\mathbf{u}), \sigma^2_{P,p}(\mathbf{u});\ p = 1,\ldots,N_P, \forall\, \mathbf{u} \in A \tag{4}$$

These distributions summarize the spatial information of surrounding data of the same variable type. The likelihood and prior distributions are then combined to get the final updated distribution. Since the two input distributions are Gaussian in shape, the resulting updated distribution will also be Gaussian. The updated distribution is defined by the updated mean and variance:

$$\left.\begin{aligned} \overline{y}_{U,p}(\mathbf{u}) &= \frac{\overline{y}_{L,p}(\mathbf{u})\cdot\sigma^2_{P,p}(\mathbf{u}) + \overline{y}_{P,p}(\mathbf{u})\cdot\sigma^2_{L,p}(\mathbf{u})}{\left(1-\sigma^2_{L,p}(\mathbf{u})\right)\left(\sigma^2_{P,p}(\mathbf{u})-1\right)+1} \\ \sigma^2_{U,p}(\mathbf{u}) &= \frac{\sigma^2_{L,p}(\mathbf{u})\cdot\sigma^2_{P,p}(\mathbf{u})}{\left(1-\sigma^2_{L,p}(\mathbf{u})\right)\left(\sigma^2_{P,p}(\mathbf{u})-1\right)+1} \end{aligned}\right\} p=1,....,N_P, \forall \mathbf{u}\in A \qquad (5)$$

The updated distributions defined above must be back-transformed to return the primary variables to their original distributions. The proposed technique is summarized as **B**ayesian Updating under a **M**ultivariate **G**aussian model - or a **BMG** model for lack of a better acronym. The elements of this technique are not new; however, this is a novel way of putting everything together for reliable and simple estimation. A Markov screening assumption is made whereby collocated secondary data screen the influence of nearby secondary data. There is a further assumption that primary data of different types at different locations are also screened. The consequences of these assumptions are not considered severe in most cases. Full cokriging could be implemented to judge their importance.

The percentiles, or arbitrary number of quantiles, could be back transformed from the local distributions (Eq. 5). Any summary statistics of the local distributions could then be calculated including the expected value, the local variance, P_{10}, P_{50} and P_{90} values and so on. These summaries could be used to appreciate local uncertainty and to assist with well placement and data collection decisions. Local uncertainty in each of the N_p variables at each location $\forall \mathbf{u}\in A$ does not permit multivariate calculations or uncertainty over larger volumes. A simulation approach is required for those calculations.

Joint Uncertainty with LU/P-Field Simulation

We are often interested in derived variables such as economic value or net calculations. Multiple variables must be combined together. The distributions of uncertainty in the input variables can sometimes be combined analytically, but only when the calculations are simple. In general, a simulation approach is required. Multiple realizations are drawn, each realization is processed to establish the derived variables and distributions of uncertainty in the derived variables are assembled.

The N_P variables are correlated with a known structure; the N_P by N_P covariance or correlation matrix **C** is known from the data. This matrix of covariance values is not used in the Bayesian Updating since we only need the correlation between the secondary and the primary; however, the goal at this time is to draw primary values with the correct correlation structure. N_P Gaussian values can be drawn from the covariance matrix very simply with the LU simulation approach: (1) the **C** matrix is decomposed into lower and upper matrices **L** and **U** using Cholesky decomposition, (2) vectors of N_P random Gaussian values are created by a random number generator (the random vector is commonly denoted **w**), and (3) correlated Gaussian values are calculated as **y**=**Lw**. The **y** vectors of correlated values have the correct covariance structure, but

they do not respect the updated local distributions of uncertainty, that is, they do not follow $\overline{y}_{U,p}(\mathbf{u}), \sigma^2_{U,p}(\mathbf{u});\ p=1,....,N_P, \forall \mathbf{u} \in A$. A post processing correction is applied:

$$y_{c,p}(\mathbf{u}) = y_p(\mathbf{u}) \square \sigma_{U,p}(\mathbf{u}) + \overline{y}_{U,p}(\mathbf{u});\ p=1,....,N_P, \forall \mathbf{u} \in A \tag{6}$$

where the $y_p(\mathbf{u})$ values are the result of LU simulation. The $y_{cp}(\mathbf{u})$ values are conditional to the updated distributions of uncertainty. This procedure can be seen as a Gaussian based P-field approach where LU simulation is used for the probabilities. There are minor concerns related to the non-stationarity of the resulting conditional covariance structure and hard data appearing as local minima and maxima.

The LU simulation approach works very well for multiple correlated variables – the 10 to 30 primary data variables we encounter in practice; however, we are often interested in joint spatial uncertainty over a large area or volume. The covariance matrix size becomes intractably large when the number of variables is greater than, say, 5000. Three techniques have evolved for large problems: (1) turning bands, (2) sequential simulation, and (3) P-field simulation. There is extensive literature available on each technique. We adopt P-field simulation for sampling large scale uncertainty. There are some minor concerns, as previously mentioned, but there is one very significant advantage. The global uncertainty calculated from P-field simulation is perfectly consistent with the local uncertainty – all data sources and spatial features contained in the local uncertainty predictions are reproduced exactly.

The P-field procedure amounts to simultaneously sampling many local distributions of uncertainty with correlated probabilities. In a Gaussian context a standard Gaussian value takes the place of a probability value since they are related through $p=G(y)$ and $y=G^{-1}(p)$. Thus, application to a large multivariate problem proceeds as follows:

1. Generate a probability field for each variable ($y_p^{p-field}(\mathbf{u});\ p=1,....,N_P, \forall \mathbf{u} \in A$) using a Gaussian simulation technique (we used sequential Gaussian simulation). The variogram for each probability field is the normal scores variogram. There is some discussion on using the rank-order variogram; however, the normal scores of the rank transform is equivalent to the normal scores of the original variable.

2. Apply the LU algorithm to $y_p^{p-field}(\mathbf{u})$ at each location to ensure that the multivariate structure is reasonable, that is, multiply $\mathbf{y}_p^{LU}(\mathbf{u}) = \mathbf{L}\mathbf{y}_p^{p-field}(\mathbf{u})$. The "LU" values at each location have the right spatial correlation structure (from step 1) and the right multivariate structure (from the product with **L**).

3. Condition the P-field/LU results to the local data by the procedure given in Eq. 6: $y_{c,p}(\mathbf{u}) = y_p^{LU}(\mathbf{u}) \square \sigma_{U,p}(\mathbf{u}) + \overline{y}_{U,p}(\mathbf{u});\ p=1,....,N_P, \forall \mathbf{u} \in A$. The final result is exactly what we need, that is, a realization of all variables with the right multivariate structure and the right spatial structure.

Multiple realizations are generated by repeating steps 1-3 with a different random number seed. The results of Eq. 6 are not needed if we proceed with the full procedure described in steps 1-3; however, we often need local uncertainty in derived variables without recourse to global uncertainty.

Discussions and Conclusions

Prediction of uncertainty with multiple primary and secondary variables is an important new area of geostatistics. Bayesian updating under a multivariate Gaussian model provides a simple and robust solution to this inference problem. LU and P-field simulation permit calculation of complex derived variables and uncertainty over large areas. There are, of course, limitations and assumptions such as representative data, statistical homogeneity and multivariate Gaussianity. The procedure advocated in this paper may appear like a hodge-podge of techniques. Each constituent technique is required for a specific purpose of data integration or accounting for multivariate or spatial structure. Simpler techniques would necessarily leave out some aspect of data structure.

References

Deutsch, C.V. and Journel, A.G., 1998: GSLIB - Geostatistical software library and users guide. Oxford University press, 2nd Edition.

Doyen, P. M., L. D. den Boer and W. R. Pillet, 1996, Seismic Porosity Mapping in the Ekofisk Field Using a New Form of Collocated Cokriging, Society of Petroleum Engineers Paper Number 36498.

Xu, W., T. T. Tran, R. M. Srivastava and A. G. Journel, 1992, Integrating Seismic Data in Reservoir Modeling: The Collocated Cokriging Alternative, Society of Petroleum Engineers Paper Number 24742.