Modeling Multivariate Multiscale Data

Oy Leuangthong (oy@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Clayton V. Deutsch (cdeutsch@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Abstract

Direct sequential simulation is an attractive alternative to conventional Gaussian simulation. It permits the consideration of multiscale data, and has the advantage of linear averaging for block simulation. Extending direct simulation to multiscale, multivariate data is the focus of this research. Applying the generalized co-kriging equations simultaneously for multiple variables is an efficient approach to account for multiple data types and different volume supports. Although this gives the distributional parameters of the conditional univariate distributions, inference of the conditional multivariate distribution is a significant challenge for simulation. An iterative updating approach is proposed, whereby the global multivariate distribution is scaled by ratios of the desired conditional univariate distribution to the marginal univariate distribution. Exercises in the validation of this approach are presented along with the proposed methodology for direct sequential cosimulation (DSCosim).

Introduction

The idea of direct sequential simulation (DSS) is to simulate in original data units, without assumptions or transformations about the data distribution. This allows the use of multiscale data in DSS. Recent advances in the inference of univariate conditional distributions [3, 8] provide a key link to the extension of DSS to multivariate multiscale data. Conditional distributions in original data units are established without the requirement for data transformation.

Thus far, development of the DSS algorithm has been limited to simulating one variable at a time. Kriging provides the mean and variance parameters of conditional univariate distributions. The focus of this research is to simultaneously simulate multiple variables by inferring a conditional *multivariate* distribution. The framework to accomplish this objective is presented, including a review of the co-kriging formalism and a discussion of anticipated theoretical and practical challenges. The proposed methodology for direct sequential cosimulation of multiple multiscale data is presented.

Generalized Co-kriging

Suppose there are P variables, $Z_p, p = 1, \ldots, P$ with mean μ_p defined on support V_p centered at location $\mathbf{u}_{\alpha p}$, where $\alpha = 1, \ldots, n_p$ and n_p is the number of available data of type p. It is not necessary that the volume supports $V_p, p = 1, \ldots, P$ be constant.

$$Z_p(\mathbf{u}_{\alpha p}) = \frac{1}{V_p} \int_{V_p} Z_p(\mathbf{u}_{\alpha p}) du$$

The residual of the original Z_p variable about its mean, $Y_p(\mathbf{u}_{\alpha p})$

$$Y_p(\mathbf{u}_{\alpha p}) = Z_p(\mathbf{u}_{\alpha p}) - \mu_p(\mathbf{u}_{\alpha p}), \forall p, \mathbf{u}_{\alpha p}$$

is also defined on support V_p. Consider estimating $Y_i^*(\mathbf{u})$ as a linear combination of the P data types (where i can be any one of the P variables):

$$Y_i^*(\mathbf{u}) = \sum_{p=1}^{P} \sum_{\alpha}^{n_p} \lambda_{\alpha p} Y_p(\mathbf{u}_{\alpha p})$$

The corresponding estimation variance is

$$\begin{aligned}
\sigma_E^2 &= E\{(Y_i(\mathbf{u}) - Y_i^*(\mathbf{u}))^2\} \\
&= E\{[Y_i(\mathbf{u})]^2 + [Y_i^*(\mathbf{u})]^2 - 2Y_i(\mathbf{u}) \cdot Y_i^*(\mathbf{u})\} \\
&= E\{[Y_i(\mathbf{u})]^2\} + E\{[Y_i^*(\mathbf{u})]^2\} - 2E\{Y_i(\mathbf{u}) \cdot Y_i^*(\mathbf{u})\} \\
&= E\{[Y_i(\mathbf{u})]^2\} + E\left\{\sum_{p=1}^{P}\sum_{p'=1}^{P}\sum_{\alpha=1}^{n_p}\sum_{\beta=1}^{n_{p'}} \lambda_{\alpha p}\lambda_{\beta p'} Y_p(\mathbf{u}_{\alpha p}) Y_{p'}(\mathbf{u}_{\beta p'})\right\} - 2E\left\{\sum_{p=1}^{P}\sum_{\alpha}^{n_p} \lambda_p(\mathbf{u}_{\alpha p}) Y_i(\mathbf{u}) Y_p(\mathbf{u}_{\alpha p})\right\} \\
&= E\{[Y_i(\mathbf{u})]^2\} + \sum_{p=1}^{P}\sum_{p'=1}^{P}\sum_{\alpha=1}^{n_p}\sum_{\beta=1}^{n_{p'}} \lambda_{\alpha p}\lambda_{\beta p'} E\{Y_p(\mathbf{u}_{\alpha p}) Y_{p'}(\mathbf{u}_{\beta p'})\} - 2\sum_{p=1}^{P}\sum_{\alpha}^{n_p} \lambda_p(\mathbf{u}_{\alpha p}) E\{Y_i(\mathbf{u}) Y_p(\mathbf{u}_{\alpha p})\} \\
&= \bar{C}(V_i(\mathbf{u}), V_i(\mathbf{u})) + \sum_{p=1}^{P}\sum_{p'=1}^{P}\sum_{\alpha=1}^{n_p}\sum_{\beta=1}^{n_{p'}} \lambda_{\alpha p}\lambda_{\beta p'} \bar{C}(V_p(\mathbf{u}_{\alpha p}), V_{p'}(\mathbf{u}_{\beta p'})) - 2\sum_{p=1}^{P}\sum_{\alpha}^{n_p} \lambda_p(\mathbf{u}_{\alpha p}) \bar{C}(V_i(\mathbf{u}), V_p(\mathbf{u}_{\alpha p}))
\end{aligned}$$

with

$$\bar{C}(V_p(\mathbf{u}_{\alpha p}), V_{p'}(\mathbf{u}_{\beta p'})) = \frac{1}{V_p \cdot V_{p'}} \int_{V_p} du \int_{V_{p'}} C(V_p(\mathbf{u}_{\alpha p}), V_{p'}(\mathbf{u}_{\beta p'}))dv$$

Minimizing the error variance with respect to the weights gives the $\sum_{p=1}^{P} n_p$ equations that constitute the simple co-kriging system of equations:

$$\sum_{p'=1}^{P}\sum_{\beta=1}^{n_{p'}} \lambda_{\beta p'} \bar{C}(V_p(\mathbf{u}_{\alpha p}), V_{p'}(\mathbf{u}_{\beta p'})) \quad = \quad \bar{C}(V_i(\mathbf{u}), V_p(\mathbf{u}_{\alpha p})) \tag{1}$$

where $p = 1, \ldots, P$ and $\alpha = 1, \ldots, n_p$. In matrix notation, the left hand side covariance matrix consists of $P \times P$ submatrices of volume to volume covariances between different data types.

$$\begin{bmatrix}
\bar{C}(V_1(\mathbf{u}_{11}), V_1(\mathbf{u}_{11})) & \cdots & \bar{C}(V_1(\mathbf{u}_{11}), V_1(\mathbf{u}_{n_1 1})) & & \bar{C}(V_1(\mathbf{u}_{11}), V_P(\mathbf{u}_{1P})) & \cdots & \bar{C}(V_1(\mathbf{u}_{11}), V_P(\mathbf{u}_{n_P P})) \\
\vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\
\bar{C}(V_1(\mathbf{u}_{n_1 1}), V_1(\mathbf{u}_{11})) & \cdots & \bar{C}(V_1(\mathbf{u}_{n_1 1}), V_1(\mathbf{u}_{n_1 1})) & & \bar{C}(V_1(\mathbf{u}_{n_1 1}), V_P(\mathbf{u}_{1P})) & \cdots & \bar{C}(V_1(\mathbf{u}_{n_1 1}), V_P(\mathbf{u}_{n_P P})) \\
 & \vdots & & \ddots & & \vdots & \\
 & \vdots & & & & \vdots & \\
\bar{C}(V_P(\mathbf{u}_{1P}), V_1(\mathbf{u}_{11})) & \cdots & \bar{C}(V_P(\mathbf{u}_{1P}), V_1(\mathbf{u}_{n_1 1})) & & \bar{C}(V_P(\mathbf{u}_{1P}), V_P(\mathbf{u}_{1P})) & \cdots & \bar{C}(V_P(\mathbf{u}_{1P}), V_P(\mathbf{u}_{n_P P})) \\
\vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\
\bar{C}(V_P(\mathbf{u}_{n_P P}), V_1(\mathbf{u}_{11})) & \cdots & \bar{C}(V_P(\mathbf{u}_{n_P P}), V_1(\mathbf{u}_{n_1 1})) & & \bar{C}(V_P(\mathbf{u}_{n_P P}), V_P(\mathbf{u}_{1P})) & \cdots & \bar{C}(V_P(\mathbf{u}_{n_P P}), V_P(\mathbf{u}_{n_P P}))
\end{bmatrix}$$

or simply

$$[[\bar{\mathbf{C}}(\mathbf{V_p}, \mathbf{V_{p'}})], p, p' = 1, \ldots, P]$$

where each submatrix consists of $n_p \times n_{p'}$ covariances between the p and the p' data.

$$\bar{\mathbf{C}}(\mathbf{V_p}, \mathbf{V_{p'}}) = \begin{bmatrix}
\bar{C}(V_p(\mathbf{u}_{1p}), V_{p'}(\mathbf{u}_{1p'})) & \cdots & \bar{C}(V_p(\mathbf{u}_{1p}), V_{p'}(\mathbf{u}_{n_{p}p'})) \\
\vdots & \ddots & \vdots \\
\bar{C}(V_p(\mathbf{u}_{n_p p}), V_{p'}(\mathbf{u}_{1p'})) & \cdots & \bar{C}(V_p(\mathbf{u}_{n_p p}), V_{p'}(\mathbf{u}_{n_p p'}))
\end{bmatrix}$$

The large covariance matrix (containing all submatrices) is symmetric.

$$[\bar{\mathbf{C}}(\mathbf{V_p}, \mathbf{V_{p'}})] = [\bar{\mathbf{C}}(\mathbf{V_{p'}}, \mathbf{V_p})]^T$$

The column vector of weights and right hand side covariances then consists of $\sum_{p=1}^{P} n_p$ elements:

$$\lambda = \begin{bmatrix} \lambda_{11} \\ \vdots \\ \lambda_{n_1 1} \\ \\ \vdots \\ \vdots \\ \\ \lambda_{1P} \\ \vdots \\ \lambda_{n_P P} \end{bmatrix} \qquad \bar{\mathbf{C}}(\mathbf{V_i}(\mathbf{u}), \mathbf{V_p}(\mathbf{u_{\alpha p}})) = \begin{bmatrix} \bar{C}(V_i(\mathbf{u}), V_1(\mathbf{u}_{11})) \\ \vdots \\ \bar{C}(V_i(\mathbf{u}), V_1(\mathbf{u}_{n_1 1})) \\ \\ \vdots \\ \vdots \\ \\ \bar{C}(V_i(\mathbf{u}), V_P(\mathbf{u}_{1P})) \\ \vdots \\ \bar{C}(V_i(\mathbf{u}), V_P(\mathbf{u}_{n_P P})) \end{bmatrix}$$

Solving for the weights in the co-kriging system of equations (1) gives the minimized error variance known as the simple co-kriging (SCK) variance

$$\sigma^2_{SCK} = \bar{C}(V_i(\mathbf{u}), V_i(\mathbf{u})) - \sum_{p=1}^{P} \sum_{\alpha=1}^{n_p} \lambda_{\alpha p} \bar{C}(V_i(\mathbf{u}), V_p(\mathbf{u}_{\alpha p}))$$

The resulting co-kriging estimate and estimation variance correspond to the conditional expectation and variance of the RV $Y_i(\mathbf{u})$. The above co-kriging system corresponds to simple kriging; however, it is straightforward to modify the above formalism to reflect the unit sum constraint of the weights for ordinary kriging.

Simultaneous Co-kriging of Multiscale Data

We can consider the simultaneous co-kriging of M multiple data types simply by changing the column vector of weights and right hand side covariance into an $M \times P$ matrix, where $M \leq P$.

$$\begin{aligned} Y_1^*(\mathbf{u}) &= \sum_{p=1}^{P} \sum_{\alpha}^{n_p} \lambda^1_{\alpha p} Y_p(\mathbf{u}_{\alpha p}) \\ &\vdots \\ Y_M^*(\mathbf{u}) &= \sum_{p=1}^{P} \sum_{\alpha}^{n_p} \lambda^M_{\alpha p} Y_p(\mathbf{u}_{\alpha p}) \end{aligned}$$

Solving for the weights of the resulting co-kriging system requires little additional effort since the large left hand side covariance only has to be inverted once. Matrix multiplication of the inverted covariance matrix with the additional $M - 1$ columns of the right hand side covariance will give the weights to estimate the other $M - 1$ additional variables. In fact, most solvers can be modified to solve systems of simultaneous equations with multiple right hand sides without explicitly solving for an inverse.

$$\lambda = \begin{bmatrix} \lambda_{11}^{1} & \cdots & \lambda_{11}^{M} \\ \vdots & & \vdots \\ \lambda_{n_1 1}^{1} & \cdots & \lambda_{n_1 1}^{M} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \lambda_{1P}^{1} & \cdots & \lambda_{1P}^{M} \\ \vdots & & \vdots \\ \lambda_{n_P P}^{1} & \cdots & \lambda_{n_P P}^{M} \end{bmatrix} \qquad \bar{\mathbf{C}}(\mathbf{V_i}(\mathbf{u}), \mathbf{V_p}(\mathbf{u_{\alpha p}})) = \begin{bmatrix} \bar{C}(V_1(\mathbf{u}), V_1(\mathbf{u}_{11})) & \cdots & \bar{C}(V_M(\mathbf{u}), V_1(\mathbf{u}_{11})) \\ \vdots & & \vdots \\ \bar{C}(V_1(\mathbf{u}), V_1(\mathbf{u}_{n_1 1})) & \cdots & \bar{C}(V_M(\mathbf{u}), V_1(\mathbf{u}_{n_1 1})) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \bar{C}(V_1(\mathbf{u}), V_P(\mathbf{u}_{1P})) & \cdots & \bar{C}(V_M(\mathbf{u}), V_P(\mathbf{u}_{1P})) \\ \vdots & & \vdots \\ \bar{C}(V_1(\mathbf{u}), V_P(\mathbf{u}_{n_P P})) & \cdots & \bar{C}(V_M(\mathbf{u}), V_P(\mathbf{u}_{n_P P})) \end{bmatrix}$$

The only additional computations required in order to simultaneously estimate the multiple data types is the determination of the right hand side volume to volume covariance between the location to be estimated and the nearby data of P types. While co-kriging of one variable gives the conditional expectation and variance of the RV, simultaneous co-kriging of multiple RVs gives the conditional mean vector and covariance matrix of the M RVs. Simulation using these distributional parameters must still be performed.

Example of Co-kriging of Multiscale Data

Consider two types of data - 7 core samples of variable Y_1 and 3 seismic data of variable Y_2. The core data are considered to be point-scale data, and the seismic data are block scale data informing a 50×50 volume. Suppose we are interested in cokriging an intermediate 10×10 volume. Without loss of generality, suppose both variables, Y_1 and Y_2, have the same direct isotropic variogram:

$$\gamma(\mathbf{h}) = 0.5Sph_{a=3}(\mathbf{h}) + 0.5Sph_{a=15}(\mathbf{h})$$

The correlation between Y_1 and Y_2 was chosen to be 0.70, with an intrinsic cross variogram:

$$\gamma(\mathbf{h}) \quad = \quad 0.35Sph_{a=3}(\mathbf{h}) + 0.35Sph_{a=15}(\mathbf{h})$$

In practice, we calculate these point-scale statistics by downscaling the seismic statistics (see Oz et. al. [7, 6]). We create a consistent data set by simulation.

Two unconditional Gaussian simulations were used to generate two reference 2-D maps at the point scale on a 150×50 grid. The first map is considered the reference map of the core data. Seven samples were drawn from this map to give the 7 core samples that will be used for cokriging. The second map was generated by cokriging with a correlation of 0.70. This "reference" map was then upscaled to the 50×50 volume to provide the 3 seismic data. Figure 1 shows both the core and seismic data on the same map.

Cokriging is performed by setting up the simple co-kriging equations (see Equation 1). Average volume to volume covariances are numerically calculated by discretizing each volume into 25 points (i.e. 5×5 discretizations), calculating the covariance values between the points in one volume and the points in the second volume, and then averaging these covariance values. To set up the left hand side covariance matrix (i.e. covariance between the data and themselves), the direct variograms were used to get the average covariance between two data of the same type, while the cross variogram is used to determine the average covariance between two different types of data. Similarly, the right hand covariance vector (i.e. covariance between the data and the volume to be estimated) relies on the direct variogram for data of the same type and the cross variogram for data of different types. The key in this latter calculation is that the intermediate scale (which is the volume that we are interested in) is the same type of "data" as the core data and not the seismic data. This means that the average covariance between the core data and the volume to be estimated uses the direct

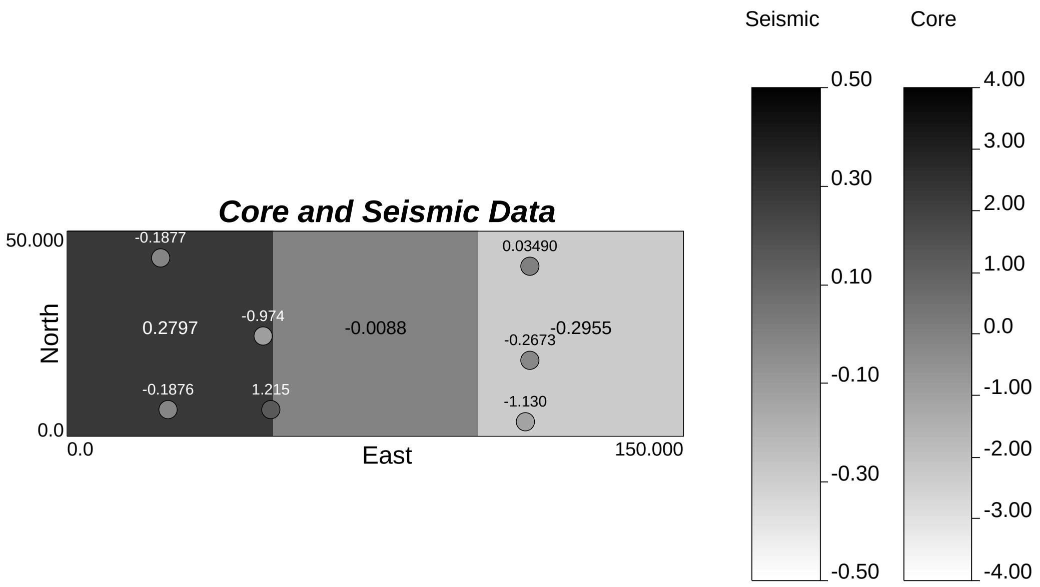

Figure 1: 2-D map of core and seismic data. Seismic data values are shown in larger font centered on the 50 × 50 volume which it informs.

variogram, while the average covariance between the seismic data and this same volume uses the cross variogram. Note that in all cases, it is the point scale variograms that are used in the numerical calculation.

For the 10×10 volume centered at coordinates (35,35), the resulting cokriged estimate is 0.15882 with an estimation variance of 0.41576. This simple exercise for one variable can easily be extended to the simultaneous cokriging of multiple variables (as discussed above) to obtain the conditional mean vector and covariance matrix. This multiscale cokriging is key to DSS for multiple variables; however, some theoretical and practical challenges still exist in the implementation of a direct sequential cosimulation (DSCosim) algorithm.

Outstanding Theoretical and Practical Challenges

Classical geostatistical simulation relies on kriging and Monte Carlo simulation to obtain simulated values. DSS is no different. Kriging (or cokriging in the case of multiple variables) is used to determine the mean and variance of the conditional univariate distribution at the location to be simulated. Assuming that the two distributional parameters (mean and variance) fully define the conditional distribution, Monte Carlo simulation is performed to obtain a simulated value.

For multivariate geostatistics using multiscale data, inference of the coregionalization model is still a challenge. In the case of multiscale data, downscaling the spatial measures of the larger scale data to the smaller scale data is especially challenging. It requires the inference of a short scale structure for a volume that is smaller than that informed by the data. In the previous cokriging example, the average covariance values were numerically determined using the point scale variograms. In the case of real data where the point scale variogram of the large scale data (such as seismic) is unknown, programs exist to downscale the block scale variogram to a point scale (or some small finite scale) variogram; however, it is the inference of the coregionalization model at the point scale that is most challenging. This requires:

1. Inference of a same-scaled cross-variogram based on different scaled data;

2. Downscaling of the direct block-scale variogram to a direct point-scale variogram; and
3. Iterating between Steps 1 and 2 to ensure (i) a legitimate model of coregionalization, and (ii) consistency between the model variograms and the variability of the natural phenomena.

The inference of the coregionalization model is a difficult issue in any conventional multivariate geostatistics that accounts for multiscale data.

Two other challenges that are a result of simulating in direct space are (1) the limitations of using kriging in simulation, and (2) inference of the multivariate distribution.

Implications due to Kriging

Kriging is a linear estimator. The kriging estimate is also the conditional expectation of the RV given the conditioning data. A consequence of linearity in the conditional expectation is the inability to reproduce complex non-linear features. Unfortunately, real data exhibit complex relations.

Kriging also provides information about the uncertainty in its estimate. This is the kriging variance. The variance is independent of the data values and the estimate, hence it is homoscedastic. In contrast, the variance of mineral grades or petrophysical properties found in a real deposit or reservoir is heteroscedastic. For example, it is common to find a low variance in a low valued area, and a correspondingly high variance in a high valued area. The use of the kriging variance does not account for heteroscedastic behaviour of the conditional distribution.

Inference of Multivariate Distribution

Simultaneous kriging of all variables yields the conditional mean vector and the conditional variance corresponding to each variable. All that is required are the correlation coefficients at $\mathbf{h} = 0$ in order to fully define the covariance matrix of the multivariate distribution at $\mathbf{h} = 0$. Given that these correlations are known or can be estimated (as in the case of non-isotopic sampling [9]), and that the multivariate distribution is fully defined by its mean vector and covariance matrix, simulation can proceed by recursive application of Bayes relation.

In the conventional Gaussian framework, the mean vector and covariance matrix provides all the information required to define the multivariate distribution. However, in direct space this information is insufficient. In fact, multivariate distributions of real data generally do not follow nice parametric forms. In these instances, knowing only the mean vector and covariance matrix is not sufficient to define the multivariate distribution.

Deutsch et. al. proposed a methodology to determine the conditional univariate distributions with only the mean and variance provided from kriging [3, 8]. Using this approach, the conditional marginal distribution of each variable can be obtained. The main challenge is then to identify the conditional multivariate distribution knowing these conditional *marginal* distributions.

Updating Technique to Obtain Conditional Multivariate Distribution

Let's review the information that is readily available: the original data distributions consisting of the global (or standard) univariate and multivariate distributions, and the conditional (or non-standard) univariate distributions obtained from solving the co-kriging system and the algorithm presented by Deutsch et. al. [3, 8]. To determine a non-standard multivariate distribution, a simple iterative updating procedure using all the available distributions is proposed.

Consider the bivariate case, where we have Y_1 and Y_2 data. The algorithm proceeds as follows:

1. Update the global bivariate distribution, $f_{Y_1,Y_2}(y_1, y_2)$, by scaling it by the ratios of the non-standard univariate conditional distribution, $f_{Y_i'}(y_i)$, to the standard univariate distribution,

$f_{Y_i}(y_i)$, $i = 1, \ldots, 2$.

$$f_{Y_1',Y_2'}(y_1, y_2) = f_{Y_1,Y_2}(y_2, y_1) \cdot \frac{f_{Y_1'}(y_1)}{f_{Y_1}(y_1)} \cdot \frac{f_{Y_2'}(y_2)}{f_{Y_2}(y_2)} \tag{2}$$

2. Calculate the new marginal Y_1 and Y_2 distributions that result from Equation 2 and reset $f_{Y_i}(y_i)$, $i = 1, \ldots, 2$ to these new marginals:

$$f_{Y_1}(y_1) = \int_{Y_2} f_{Y_1',Y_2'}(y_1, y_2) dy_2$$
$$f_{Y_2}(y_2) = \int_{Y_1} f_{Y_1',Y_2'}(y_1, y_2) dy_1$$

   Also, reset $f_{Y_1,Y_2}(y_1, y_2)$ to the new updated global distribution of Equation 2:

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1',Y_2'}(y_1, y_2)$$

3. Calculate corresponding summary statistics: mean, variance of marginal distributions, and covariance and correlation of (updated) bivariate distribution.

4. Check that the mean and variance of new marginal distributions $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ match those of the desired conditional distributions, $f_{Y_1'}(y_1)$ and $f_{Y_2'}(y_2)$, within some acceptable margin, ϵ. If this condition is not met, go to Step 1.

Scaling of the multivariate distribution to obtain a conditional multivariate distribution should reproduce complex, non-linear and/or heteroscedastic properties. Note that spatial heteroscedasticity of the simulated values (as mentioned in the section on kriging implications) is different from heteroscedasticity in the multivariate distribution at $\mathbf{h} = 0$ (for which this updating process will account).

Validation of Updating Approach

Consider a simple bivariate Gaussian global distribution with correlation ρ. Given parameters that specify a conditional distribution, the above updating methodology can be applied. For this purpose, a numerical exercise was devised with the following parameters:

- Global Distribution is standard bivariate Gaussian with correlation of $\rho_{global} = 0.30$.
- Conditional Distribution parameters:

$$\begin{aligned} E\{X\} &= 1.2 \\ Var\{X\} &= 0.25 \\ E\{Y\} &= 0.0 \\ Var\{Y\} &= 1.0 \end{aligned}$$

Furthermore, we know that restandardizing the marginal X and Y distributions to non-standard means and variances will not change the correlation of the resulting bivariate distribution.

The updating approach was applied and the corresponding results are shown in Figure 2. The reference global and desired conditional distributions are shown at the top (left and right, respectively). Three iterations were required to obtain an updated conditional distribution that reproduced the desired conditional univariate statistics; however, the resulting updated conditional distribution has a correlation of 0.161, not the 0.30 that would be found by rescaling a bivariate Gaussian distribution.

Unfortunately, re-standardizing the marginal distribution of X to some other non-standard Gaussian distribution and then calculating the resulting bivariate distribution does not amount to a conditional distribution in the conventional sense of the term. Rather, this generates a new global bivariate distribution with the desired univariate marginal distributions. Thus, the reference conditional distribution may be incorrect, that is, the expected correlation of 0.30 may not be correct. To investigate the difference between the correlation of the conditional distribution with that resulting from the iterative scaling approach, a slight shift in the thought process is proposed.

Building a Global Distribution from Conditional Distributions. Suppose that the global multivariate distribution is a linear combination of conditional multivariate distributions. This is analogous to supposing that these conditional distributions are subsets of the multivariate distribution, resulting from considering only a subset of the data in the domain. This idea is consistent with the practical application of (co)kriging, which only considers a subset of the data that are within some neighbourhood of the location of interest.

Without loss of generality, consider that there are m bivariate Gaussian conditional distributions, all with common correlation, ρ. The resulting global bivariate distribution is a linear combination of these m distributions (and will be non-Gaussian unless $m \rightarrow \infty$):

$$f(x,y) = \sum_{i=1}^{m} p_i \cdot f_i(x,y)$$

where $p_i, i = 1, \ldots, m$ are weights corresponding to $f_i(x,y), i = 1, \ldots, m$. The weights represent the contribution of each subset bivariate distribution to the global bivariate distribution.
The new global X marginal distribution, $f(x)$, has the following summary statistics:

$$\begin{aligned} E\{X\} &= \sum_{i=1}^{m} p_i \cdot E\{X_i\} && (3) \\ E\{X^2\} &= \sum_{i=1}^{m} p_i \cdot (\sigma_i^2 + E\{X_i\}^2) && (4) \\ Var\{X\} &= E\{X^2\} - (E\{X\})^2 && (5) \end{aligned}$$

Similarly, for the Y marginal distribution, $f_3(y)$,

$$\begin{aligned} E\{Y\} &= \sum_{i=1}^{m} p_i \cdot E\{Y_i\} && (6) \\ E\{Y^2\} &= \sum_{i=1}^{m} p_i \cdot (\sigma_i^2 + E\{Y_i\}^2) && (7) \\ Var\{Y\} &= E\{Y^2\} - (E\{Y\})^2 && (8) \end{aligned}$$

The resulting global bivariate distribution has the following covariance,

$$\begin{aligned} E\{XY\} &= \sum_{i=1}^{m} p_i \cdot E\{X_i Y_i\} && (9) \\ &= \sum_{i=1}^{m} p_i \cdot (Cov\{X_i Y_i\} - E\{X_i\}E\{Y_i\}) \\ Cov\{XY\} &= E\{XY\} - E\{X\}E\{Y\} && (10) \end{aligned}$$

$$\rho = \frac{Cov\{XY\}}{\sqrt{Var\{X\}Var\{Y\}}} \qquad (11)$$

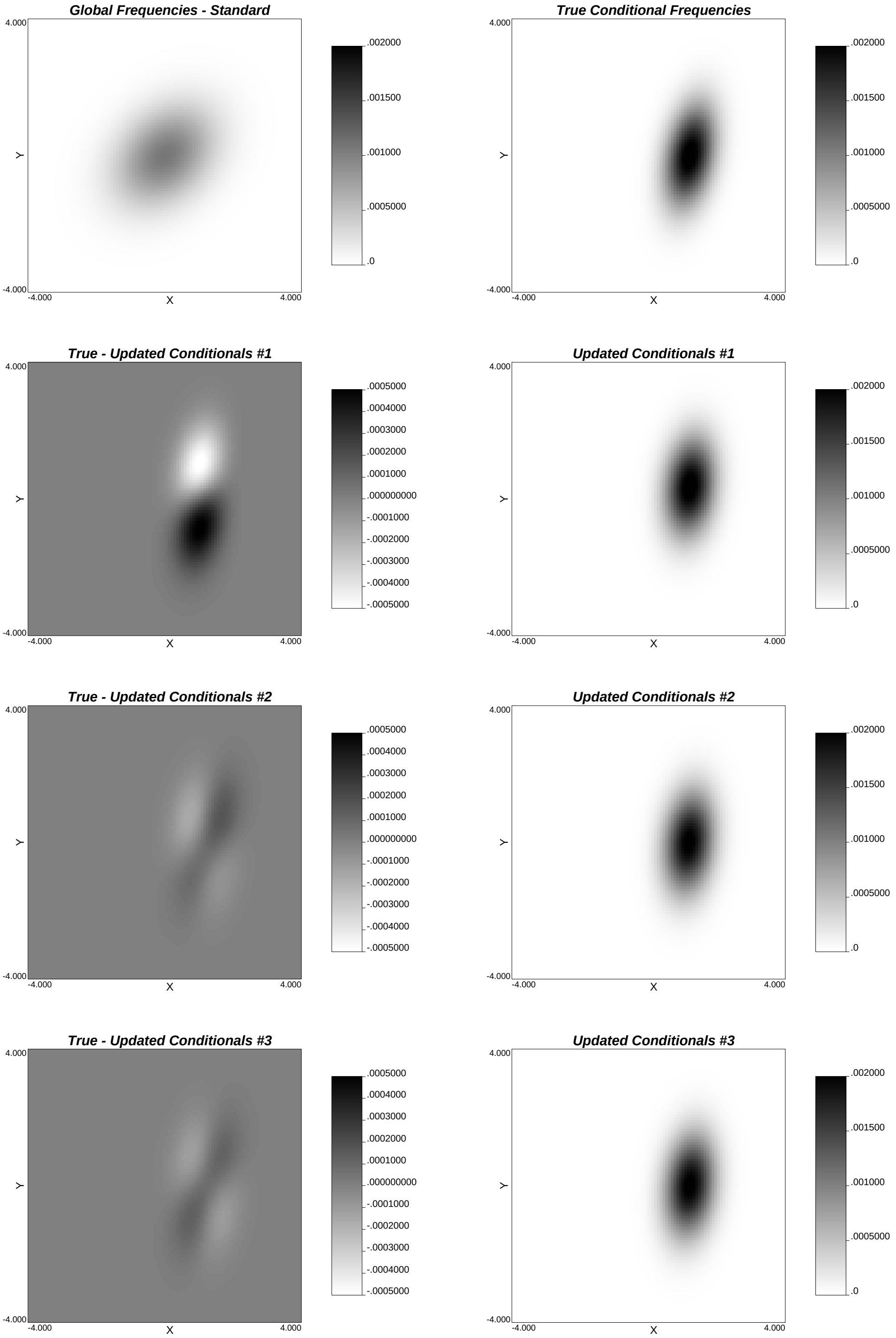

Figure 2: Results from applying iterative updating approach to a global standard bivariate Gaussian distribution with correlation of 0.30 (top right). Reference conditional univariate distribution for X is set with a mean of 1.20 and a variance of 0.25 (top left). Difference in bivariate probability distribution between reference conditional distribution and corresponding updated conditional distribution are shown in the remaining left side plots. The updated conditional distributions (for each iteration) are shown in the bottom 3 right side plots.

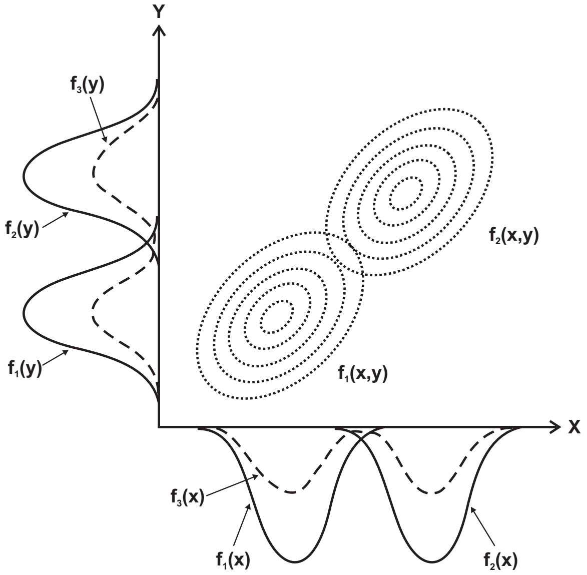

Figure 3: Schematic illustration of combining subsets of bivariate distributions, $f_1(x, y)$ and $f_2(x, y)$. The marginal X and Y distributions corresponding to each bivariate distribution are given as $f_i(x)$ and $f_i(y)$, with i corresponding to that distribution. The result of combining these distributions are two marginal distributions, $f_3(x)$ and $f_3(y)$.

Figure 3 is a schematic illustration for the case of $m = 2$; the resulting global bivariate distribution is bimodal and non-Gaussian.

Supposing that a global multivariate distribution can be constructed as a combination of conditional distributions, the decomposition of the global distribution to obtain a particular conditional distribution using the updating approach should be straightforward.

Numerical Example For this exercise, a bivariate global distribution was constructed by populating a bivariate field with 10 bivariate Gaussian kernels that will act as conditional distributions (that is, $m = 10$). Each kernel has a specified variance of 0.25 and a correlation of 0.5. The location of each kernel is chosen by drawing 10 locations from a bivariate standard Gaussian distribution with a correlation of 0.8.

In reality, the weights ($p_i, i = 1, \ldots, m$) required to calculate Equations 3 to 11 are unknown, and so some assumptions must be made. For illustrative purposes, let us assume that they are all equal, that is $p_i = 1/m, i = 1, \ldots, m$,

$$f(x, y) = \frac{\sum_{i=1}^{m} f_i(x, y)}{m} \tag{12}$$

Note that although the constituent kernel distributions are Gaussian, by Equation 12, the reference global bivariate distribution is not.

Using the updating algorithm, the global bivariate distribution can be scaled to reproduce the univariate statistics of each of the kernels. The resulting correlation of each conditional distribution (corresponding to each specified kernel) can be compared. Figures 4 and 5 shows the reference conditional distributions (or the kernels) on the right and its corresponding updated conditional on the left side (first and second set of five kernels are shown in each Figure respectively). Table 1 shows the correlation coefficient of the updated conditional distributions . From Figures 4 and

Kernel	**Reference Statistics**					**Updated Statistics**				
i	$E\{X_i\}$	$E\{Y_i\}$	$Var\{X_i\}$	$Var\{Y_i\}$	ρ	$E\{X_i\}$	$E\{Y_i\}$	$Var\{X_i\}$	$Var\{Y_i\}$	ρ
1	-1.24	-0.28	0.50	0.50	0.50	-1.2397	-0.2806	0.5000	0.5000	0.392
2	-1.64	-0.68	0.50	0.50	0.50	-1.6397	-0.6808	0.5000	0.5000	0.409
3	-0.44	-0.12	0.50	0.50	0.50	-0.4395	-0.1207	0.5001	0.4998	0.314
4	1.00	0.76	0.50	0.50	0.50	0.9993	0.7596	0.4999	0.4999	0.471
5	0.84	-0.12	0.50	0.50	0.50	0.8397	-0.1194	0.5000	0.5000	0.367
6	0.28	0.12	0.50	0.50	0.50	0.2804	0.1207	0.5000	0.5001	0.367
7	-0.60	-1.24	0.50	0.50	0.50	-0.6003	-1.2396	0.4999	0.5000	0.294
8	-2.04	-1.72	0.50	0.50	0.50	-2.0393	-1.7193	0.4997	0.4999	0.424
9	1.56	1.64	0.50	0.50	0.50	1.5594	1.6391	0.5000	0.5000	0.530
10	-1.64	-1.24	0.50	0.50	0.50	-1.6395	-1.2395	0.4999	0.4999	0.341

Table 1: Comparison of reference statistics with updated statistics as a result of setting each kernel distribution as the conditional distribution to be reproduced.

5, we see that the approximate shapes of the kernels are reproduced along with the conditional univariate summary statistics; however Table 1 shows that, as before, the reference correlations are not reproduced.

Although the conditional correlations are not exactly reproduced, a further check on the resulting global distribution was performed. For this task, the new global distribution is calculated as a linear combination of the updated conditionals via Equation 12. Figure 6 shows that the global bivariate and univariate distributions are reproduced, despite the fact that the reference conditional distributions were not reproduced exactly.

Proposed Methodology

The proposed algorithm for direct sequential cosimulation incorporates (1) the co-kriging formalism for multiscale data, (2) a conversion from conditional means and variances to conditional distributions, (3) an iterative updating technique to obtain the conditional multivariate distribution, and (4) the stepwise decomposition of this distribution for cosimulation of multiscale data. Specifically, the main steps of the sequential algorithm are:

1. Pick a random path visiting all locations.
2. At each location:
   (a) Search for all nearby data of different types and/or scale and previously simulated nodes.
   (b) Perform simultaneous cokriging (collocated or full) to determine the parameters corresponding to the conditional univariate distribution for each variable.
   (c) Using the cokriged parameters, determine the conditional univariate distribution for each variable using the approach proposed by Deutsch et. al. [3, 8]. These distributions will be referred to as the non-standard marginal distributions.
   (d) Determine a non-standard multivariate distribution via the iterative updating approach:

$$f_{Y_1',Y_2'}(y_1,y_2) = f_{Y_2,Y_1}(y_2,y_1)\cdot\frac{f_{Y_1'}(y_1)}{f_{Y_1}(y_1)}\cdot\frac{f_{Y_2'}(y_2)}{f_{Y_2}(y_2)}$$

   Resetting the global univariate and bivariate distributions to the previously updated distributions allows for iterative updating until the desired conditional univariate distributions are reproduced.

   (e) Draw from the non-standard multivariate distribution in a stepwise manner:
      i. Draw a simulated value y_1 from the conditional marginal distribution of $Y_1(y_1)$.

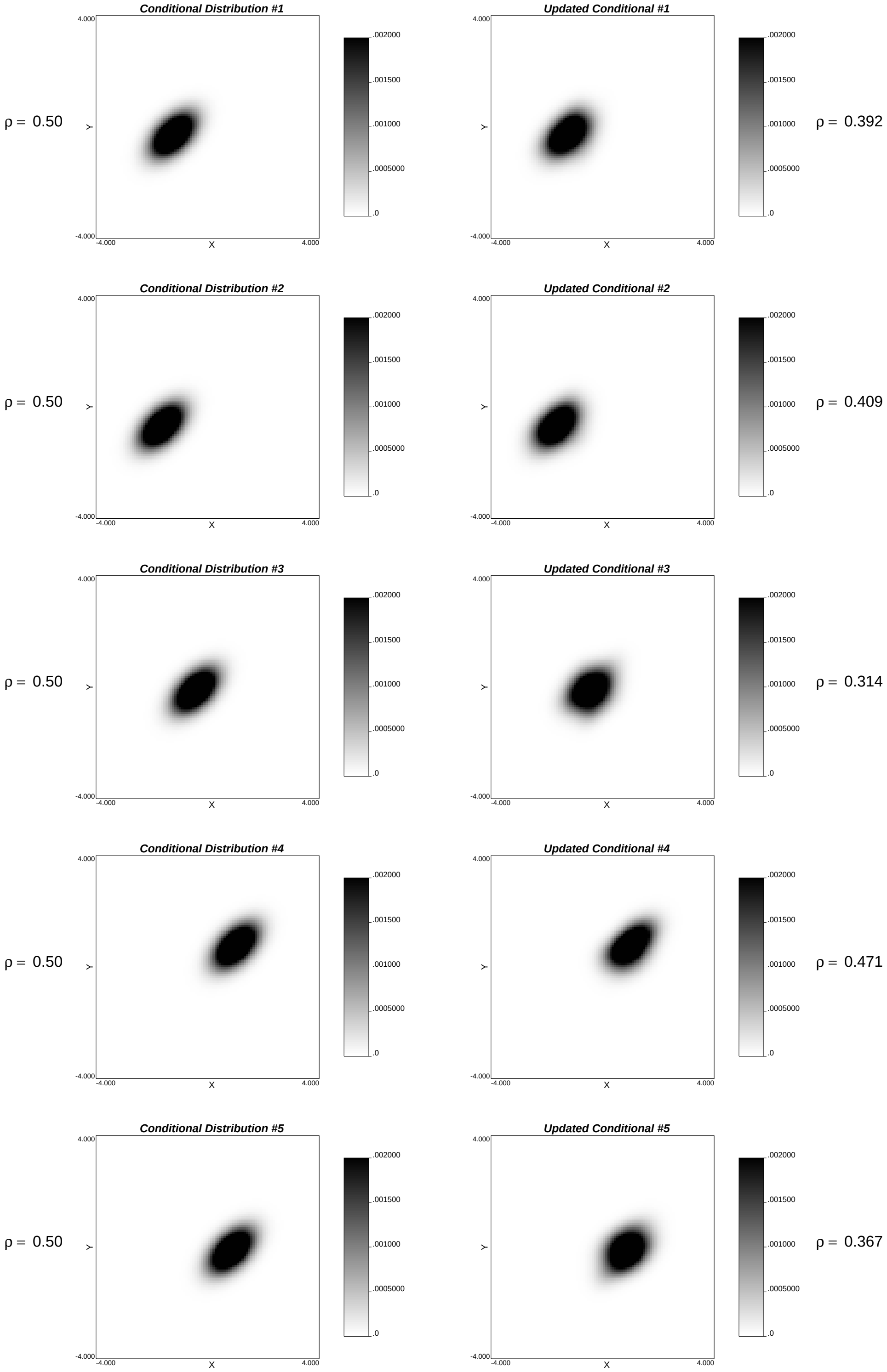

Figure 4: Kernels 1 to 5: Comparison of the reference conditional (kernel) distribution (left) and the updated conditional distribution (right).

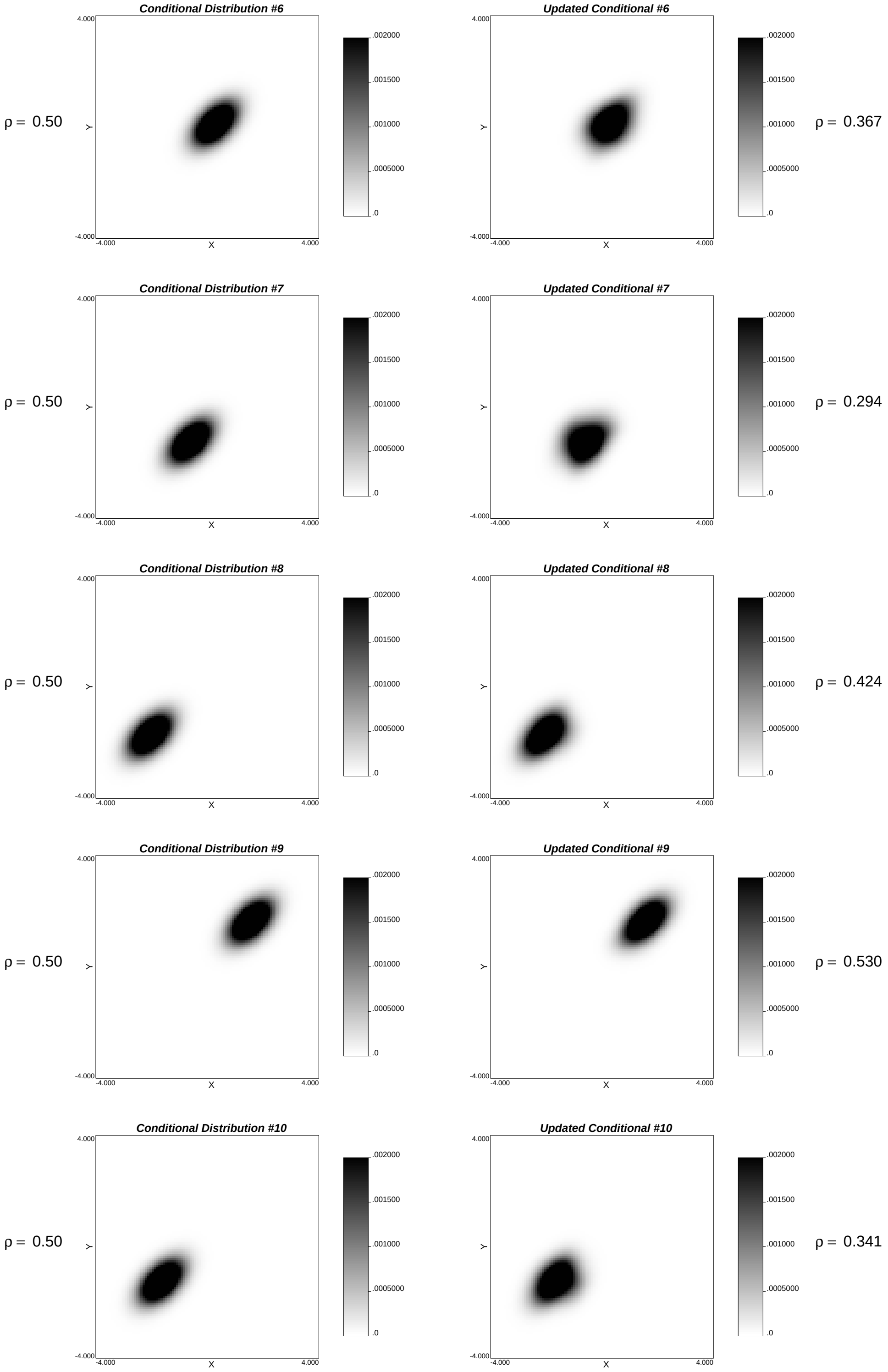

Figure 5: Kernels 6 to 10: Comparison of the reference conditional (kernel) distribution (left) and the updated conditional distribution (right).

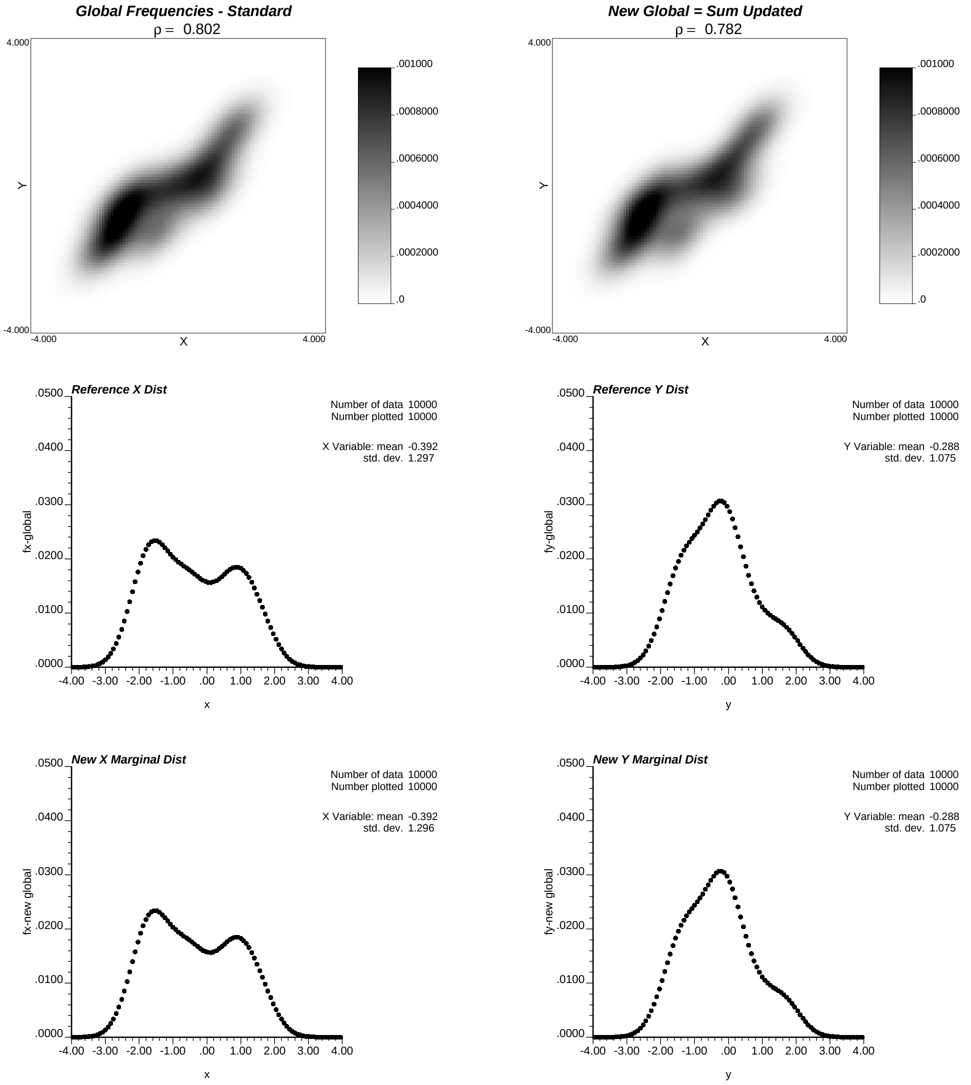

Figure 6: Comparison of reference global bivariate distribution (left) and new global distribution (right) resulting from weighted sum of updated conditional distributions. The middle row of histograms shows the reference X (left) and Y (right) histograms, and the bottom row shows the corresponding X and Y distributions determined from the new global distribution.

ii. From the conditional multivariate distribution determined in Step 2d, determine the conditional univariate distribution of $Y_2(y_2)$ given $Y_1 = y_1$, $f_{Y_2'|Y_1'} = y_1$. Draw y_2 from this conditional marginal distribution.

iii. Repeat Step (b) until a simulated value for each p variable is drawn.

(f) Proceed to next node.

Future Work

Naturally, implementation of the proposed methodology is the first item on the agenda. It will be interesting to see how this procedure performs when it is applied to real, complex, multivariate data. The scaling approach should produce conditional multivariate distributions that respect the features inherent in the global multivariate distribution. These features may include non-linearity, constraints, and/or heteroscedasticity.

This will also serve to illustrate the effect of *not* reproducing the conditional correlations exactly. Of course, with the use of real data, the reference conditional distributions are not known. The only real comparison that will be possible is to determine whether the global multivariate distribution is reproduced. Should these results be encouraging, further theoretical and practical issues associated with the implementation will be explored. Simulation results will also be compared to those produced by some of the more conventional multivariate simulation techniques.

Future research in determining a multivariate conditional distribution will continue. The consequences of relying on a simple assumption of a stationary ratio (as in the proposed updating equation) will be explored. Future work will involve identification, exploration and validation of other methodologies that may be applied in order to achieve this same objective.

References

[1] G. Bourgault. Using non-gaussian distributions in geostatistical simulations. *Mathematical Geology*, 29:315–334, 1997.

[2] J. Caers. Adding local accuracy to direct sequential simulation. *Mathematical Geology*, 32:815–850, 2000.

[3] C. Deutsch, T. Tran, and Y. Xie. A preliminary report on: An approach to ensure histogram reproduction in direct sequential simulation. Technical report, Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, March 2001.

[4] C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, New York, 2nd edition, 1998.

[5] A. G. Journel and C. J. Huijbregts. *Mining Geostatistics.* Academic Press, New York, 1978.

[6] B. Oz and C. Deutsch. Size scaling of cross-correlation between multiple variables. Technical report, Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, March 2001.

[7] B. Oz, C. Deutsch, and P. Frykman. A visual basic program for histogram and variogram scaling. Technical report, Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, March 2000.

[8] B. Oz, C. Deutsch, T. Tran, and Y. Xie. A fortran 90 program for direct sequential simulation with histogram reproduction. *Computers & Geosciences*, page submitted, 2001.

[9] T. Wawruch, C. Deutsch, and J. McLennan. Geostatistical analysis of multiple data types that are not available at the same locations. Technical report, Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, March 2002.

[10] W. Xu and A. G. Journel. Dssim: A general sequential simulation algorithm. In *Report 7, Stanford Center for Reservoir Forecasting*, Stanford, CA, May 1994.