Programs for MDE Modeling and Conditional Distribution Calculation

Sahyun Hong and Clayton V. Deutsch

Improved numerical reservoir models are constructed when all available diverse data sources are accounted for to the maximum extent possible. Integrating diverse data is not a simple problem because data show different precision and relevance to the primary variables being modeled, nonlinear relations and different types. Previous approaches rely on a strong Gaussian assumption or the combination of the source-specific probabilities that are individually calibrated from each data source. Gaussian techniques may be inappropriate when applying to data that show strong non-Gaussian characteristics. Probability combination approaches are vulnerable in theory and may require ad-hoc weight calibration. This work develops a multivariate analysis technique for data integration and implements the proposed method in a standalone GSLib type program. The method models the multivariate distribution without any distribution assumption. Issues such as nonlinearity, redundancy and data types are implicitly accounted for during the joint pdf modeling. The modeled multivariate distribution may violate the marginal constraints that the distribution should conform to. A sequential iteration algorithm was proposed to impose the constraints on the multivariate distribution. The robustness of the iterative algorithm is addressed. The developed methodology is applied to examples using the program. Parameter files used for the examples are specified in details. The application results show that the proposed method and program effectively integrate various secondary data.

1. Introduction

Building numerical reservoir models is an intermediate but essential step for reservoir management. Numerical models are used to plan new wells, calculate overall hydrocarbon reserves, predict the reservoir performance in a flow simulator, and analyze the uncertainty in reservoir performance forecasts. Thus, constructing geological model is an important step in reservoir management. Accurate reservoir modeling, however, is difficult to achieve given few data; the key reservoir properties such as facies, porosities, permeabilities and hydrocarbon saturations are typically sampled at very few well locations. These reservoir properties are heterogeneous and the distribution is never known exactly. Moreover, these properties are highly coupled with complex geological structures. For these reasons, all available diverse data should be integrated to the maximum extent possible for reservoir modeling.

The uncertainty in the reservoir model would generally decrease with additional data sources. Fortunately, various data are commonly available in petroleum applications; drilled well is the primary data source and seismic data or inverted seismic attributes are typical examples of supplementary data source in reservoir modeling. Geologic map data is also an important data. Data from wells are referred to the primary data and seismic data, geology map or other noncritical reservoir properties such as reservoir thickness and volume of shale are referred to the secondary data. The characteristics of the secondary data are that the data are exhaustively sampled over the modeling area and they need a calibration using the primary data because the secondary data are surrogate variables somewhat relating to the reservoir properties being estimated. Calibration or integration of the secondary data, however, is not easy because the available data are not only different in scales and types but also they are redundant and the data redundancy is not explicitly quantified. These difficulties prevent us from straightforwardly integrating the diverse data.

This paper presents a robust methodology and program for integrating multiple secondary data that have potentially varying types, scales, nonlinear relation and redundancy. The proposed methodology is based on the joint distribution modeling of the relevant variables in a nonparametric way. During the direct joint modeling, nonlinear relations between the variables and data redundancy are implicitly accounted for. Besides, mixed types of continuous and categorical variables can be naturally modeled by considering a categorical variable as continuous variable with particular outcomes. The conditional probability or probability distribution of our interest can be immediately derived if the joint probability distribution is modeled.

A nonparametric joint pdf modeling for data integration is not new (Duma and Fournier, 1988; Fournier and Derain, 1995; Saggaf et al., 2003). Previous studies investigated the applicability of different nonparametric modeling methods for reservoir property modeling using numerous seismic attributes. In this paper, we proposed to evaluate marginal constraints of the modeled joint distribution. These constraints are the conditions that make

the joint pdf legitimate. However, we demonstrated cases where the conditions are violated. To make a legitimate joint pdf, we proposed a simple and robust marginal fitting algorithm. The effectiveness of the proposed method is evaluated through the examples.

2. Method

Consider that X and Y_i, i=1,…,*m*, are the random variables representing the primary and *m* secondary variables. The random variable X can be continuous or categorical variable depending on what we want to estimate is continuous or categorical reservoir property. Y_i can be continuous or categorical variable depending on the type of given secondary data as well. The essential idea of the method is to nonparametrically model the joint relation of the primary and secondary variables. Once the multivariate distribution is modeled, the conditional probability or conditional estimate of the primary variable can be immediately derived by Bayes law:

$$f_{X|\mathbf{Y}}(x \mid y_1,...,y_m) = \frac{f_{X\mathbf{Y}}(x, y_1,...,y_m)}{f_{\mathbf{Y}}(y_1,...,y_m)} \tag{1}$$

The modeling of the multivariate distribution $f_{X\mathbf{Y}}(x,y_1,…,y_m)$ is based on the collocated samples of $(X,Y_1,…,Y_m)$ from wells. The kernel density estimation is used for nonparametric modeling. $f_{X\mathbf{Y}}(x,y_1,…,y_m)$ can be modeled as following by the kernel estimator (Cacoullos, 1966; Sain et al., 1992),

$$f_{X\mathbf{Y}}(x, y_1,...,y_m) = \frac{1}{nh_0 \times h_1 \times \cdots \times h_m} \sum_{i=1}^{n} K\left(\frac{x_i - X}{h_0}\right) \times \cdots \times K\left(\frac{y_{m,i} - Y_m}{h_m}\right) \tag{2}$$

where *n* is the number of collocated samples, $(h_0,…,h_m)$ are kernel bandwidths for each of $(X,Y_1,…,Y_m)$, K(·) is a univariate Gaussian kernel function. Equation (2) is known as the product kernel estimator for the multivariate modeling. The product kernel method is widely adopted in nonparametric multivariate pdf estimation (Scott, 1992). The choice of kernel size $(h_0,…,h_m)$ is crucial to the final multivariate distribution. Theoretical suggestion about the optimal kernel size is based on the number of samples, sample variance and dimension (Scott,1992):

$$h = \hat{\sigma} n^{-1/(d+4)} \tag{3}$$

The program `DataIntMDE` has a flexibility that allows users to adjust the kernel size. The kernel bandwidth becomes 1 if the given variable is categorical type.

Marginal Conditions of the Multivariate Distribution

The next step is for checking axioms of the modeled multivariate probability distributions: non-negative density functions, closure condition and reproduction of lower order marginal distributions. The kernel density estimator meets the first two axioms if the used kernel function K(·) follows $K(x)\geq 0$ and $\int K(x)dx=1$. The third condition, reproduction of lower order marginal distribution, is a marginality condition stating that *p*-variate joint distribution should reproduce *p′*-variate distribution where $p' < p$ and *p′*-variate distributions are the very well known. There are two very reliable *p*-variate lower order distributions in reservoir data integration: the univariate pdf of the primary variable $f_X(x)$ and the multivariate pdf of the secondary variables $f_{\mathbf{Y}}(y_1,…,y_m)$. The followings are possible marginal conditions that the modeled multivariate distribution $f_{X\mathbf{Y}}(x,y_1,…,y_m)$ must meet:

$$\int \cdots \int f_{X\mathbf{Y}}(x, y_1,...,y_m)dx = f_{\mathbf{Y}}(y_1,...,y_m) \tag{4}$$

$$\int \cdots \int f_{X\mathbf{Y}}(x, y_1,...,y_m)dy_1 \cdots dy_m = f_X(x) \tag{5}$$

The marginal condition in Eq (4) states that sum of the multivariate probability densities over the primary variable should amount to $f_Y(y_1,\ldots,y_m)$. The second marginal condition in Eq. (5) states that sum of the probability densities over the secondary variable should amount to $f_X(x)$. The global distribution of the primary variable $f_X(x)$ is achieved using well samples. Representative global distribution can be obtained by applying the declustering or debiasing technique if there is a bias in $f_X(x)$ caused by spatial clustering of well locations (Deutsch, 2002). The distribution of secondary variables $f_Y(y_1,\ldots,y_m)$ is modeled using the large number of samples that are exhaustively sampled and thus, the modeled $f_Y(y_1,\ldots,y_m)$ is very reliable.

Previous CCG papers (Hong and Deutsch, 2009; Hong, 2010) investigated the marginal constraints and they showed these conditions are not always met. Figure 1 illustrates this case. The iterative algorithm to match with the marginal conditions was proposed and its convergence was proved. This paper does not describe the details.

3. Program Details

The program is implemented as the Fortran 90 and based on the GSLib type that reads parameter file and run the executable file with the input parameter information. The following presents the details specific to input, output files and user input parameters. Table 1 shows the required input data files for the program. All input and output files are simplified Geo-EAS format constituting the header and data parts. Two input data files are needed for the program: well data file and secondary data file. Well data must contain the primary and all of secondary data samples and they should be identified by spatial coordinates. Secondary data file must be exhaustive type and each secondary variable is separated by column in the same file. Table 2 lists the user input parameters. Number of secondary variable is limited to the maximum five. Practically merging secondary variables more than 5 would be recommended because a large number of secondary variables are unnecessarily redundant without a significant improvement in the results. Type of the primary and secondary variable to be integrated should be specified; integer number either 0 or 1 identifies continuous or categorical variable. The number of modeling grids (n_x, n_y, n_z) should be specified and total number of grids must be same as the number of data line in the secondary data file. The representative global proportions should be input if the primary variable is categorical. The program models the global distribution with declustering weights if the primary variable is continuous. The program calculates the optimum kernel bandwidth using Eq. (3), however, users also can change the kernel size by kernel smoothing factors in Table 2. Table 3 summarizes the output files from the program: multivariate pdf without marginal correction, multivariate pdf with marginal correction, error report and result files. The initial and corrected pdf files contain joint probability densities given the binned value of the variables. Error file is a summary of the marginal errors at each iteration step. The final conditional probability (if primary variable is categorical) or conditional estimate and estimation variance (if primary variable is continuous) deriving from the updated joint distribution are saved in the result file. Some parameters other than user input are hard coded in the program. Maximum number of secondary variables, maximum bin number and maximum categories are reasonably limited as shown in Table 4. Iteration number for marginal correction is set as 100. Practices showed that the marginal errors were rapidly dropped during the first few iteration and the errors are usually converged into nearly zero % before 100 iterations. Figure 2 illustrates the input, output files and parameter information required for the program.

3. Examples

Figure 3 shows the simulated well data and secondary data for the example. Facies is sampled at four well locations with 1m vertical interval. Secondary data are simulated at every 1m×1m×1m over 30m×50m×20m. Figure 4 visualizes the results from each step of the method: (1) preparing the marginal distributions, (2) modeling the initial multivariate distributions, (3) updating the initial distributions by the marginal distributions, and (4) deriving facies probability. The updated distributions became stable after four iterations for matching marginal conditions. The final results are the facies probability cubes and they are shown in the bottom of Figure 7. The parameter file used for this particular example is shown in Figure 5. The choice of kernel smoothing factor in Line 14 is more influential than other parameters; the first value is used for the modeling of $f_{XY}(x,y_1,y_2,y_3)$, and the second value is used for the modeling of $f_Y(y_1,y_2,y_3)$. Although the program calculates the optimal kernel size, several attempts with $\pm$30% changes in the optimal values are recommended.

This paper focuses on introducing the MDE program. More elaborated examples can be seen in the previous CCG papers.

9. Conclusions

In reservoir modeling applications, diverse secondary data are frequently available for the modeling. The secondary data are not in the unit of the primary variable being estimated and thus, they should be calibrated or integrated with the primary variable. Challenges for integrating numerous data are in that they have different types, data redundancy and complex joint relations. This paper presents a robust method for secondary data integration and the related program. The proposed method is based on the nonparametric multivariate probability distribution modeling. By directly modeling the joint distribution between diverse data, different types are naturally accounted for and data redundancy is implicitly considered. The multivariate pdf modeling of the diverse data is not new. This paper addressed some marginal conditions that the multivariate pdf must meet and illustrated the case where the modeled multivariate distribution normally does not conform to the marginal conditions. To make the multivariate distribution legitimate, this paper proposed an iterative marginal fitting algorithm. The algorithm calculates the marginal differences and directly applies the differences to the initial joint distribution resulting in an updated distribution. The marginal fitting procedure is repeated with respect to the marginal distributions of the primary, and the secondary variables. The robustness of the iterative method is theoretically proved and practically tested through examples.

The idea of the proposed method is implemented in `DataIntMDE` based on GSLib type. Parameter files used for the demonstrated examples were described in details. The proposed method assumes that the multiple secondary data have the same grid size each other and they are homotopically sampled. The paper showed a categorical variable modeling as the primary variable from integrating secondary data, but a continuous variable modeling can be done as well by setting the parameter file appropriately.

References

Cacoullos, T., 1966, Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18:178-189.

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

Deutsch, C. V. and Journel, A. G., 1998, *GSLIB: geostatistical software library and user's guide*, Oxford Univeristy Press, New York.

Duma, J. and Fournier, F., 1988, Multivariate statistical analyses applied to seismic facies recognition Geophysics, 53, 1151–1159.

Fournier, F. and Derain, J. F., 1995, A statistical methodology for deriving reservoir properties from seismic data, Geophysics, 60, 1437–1450.

Krishnan, S., 2004, Combining diverse and partially redundant information in the earth sciences. PhD dissertation, Stanford University, Stanford, CA.

Saggaf, M. M., M, Toksoz, M. N., and Marhoon, M. I., 2003, Seismic facies classification and identification by competitive neural networks, Geophysics, 68, 1984–1999.

Scott, D. W., 1992, Multivariate density estimation: theory, practice, and visualization. John Wiley and Sons, Inc., New York.

Stain, S. R., Baggerly, K. A. and Scott, D. W., 1994, Cross-validation of multivariate densities. Journal of the American Statistical Association, 89:807-817.

Hong, S. and Deutsch, C.V., 2009, On secondary data integration, *Annual Report of Centre for Computational Geostatistics*

Hong, S., 2010, Multivariate analysis of diverse data for improved geostatistical reservoir modeling, *PhD. Thesis*, University of Alberta.

Tables

Table 1: Input files for the program

well data	• The well data should contain collocated primary and secondary data samples at well locations.
Secondary data	• This file should contain all of secondary variables being integrated and they should be separated by column. • The number of data line should be same as the number of modeling grids.

Table 2: User input parameters

Number of secondary variables	• The number of secondary variable is limited to the maximum 5.
Type of primary variable	• Integer number 1 indicates categorical type and 0 indicates continuous type
Type of secondary variable	• Integer number 1 indicates categorical type and 0 indicates continuous type
Grids definition	• Number of modeling grids in X,Y,Z directions is specified.
Global proportion	• Global proportions of categorical variable should be input and they should be declustered. • If the primary variable is continuous then the program skips reading this part.
Trimming limits	• User can define the range of eligible data to be used by setting trimming minimum and maximum
Kernel smoothing factor	• The program calculates the optimal kernel bandwidths using Equation (3) • Smoothing factors are multiplied with the calculated kernel bandwidths; factor being greater than 1 makes the distribution smoother and vice versa. • The program requires two smoothing factors. The first factor is for the multivariate pdf modeling of the secondary variables, and the second factor is for the multivariate pdf modeling of the primary and secondary variables.
Bin number	• The level of binning continuous variable is defined by the number of bins.

Table 3: Output files for the program

Initial MV pdf	• The joint pdf of the primary and all of secondary variables is saved in this file. • The marginal constraints are not applied. User can compare this file to the updated MV pdf file.

Updated MV pdf	• The updated joint pdf is saved in this file.
Marginal errors	• Calculated marginal errors are reported in this file
Result	• Conditional probability of facies or conditional mean and variances are extracted from the updated joint pdf. This results are in (X,Y,Z) space and thus, user can directly import this file in any 3D visualization software.

Table 4: Hard coded parameters of the program

Maximum bin number	100
Maximum number of secondary variable	5
Maximum iteration number	100
Maximum number of categories	6

Figures

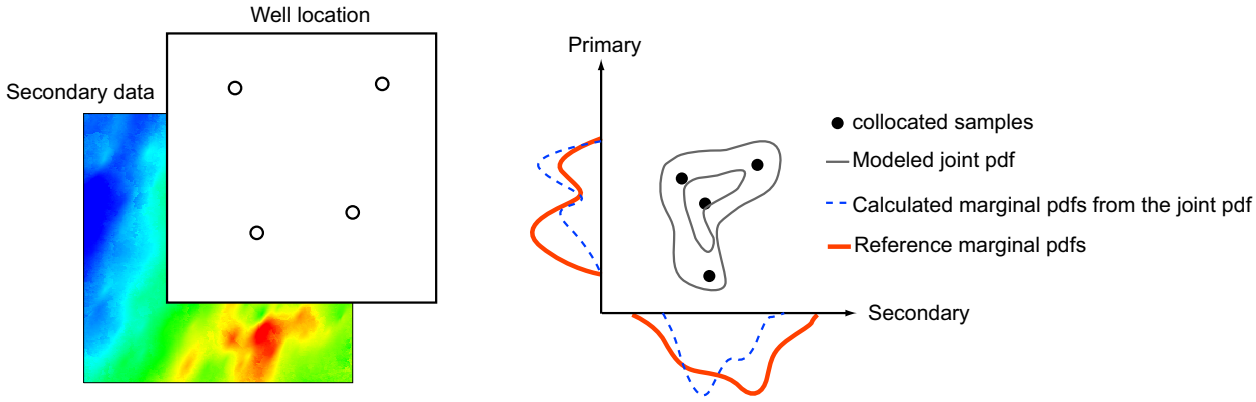

Figure 1: Schematic illustration for the inconsistency between the reproduced marginal distributions and the reference marginal distributions.

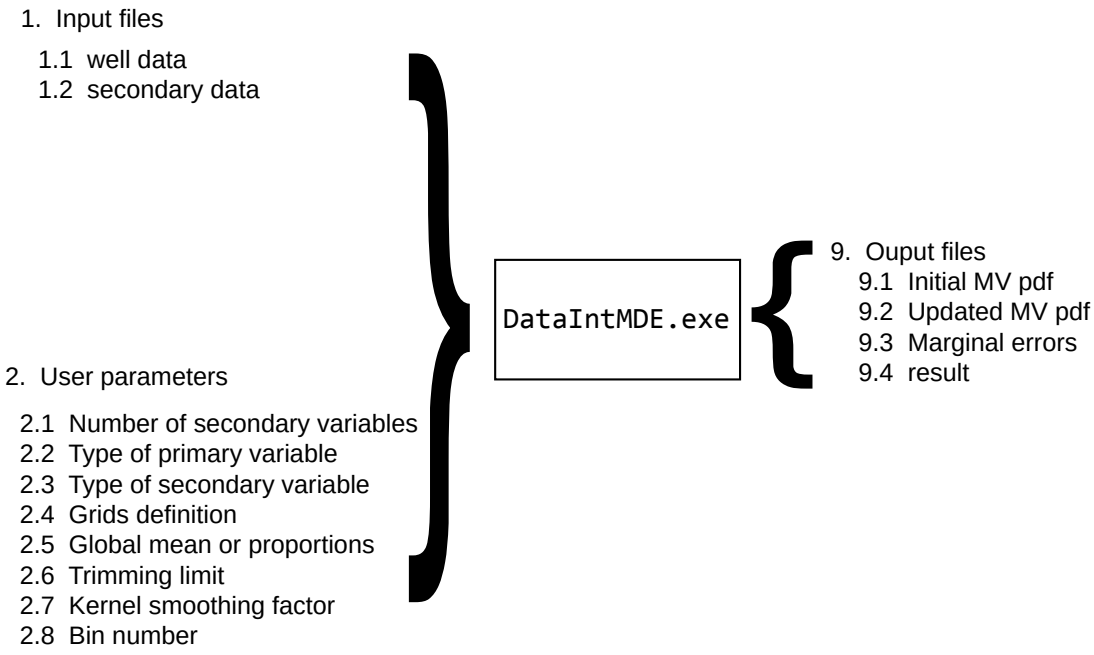

Figure 2: Input files, output files and user input parameters for the program.

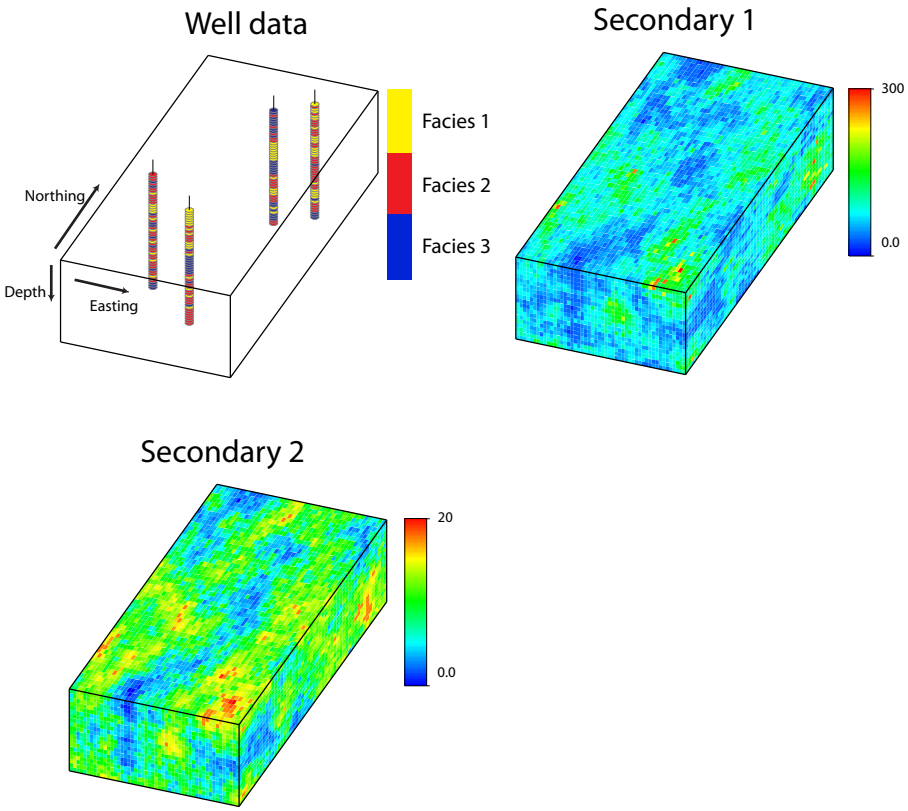

Figure 3: Simulated well data and two continuous secondary data.

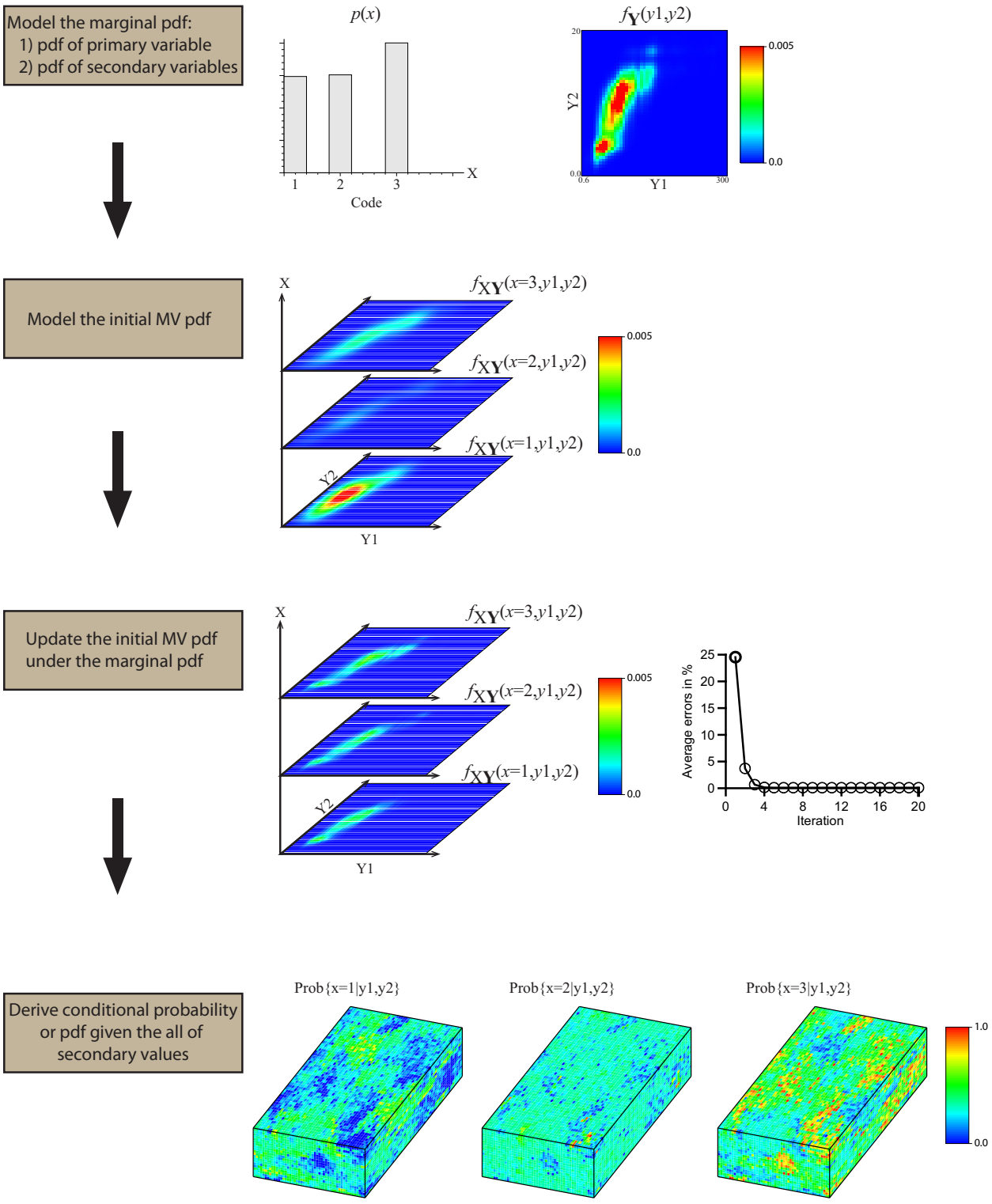

Figure 4: Procedure of the proposed method (diagrams in the left) and results obtained from each process (in the right).

```
                        Parameters for Example 1
                        **********************
Line1:2               -Number of Sec Variables
Line2:1               -Primary variable type; 0-cont, 1-cat
Line3:3               -Number of categories
Line4:1 2 3           -Code if categorical var is specified
Line5:0.3 0.3 0.4     -Declustered proportions
Line6:secodnary.dat   -Secondary data
Line7:1 2             -Column for each secondary variable
Line8:0 0             -Type of each secondary var;1-cat,0-cont
Line9:30 50 20        -Grids definition
Line10:well.dat       -Well data
Line11:1 2  4  0      -Column for sec, pri and declustering weights
Line12:-99 99         -Trimming min, trimming max
Line13:50             -Bin number
Line14:1.0  0.8       -Smoothing factor for kernel window
Line15:iMV.out        -MV pdf without marg correct
Line16:uMV.out        -MV pdf with marg correct
Line17:error.out      -Error vs iteration
Line18:result.out     -Final result
```

Figure 5: Parameter file for the example.