Programs for facies modeling with DMPE

Yupeng Li and Clayton V. Deutsch

The programs needed for direct multivariate probability estimation (DMPE) are introduced in detail. The first program (TPcalc) is used to calculate the bivariate probability from a vertical profile or from a training image along a certain directions. The second program (TPdiagram) is used to plot the bivariate probability or transition probability. The calculated bivariate probability will be used as input for the programs to do estimation/simulation with DMPE. The program (DMPEest) is specially for doing estimation and cross validate. The program (DMPEsim) can do estimation and simulation when the hard data can be moved into the grid nodes. Also the core algorithm of the DMPE is in the program of the (DMPEsingle) which can be used to estimate one discrete multivariate probability from the input constraints. The single program that can do marginalization with a known discrete multivariate probability is also given in a separate program (MVmarg). Using the these last two programs, the marginalization and its inverse procedure can be tested with a given data set or to estimate the multivariate probability for other purposes.

1 Bivariate Probability Calculation

The bivariate probability for two locations $\mathbf{u}_\alpha, \mathbf{u}_\beta$ is defined as $p(\mathbf{u}_\alpha, \mathbf{u}_\beta)$. If there are K categories for each location, there would be a matrix to describe all the bivariate probabilities $p(\mathbf{u}_\alpha, \mathbf{u}_\beta)$, named as bivariate probability matrix. The bivariate probability for two locations would be different from the classical bivariate probability in mathematics. In mathematics, the bivariate probability for two variable a, b would be $p(a, b)$ which is symmetric. While for spatial statistics, it is asymmetric character. The location $\mathbf{u}_\alpha$ and location $\mathbf{u}_\beta$ are not inter-changeable. The transition probability would be defined as a conditioning probability $p(\mathbf{u}_\alpha|\mathbf{u}_\beta)$. The same as bivariate probability, it would be a $K * K$ matrix for each pair of locations[1, 2].

Given the stationary assumption, the bivariate probability matrix and transition probability matrix for two locations separated by distance $\mathbf{h}$ would compose a diagram as those two locations departing away from each other. Also, the calculation would only be performed in highly density sampled data set such as vertical well profile or training image in order to get enough replication for a reliable probability reference.

In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables[3] and it is used in many research areas such as knowledge discovery and data mining[4]. The mutual information of two discrete random variables S_i and S_j is defined as:

$$R(S_i; S_j) = \sum_{s_i} \sum_{s_j} p(s_i, s_j) \log \left(\frac{p(s_i, s_j)}{p(s_i)p(s_j)} \right) \tag{1}$$

where $p(s_i, s_j)$ is the joint probability of s_i and s_j, $p(s_i)$ and $p(s_j)$ are the univariate marginal probability of $\mathbf{u}_i$ and $\mathbf{u}_j$ respectively. More details on this point are in paper 122 of this volume.

A *GSLIB*[5] style of program *TPcalc* is developed to calculate the bivariate probability matrix(Transition probability matrix). The definition of bivariate probability matrix diagram is explained in the paper 123 of this volume. The parameter file is shown in Figure 1.

The program will calculate the bivariate probability/transition probability from profile or from picked direction of a training image. In line 1 and 2 are the category number and category types that

exist in the domain.
In line 3 is an indicator to define the bivariate probability/transition probability will calculate form well data or from training image.
If it will be calculated from well data, in line 4 and 5 will be the file name and the specific column number(well ID, depth and category) needed in the program.
If it will be calculated from the training image, in line 6 , 7 and 8 will be the training dimension definition and its file name.
Line 9 will be the scan direction. It could be along x or along y or both of them. The test shows that scan from X and Y directions would improve the reproduction of training image pattern.
Line 10 is length and number of the count interval for well data and training image.
The left lines are different output files from the program. Line 11 is the output of the bivariate probability $p(\mathbf{u}_\alpha, \mathbf{u}_\beta)$ and transition probability $p(\mathbf{u}_\alpha|\mathbf{u}_\beta)$ for DMPE program.
Line 12 is the out put of bivariate probability $p(\mathbf{u}_\alpha, \mathbf{u}_\beta)$ for plotting in program *TPdiagram*.
Line 13 is the transition probability $p(\mathbf{u}_\alpha|\mathbf{u}_\beta)$ output for plotting. The output for estimation/simulation from line 12 is in different format as those used for plotting from line 13.
Line 14 is the debug information output file name.
Line 15 is specially for the mutual information calculation along the picked direction in plotting format.

2 Bivariate Probability Plotting

With the increase of distance between a locations pair $(\mathbf{u}_\alpha, \mathbf{u}_\beta)$, the direct bivariate probability $p(S(\mathbf{u}_\alpha) = s_\alpha, S(\mathbf{u}_\beta) = s_\beta), s_\alpha = s_\beta$ and cross bivariate probability $p(S(\mathbf{u}_\alpha) = s_\alpha, S(\mathbf{u}_\beta) = s_\beta), s_\alpha \neq s_\beta$ has different increase characteristics which is dependent on the geological pattern of the domain[6].

In the DMPE program, the experimental bivariate probability or transition probability matrix would be used directly in the program without modelling as the variogram modeling. So the plotting and doing some visual checking is very important. The plotting for visual checking or geological pattern recognition can be done using the program *TPdiagram*. The plotting file is modified from the codes *vargplt*. The parameter file is shown as in Figure 2.

The output of the Bivariate Probability and Transition Probability from program *TPcalc* is ordered as the head category first then it is the tail category. For example it there are three categories $1, 2, 3$. The bivariate probability/transition probability would output in an order of $1 \rightarrow 1, 1 \rightarrow 2, 1 \rightarrow$

```
                     parameter file for TPcalc
                    *************************
Line number     START OF PARAMETERS:
1                  3                   -number of catgory
2                 1   2  3            -category types
3                 0                   -0: from well data;1: from training image
4                 wells.out           - input markov chain data file
5                 1   2   3           - well ID,depth, category codes column
6                 50   50   1         -the dimension of training image you would scan
7                 0.5  0.5            -the origin coordinated of training image
8                 true_cat.dat        - input training image you want to scan
9                 0                   -which direction do you want to scan? 0:x; 1:y.
10                0.1      40         -transition count interval and  interval number
11                mde3.out            -output  bivariate for DMPE
12                    biv3.out            -output bivariate probability for diagram plotting
13                tp3.out                 -output transition probability for diagram plotting
14                    dbg3.out             -debug file
15                    mui.out              -the calculated mutual information
```

Figure 1: The parameter file used in bivariate diagram calculation program

$3, 2 \to 1, 2 \to 2, \ldots, 3 \to 2, 3 \to 3$. For example one output file of bivariate probability for plotting is shown in Figure 3.

3 DMPE Estimation Programs

In geostatistics, the core problem is calculation a conditional probability distribution $p(\mathbf{u}_0|\mathbf{u}_1, \cdots, \mathbf{u}_n)$. In DMPE, it is obtained from its definition as :

$$p(\mathbf{u}_0|\mathbf{u}_1, \cdots, \mathbf{u}_n) = \frac{p(\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_n)}{p(\mathbf{u}_1, \cdots, \mathbf{u}_n)} \tag{2}$$

In Equation (2), the multivariate probability $p(\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_n)$ is estimated directly from its constrains of bivariate probability between all the data pairs as $p(\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_n) = F(p(\mathbf{u}_\alpha, \mathbf{u}_\beta), \alpha, \beta = 0, \cdots, n)$. The iterative scaling solution is a nonlinear approach compared to the traditional indicator kriging approach. The estimated multivariate probability also has the Maximum Entropy property from the input marginal constraints. The details of this point is illustrated in paper 119,120 and paper 121 in this volume. After the bivariate probability/transition probability is calculated, the output file from *TPcalc* is ready to use in the spatial estimation and simulation. The DMPE are implemented into two programs: *DMPEest* and *DMPEsim*.

The program *DMPEest* is used to do estimation or cross validation with the input hard data. The parameter file of *DMPEest* is shown in Figure 4. Here are some explanations of the parameter files.
Line 1 is the estimation option. 0 is for doing estimation on grid, 1 is doing cross validate for all the hard data locations.
Line 2 , 3 and 4 are the number of category type, and its proportion. One note here is that the order of category in line 3 should be the same as the order of category in line 2 of parameter file for program *TPcalc* shown in Figure 1.
Line 5 and 6 are the information of hard data file.
Line 7 is the debug level and the debug information will output to the file name listed in line 8.
Line 9 is the estimation/cross validate results file name.
Line 10, 11 and 12 are the grid definition.
Line 13 is the conditioning data number for DMPE. This number should be small than 8 because of huge multivariate probability space.
Line 14 and 15 are the searching radius and searching angle.
Line 16 would be used when octant searching is used.

```
                                 Parameters for TPdiagram
                                **************************
Line #       START OF PARAMETERS:
1            bivariate.ps         -file for PostScript output
2            1                    -number of variograms to plot
3            0.0   -20.0          -distance  limits (from data if max<min)
4            0.0    -1.2          -bivariate limits (from data if max<min)
5            0      1.0            -plot sill (0=no,1=yes), sill value)
6            bivariate probability  -Title for bivariate probability
7            biv3.out              -1 file with variogram data
8            1    1   1   1    1    -TP #, dash #, pts?, line?, color
9            biv3.out              -2 file with variogram data
10           3    3   0   1   10    -TP #, dash #, pts?, line?, color
```

Figure 2: The bivariate probability/transition probability diagram plotting program

Line 17 is the bivariate probability matrix file name'
Line 18 is the lag interval number and its length used in transition probability calculation.
Line 19 is the anisotropy ratio used to transform the spatial distance to the distance used in Transition probability calculation.
Line 20 is the minimum iteration time used to do the multivariate probability estimation.

4 DMPE Simulation Programs

The classical sequential simulation approach is used. Assuming the desired multivariate probability is $p(\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_n)$, the sequential simulation is based on that the multivariate probability can be expressed as a chain of only conditional probabilities as:

$$\begin{aligned} p(\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_n) &= p(\mathbf{u}_n|\mathbf{u}_{n-1}, \cdots, \mathbf{u}_1, \mathbf{u}_0) \times p(\mathbf{u}_{n-1}, \cdots, \mathbf{u}_1, \mathbf{u}_0) \\ &= p(\mathbf{u}_n|\mathbf{u}_{n-1}, \cdots, \mathbf{u}_1, \mathbf{u}_0) \times p(\mathbf{u}_{n-1}|\mathbf{u}_{n-2}, \cdots, \mathbf{u}_1, \mathbf{u}_0) \times p(\mathbf{u}_{n-2} \cdots \mathbf{u}_1, \mathbf{u}_0) \\ &= \cdots \\ &= \prod_{i=0}^{n} p(\mathbf{u}_i|\mathbf{u}_{i-1}, \cdots, \mathbf{u}_0) \end{aligned} \tag{3}$$

During the simulation, the previous simulated nodes will be used as hard data for later unsampled location conditional probability calculation and the Monte-Carlo simulation. Because of the huge multivariate probability space requirement. In this program the conditioning data number for one unsampled location should be no more than 10. Also, in the parameter file the category list in Line 3 as in Figure 5 should be in the same order at that listed in parameter file for bivariate probability calculation program.

In this program *DMPEsim*, the hard data are enforced to grid nodes center in simulation. When the move of hard data to grid nodes is permitted or can be neglected, the program can also do estimation. The parameter file is shown in Figure 5.

Line 1 is given the indicator of estimation of simulation. The number 0 will do estimation. Other great than 0 will be the simulation realization number.
Line 2 and 3 are the category number and its value of the discrete random variable. Line 4 are the global proportion of the categories in the domain.
Line 5 and 6 are the input of hard data file name and the relative column number.
Line 7 is the bivariate probability file name and line 8 is the counting interval total number and its length *xlag* in building the transition probability/transition probability matrix.
Line 9 is the anisotropy ratio for three directions which is used to relate the 3D spatial distance to the length of TP calculation interval. For example, in one dimension, if it is said that $anis1 = 30$, which means that along x direction, the effective distance of h_x would be equal to $h_x/anis1$ in TP calculation space. The details of this point is illustrated in paper 123 of this volume.
Line 10 is the iterative time used in iterative scaling (**IS**) approach when doing the full multivariate probability estimation. Usually, it will convergence to the stable solution after 30 times iteration.
Line 11 and 12 are the debug level and output file for the program
Line 13 is the estimation/simulation output file of the DMPE;
Line 14, 15, and 16 is the estimation.simulation grid definition; Line 17 is the random number seed for random path build when doing simulation;
Line 18 is the maximum conditioning nodes in estimation and simulation. As the constraint of huge probability data event space, it should be no more than 11 given the category is less than three.

Line 19 is the set for octant searching setting;
Line 20 is the minimum search radius;
Line 21 is the search angles for search ellipsoid.
These six parameters in line 20 and line 21 will decide how far the conditioning data will be used in the program;
Line 20 is the searching table built for searching. In the program implementing, the searching table is used to quick searching conditioning data taking considering the searching radius and searching ellipsoid directions.

5 DMPE Subroutines

In the implementation of DMPE for spatial estimation and simulation, there are some subroutine that could be useful for other purposes. Here are two of them that are given separately. The first one is one discrete multivariate probability estimation from the bivariate probability constraints. In the *DMPEest* and *DMPEsim*, the multivariate probability is explicitly estimated and used to calculate the conditional probability. It could be used outside the program.

In this program, one discrete multivariate probability can be obtained from the input bivariate probability. The pair-wise bivariate probabilities are input for all the variables. The output is the estimated multivariate probability. The order of bivariate probability is the data variables index. This program can be used to test the iterative scaling method in the multivariate probability estimation. The program *DMPEsingle* is used to do this estimation. The parameter is shown in Figure 6.

The second one is to consider any order marginal probability operation with the known multivariate probability. The most often used one is the univariate probability and bivariate probability. The univariate probability distribution $P(\mathbf{u}_\alpha)$ which will characterize the distribution of random variable $S(\mathbf{u}_\alpha)$, can be calculated from the multivariate probability distribution $p(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ as:

$$P(\mathbf{u}_\alpha) = \sum_{\mathbf{u}_1=s_1}^{s_K} \cdots \sum_{\mathbf{u}_\alpha} \cdots \sum_{\mathbf{u}_n=s_1}^{s_K} P(\mathbf{u}_1, \cdots, \mathbf{u}_n) \quad \alpha = 1, \cdots, n \tag{4}$$

A second order marginal probability distribution $P(\mathbf{u}_\alpha, \mathbf{u}_\beta), \alpha \neq \beta$ and $\alpha, \beta = 1, \cdots, n$ is calculated from multivariate probability distribution as:

$$P(\mathbf{u}_\alpha, \mathbf{u}_\beta) = \sum_{\mathbf{u}_1=s_1}^{s_K} \cdots \sum_{\mathbf{u}_\alpha} \cdots \sum_{\mathbf{u}_\beta} \cdots \sum_{\mathbf{u}_n=s_1}^{s_K} P(\mathbf{u}_1, \cdots, \mathbf{u}_n) \tag{5}$$

The bivariate marginal probability will satisfy $0 \leq p(\mathbf{u}_\alpha, \mathbf{u}_\beta) \leq 1$ and $\sum_{\mathbf{u}_\alpha=e_1}^{e_K} \sum_{\mathbf{u}_\beta=e_1}^{e_K} p(\mathbf{u}_\alpha, \mathbf{u}_\beta) = 1$. The program *MVmarg* is used to do marginalization with a known multivariate probability. The parameter file is shown in Figure 7. The category number and the variable number are input in line 1 and line 2. The univariate and bivariate probability will be written into the specified file.

References

[1] Steven Carle and Graham Fogg. Transition probability-based indicator geostatistics. *Mathematical Geology*, 28:453–476, 1996. 10.1007/BF02083656.

[2] Steven Carle and Graham Fogg. Modeling spatial variability with one and multidimensional continuous-lag markov chains. *Mathematical Geology*, 29:891–918, 1997. 10.1023/A:1022303706942.

[3] Themas M. Cover and Joy A Thomas. *Elements of information theory.* Wiley-Interscience, 2006.

[4] Y. Y. Yao. Information-theoretic measures for knowledge discovery and data mining. In Karmeshu, editor, *Entropy Measures, Maximum Entropy and Emerging Applications*, pages 115–136. Springer, 2003.

[5] C.V. Deutsch and A.G. Journel. *GSLIB: Geostatistical software library and user's guide.* Oxford University Press, 1998.

[6] Weidong Li. Transiogram: A spatial relationship measure for categorical data. *International Journal of Geographical Information Science*, 20(6):693–699, 2006.

```
Bivariate Probability for category   1   1
   1     1.000  0.1236734693877551
    .......................
   40    40.00  0.0747826086956522
Bivariate Probability for category   1   2
   1     1.000  0.0681632653061224
    .......................
   40     40.0  0.0943478260869565
Bivariate Probability for category   1   3
   1     1.000  0.0028571428571429
    .......................
   4     40.00  0.0221276595744681
Bivariate Probability for category   2   1
   1     1.000  0.0612244897959184
    .......................
   40     40.0  0.0943478260869565
Bivariate Probability for category   2   2
   1     1.000  0.2151020408163265
    .......................
   40     40.0  0.0943478260869565
Bivariate Probability for category   2   3
   1     1.000  0.0906122448979592
    .......................
   40     40.0  0.0943478260869565
Bivariate Probability for category   3   1
   1     1.000  0.0053061224489796
    .......................
   40     40.0  0.0943478260869565
Bivariate Probability for category   3   2
   1     1.000  0.0816326530612245
    .......................
   40     40.0  0.0943478260869565
Bivariate Probability for category   3   3
   1     1.000  0.3514285714285714
    .......................
   40     40.0  0.0943478260869565
```

Figure 3: One example of the bivariate probability output file for plotting

```
                             Parameters for DMPEest
                            ***********************
Line #   START OF PARAMETERS:
1          1                            -option: 0=grid, 1=cross
2          3                            -number thresholds/categories
3            1   2    3                 -categories
4          0.21  0.29  0.50             -global pdf
5          cluster.out                  -file with data
6          1  2    3    4   5           -columns for DH,X,Y,Z and variable
7          0                            -debugging level: 0,1,2,3
8          dmpe_est.dbg                 -file for debugging output
9          dmpe_est.out                 -file for DMPE output
10         10   0.5    1.0              -nx,xmn,xsiz
11         10   0.5    1.0              -ny,ymn,ysiz
12         1    0.0    1.0              -nz,zmn,zsiz
13         6                            -maximum conditioning data for DMPE
14         10.0  10.0  2.0              -maximum search radii
15         0.0   0.0   0.0              -angles for search ellipsoid
16         0                            -max per octant (0=not used)
17         truexy_mde3.out              -the bivariate marginal file name for DMPE
18         40  1                        -the total lag in TP calculation & interval length
19         3.0   3.0   1                -the anisotropy transform ratio
20         20.0                         -minimum iteration time for DMPE
```

Figure 4: The DMPEest program parameter file

```
                                   Parameters for DMPEsim
                                  **********************
Line   #     START OF PARAMETERS:
1             0                         -operation(0:estimation; >=1:simulation number)
2             3                         -number of categories
3             1   2   3                 -category types
4             0.20   0.3   0.5          -global proportions  of each category
5             sample_true.dat           -file with local data
6              1  2  3   4              -columns for X,Y,Z, and category
7             truexy_mde3.out           -the bivariate marginal file name for DMPE
8             40  1                     -total counting interval number in  TP calculation & its length
9             30.0   30.0   1           -the anisotropy transform ratio
10            3.0                       -minimum iteration time for DMPE
11            0                         -debugging level: 0,1,2,3,4
12            DMPE_rs.dbg               -file for debugging output
13            rs0813-est-1.out          -file for estimation/simulation output
14            50    0.5   1.0           -nx,xmn,xsiz
15            50    0.5   1.0           -ny,ymn,ysiz
16            1.0    1    1             -nz,zmn,zsiz
17            31199                     -random number seed
18            5                         -maximum conditioning nodes for DMPE
19            0                         -maximum data number per octant    (0=not used)
20            100.0   100.0   1.0       -maximum search radius
21            0.0    0.0    0.          -angles for search ellipsoid
22            21     21     1           -size of searching table
```

Figure 5: The DMPE simulation program program parameter file

```
        parameter file for DMPEsingle
       *************************
START OF PARAMETERS:
3                       -number of catgory
6                        -category types
bivprob.out                   -input known mv at here
dmpe.dbg                      -output the debuging information
mvprob.out                       -univariate probability marginal file
```

Figure 6: The parameter file of one multivariate probability estimation program

```
        parameter file for MVmarg
       *************************
START OF PARAMETERS:
3                       -number of catgory
6                        -category types
mvprob.out                     -input known mv at here
marg.dbg                      -output the debuging information
bivprob.out                     -bivariate probability marginal file
uivprob.out                     -univariate probability marginal file
```

Figure 7: The parameter file of marginal program