Integrating Secondary Data and Training Images with Direct Density Estimation

Sahyun Hong and Clayton V. Deutsch

Data integration requires the multivariate distribution between the considered data sources. Probability combination schemes including permanence of ratios, the tau model, the nu model and the lamda model have received much attention recently. This approach involves the combination of each calibrated probability conditioned to individual data source to approximate the joint probability, which is termed indirect estimation method. The main challenge with these probability combination schemes is fair consideration of redundant data. Directly estimating the joint probability between variables meets this challenge. A procedure for integrating soft secondary data and training image (TI) is presented. Previous studies have mainly focused on probability combination approaches in order to integrate secondary and TI. However, in this paper we suggested an idea of directly estimating multivariate probability conditioned to secondary data and TI simultaneously.

Introduction

Geostatistical data integration is an important subject in petroleum reservoir characterization. It is desirable to reproduce all of data to model reservoir with less uncertainty and high accuracy. The data available for geostatistical modeling can often be divided into (1) direct measurements of the primary variable being predicted, (2) secondary data such as seismic attributes, and (3) analogue or geologic information in the form of training images in geostatistics. The resulting probability is a joint probability conditioned to the relevant data and it will be calculated at every visited node.

Previous probabilistic data integration studies including permanence of ratios (PR model), tau model, nu model and lamda model involve two step approaches: each datum is individually calibrated, and they are combined to approximate the joint probability conditioned to all data sources. Naïve combination function is PR model that used conditional independence assumption among the considered data. This may lead to biased integration results since considered data has inevitably related with the primary variable of interest. Advanced models such as tau, nu and lamda model impose weights to each calibrated probability that might allow considering data inter-relation or redundancy, and allow weighting the more reliable datum. The main challenge with probability combination schemes is fair consideration of redundant data which is a critical step.

In this paper, we propose a direct estimation technique to infer the joint probability rather than using probability combination approach which contains complicated redundancy weight calibration step. Besides, there is no guarantee that indirectly approximated probability satisfies basic probability requirements such as marginalization property. In the proposed method, the multivariate probability distribution is directly estimated, and they are updated constrained with the known marginal probabilities. We applied the direct estimation method to integrate soft secondary data and training images.

Probability Combination Approaches

We introduce a probability combination method briefly and discuss drawbacks. When combining soft secondary data and training image, previous studies have mainly focused on the use of probability combination approach. Suppose we have well data (termed D_1), training image (termed D_2), and seismic data (termed D_3). Each datum is calibrated to provide conditional probability of facies k=1,…,K, for example k = 1 is "being shale" and k = 2 is "being sand". Correct joint probability of interest should be constructed conditioned to all data sources D_1,D_2,D_3, such as $p(k| D_1,D_2,D_3,)$. Probability combination method, however, approximates joint probability through combining each calibrated probability in the followings:

$$\frac{p(k \mid D_1, D_2, D_3)}{p(k)} \approx \left[\frac{p(k \mid D_1, D_2)}{p(k)} \right]^{\lambda_1} \times \left[\frac{p(k \mid D_3)}{p(k)} \right]^{\lambda_2} \times C \qquad (1)$$

where C is a normalizing factor. Because MPS technique accounts for well hard data and training image data source D_1 and D_2 were aggregated to give $p(k|D_1,D_2)$. $p(k|D_3)$ is resulted from seismic data calibration with primary variable. Exponential terms λ_1 and λ_2 are interpreted as redundancy weights which should be varied based on how much redundant data are: higher redundant data get lower weight and lower redundant data get higher weight. Various weighting schemes have been developed to find optimum data redundancy measures. Data redundancy, however, is not simple function of linear relation between data but joint function of (D_1,D_2,D_3). Moreover, it is more difficult to quantify data redundancy between (well data + TI) and soft secondary data.

Major challenge of indirect estimation through combination function is not only to find optimum weights, but also to satisfy basic probability requirement such as marginal conditions. Approximated probability from eq. (1) must satisfy two known marginal conditions. Paper-123 in CCG report 10 introduces drawbacks of PCS approach in sense of difficulty of redundancy quantification and marginal conditions.

Integration of Multiple Secondary Data and TIs

CCG paper-101 in this report demonstrates the direct multivariate density estimation to integrate multiple secondary data. The method is to infer multivariate pdf under the known marginal pdf constraints. After constructing multivariate pdf the conditional probability of interest is directly extracted from the pdf. Multiple secondary data is assumed to be numeric so that the joint modeling of $(D_1,\ldots,D_m)$ is accessible, where $(D_1,\ldots,D_m)$ represent generically m soft secondary data. The multivariate pdf $f(k,D_1,\ldots,D_m)$ was initialized and then it can be modified under two marginality constraints in the below,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(k, D_1, ..., D_m) dD_1, ..., dD_m = p(k)$$

$$\sum_{k=1}^{K} f(k, D_1, ..., D_m) = f(D_1, ..., D_m)$$

since global information $p(k)$ can be computed using hard primary data, and the joint distribution of m secondary data $f(D_1,\ldots,D_m)$ is fairly accessible. Above marginal constraints were implemented with iterative updating scheme in direct multivariate density estimation.

Direct construction of multivariate pdf can be applicable to integrate secondary data and training image as well because: (1) we can infer MPS conditioned to training images and (2) we have already estimated multivariate pdf conditioned to secondary data. Figure-1 shows the construction of multivariate pdf using a specific hard conditioning data configuration and training image. Multivariate probability can be expressed such as in case of binary facies with two conditioning hard data,

$$p(\mathrm{u}_1 = k, \mathrm{u}_2 = k', \mathrm{u}_3 = k''), \quad k, k', k'' = 1, ..., K \qquad (2)$$

where u_1 is the simulation node, u_2 and u_3 are conditioning hard data location. Multivariate space is characterized by $K^{(N+1)}$ probability points, where N is number of conditioning data (see the right 3D figure in Figure-1). *snesim* is a noble algorithm to construct multivariate probability. Our challenge is to integrate soft secondary data and training image, ultimately to build multivariate pdf conditioned to MPS and secondary data such that,

$$p(\mathrm{u}_1 = k, \mathrm{u}_2 = k', \mathrm{u}_3 = k'', Y = y), \quad k, k', k'' = 1, ..., K \qquad (3)$$

Soft secondary data exists everywhere within reservoir area and collocated data is only used for inferring the multivariate probability eq. (3) at the visited node u_1. Figure-2 illustrates multivariate space jointly conditioned to multipoint data and collocated secondary data. One dimensional axis is added to multivariate space shown in Figure-1 due to incorporating secondary variable. Joint modeling of primary and secondary variable, *f*(k,y), is established from the secondary data integration with direct multivariate density estimation (see CCG paper-101 in this volume). Now, marginalization of the joint probability (3) provides the following relations:

$$\sum_{k''}^{K}\sum_{k'}^{K} p(\mathrm{u}_1 = k, \mathrm{u}_2 = k', \mathrm{u}_3 = k'', Y = y) = f(k, Y = y) \tag{4}$$

$$\int_{-\infty}^{\infty} p(\mathrm{u}_1 = k, \mathrm{u}_2 = k', \mathrm{u}_3 = k'', Y = y)dy = p(\mathrm{u}_1 = k, \mathrm{u}_2 = k', \mathrm{u}_3 = k'') \tag{5}$$

The first relation (4) shows summing up of joint probability over possible outcomes of primary variable at conditioning data location u_2 and u_3 should amount to the joint probability *f*(k,Y=y). The second relation (5) gives the summing up of joint probability over possible outcomes of secondary variable should amount to the multipoint statistics. Joint probability eq. (3), thus, must be constructed under marginal relation (4) and (5). Two marginality constraints specified above are implemented by alternating iterative process:

Step 1. Initialize joint probability $p^{(0)}(\mathrm{u}_1 = k, \mathrm{u}_2 = k', \mathrm{u}_3 = k'', Y = y), \quad \forall k, k', k'', \forall y$

Step 2. Update using the first marginality constraint eq. (4)

$$p^{(1)}(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'', Y{=}y) \Leftarrow \frac{f(k, Y{=}y)}{\sum_{k''}^{K}\sum_{k'}^{K} p^{(0)}(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'', Y{=}y)} p^{(0)}(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'', Y{=}y)$$

Step 3. Update using the second marginality constraint eq. (5)

$$p^{(2)}(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'', Y{=}y) \Leftarrow \frac{p(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'')}{\int_{-\infty}^{\infty} p^{(1)}(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'', Y{=}y)dy} p^{(1)}(\mathrm{u}_1{=}k, \mathrm{u}_2{=}k', \mathrm{u}_3{=}k'', Y{=}y)$$

Step 4. Set $p^{(2)}$ as $p^{(0)}$ and go to step 2 until there is no change in multivariate pdf.

Initial guess $p^{(0)}$ in step 1 was initialized with independence assumption such as $p^{(0)}=p(\mathrm{u}_1=k,\mathrm{u}_2=k',\mathrm{u}_3=k'') \times f(k,Y=\mathrm{y})$. Marginal conditions are honored in step 2 and 3. The above procedure is applied at every visited simulation nodes to get multivariate pdf.

Figure-3 shows the true reference image containing two faces: sand and mudstone colored by black and white, respectively. 72 well data is randomly sampled from the true image, and it was used as hard conditioning data. Training image used for MPS is also shown most right in Figure-3. Single secondary data was simulated over the entire nodes and calibrated with direct density estimation to provide facies probability maps as shown in the bottom of Figure-3. 72 samples were used for calibrating secondary data. To focus on the influence of secondary data integration, we generated very high correlated secondary data with $\rho_{\mathrm{sec,sand}}$=0.82: a certain seismic attribute is quite useful to detect sand channel in deep water clastic reservoir. Secondary data derived facies probability shows clear recognition of sand and mud facies.

MPS simulation was performed with small number of conditioning hard (we limited maximum conditioning data as 3) data and given TI. *snesim* algorithm was used for MPS simulation and two facies realizations are shown in the top of Figure-4. MPS realizations do not show good reproduction of realistic heterogeneity compared with true image possibly due to small number of conditioning data and/or the selected TI. The proposed direct density method was applied to integrate MPS with calibrated secondary probability. Two facies realizations are shown in the bottom of Figure-4. Integration of highly correlated secondary data produces better reproduction of channel pattern.

Discussion and Future Work

The main goal of this study is to assimilate secondary data and training image. One way to integrate those data is for combining data derived probabilities in order to estimate the joint probability. Probability combination schemes, however, have a few disadvantages: they require a data redundancy weight estimation process, and there is no guarantee of satisfying marginal conditions. In this paper, we proposed an idea of directly estimating multivariate distribution under the known marginal conditions. The method was applied to combine secondary data and MPS. The target joint probability is directly achieved through the alternating update process. Facies realizations with integrating secondary and MPS represented better reproduction of real heterogeneity.

The approach presented here gives good results, however, there are some aspects of the technique that need to be further explored and documented. These are including: (1) analysis of computational cost, (2) convergence problem of the iterative procedure, and (3) honoring hard data.

References

Benediktsson, J. A. and Swain, P. H., 1992, Consensus Theoretic Classification Methods, *IEEE Transactions on Systems, Man, And Cybernetics*, Vol. 22, No. 4.

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

Polyakova, E. I. and Journel, A. G., 2007, The Nu Expression for Probabilistic Data Integration, *Mathematical Geology*, Vol. 39, No. 8.

Silverman, B. W., 1993, *Density Estimation for Statistical and Data Analysis*, Chapman and Hall.

Strebelle, S., 2002, Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics, *Mathematical Geology*, Vol. 34, No. 1.

Hong, S. and Deutsch, C. V., 2008, The Place of Probability Combination Schemes, *Centre for Computational Geostatistics Report 10*.

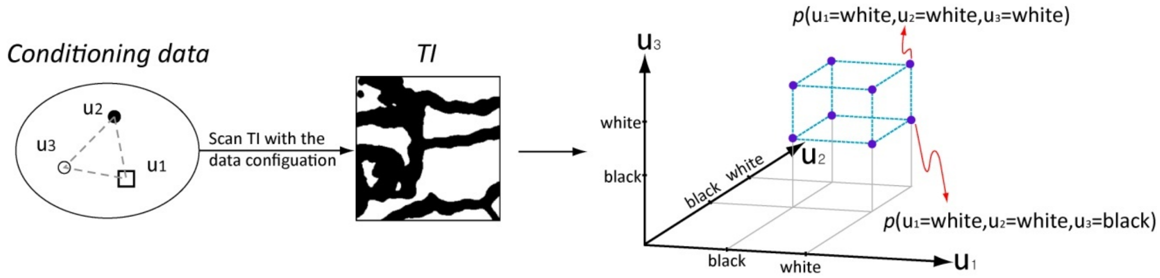

Figure 1: Schematic illustration of building multivariate pdf conditioned to training image.

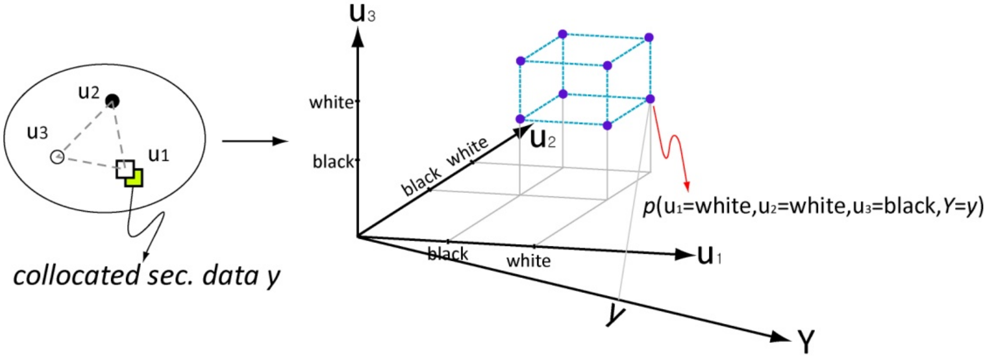

Figure 2: Schematic illustration of building multivariate pdf conditioned to MPS and secondary variable.

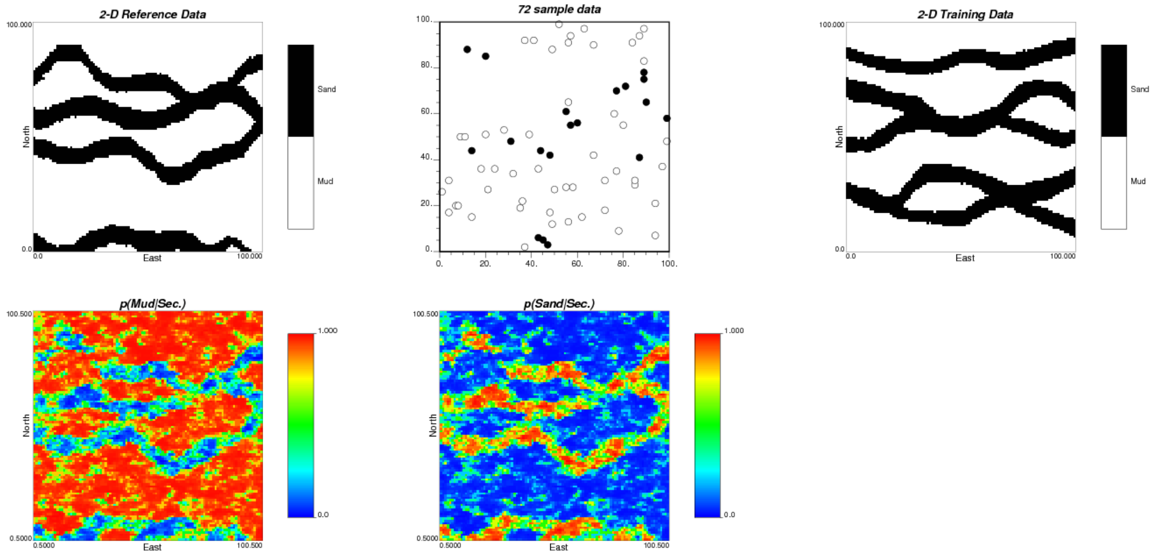

Figure 3: True image, 72 sample data extracted from true image and training image are shown in the top. Calibrated secondary data are shown in the below, p(facies|sec).

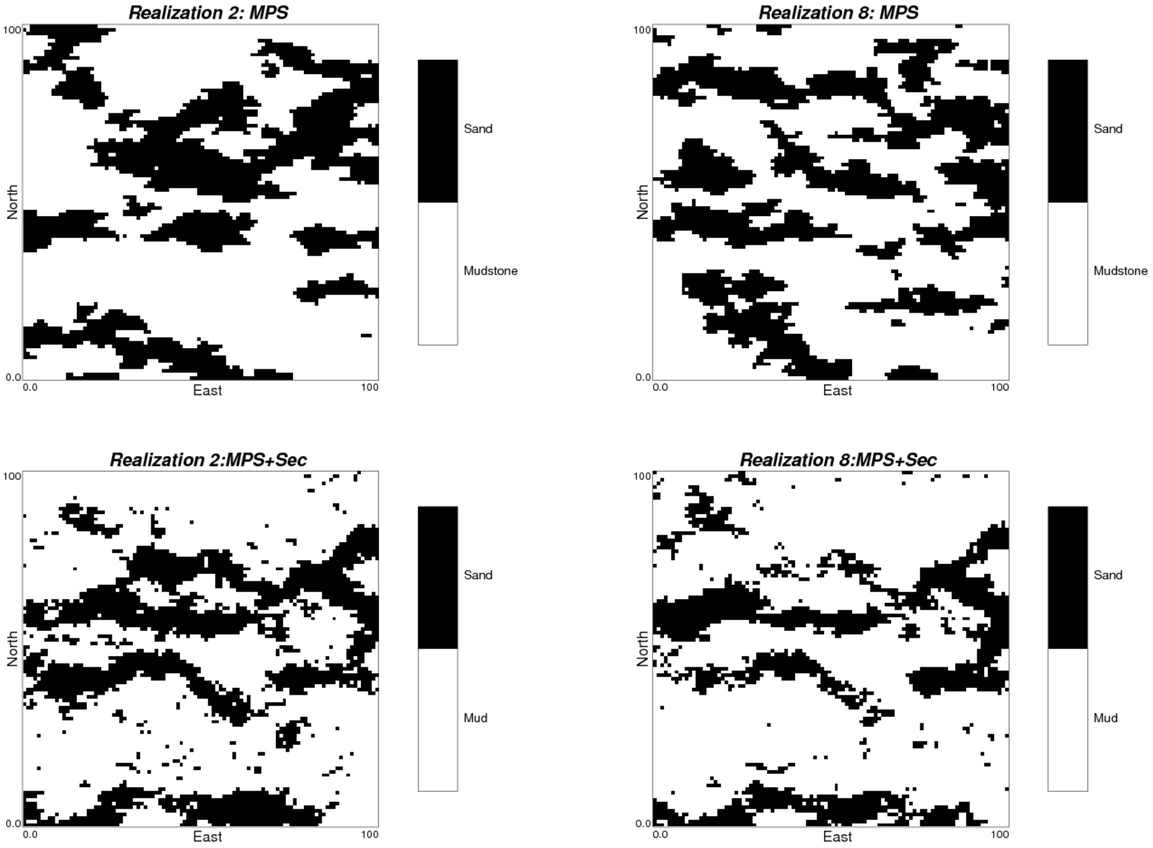

Figure 4: Two facies realizations obtained from MPS simulation are shown in the upper, and facies realizations with the integration of secondary data and MPS are shown bottom.