Paper 119, CCG Annual Report 12, 2010 (©2010)

Discrete Multivariate Probability Estimation Principles and Estimation Results Evaluation

Yupeng Li and Clayton V. Deutsch

Two questions are addressed in this paper. One is how to correctly estimate a desired probability distribution given the constraints in different order of moments? The second question is how to evaluate the estimation results from different techniques? The discrete multivariate probability is very common in geoscience research, such as the facies outcome for a group of spatial locations. While how to explicitly estimate such a multivariate probability distribution is still not well addressed. In this paper, the traditional maximum entropy principle is implemented in discrete multivariate probability estimation. The entropy concentration theorem (ECT)is revisited in multivariate probability estimation contents. It justifies that the multivariate probability estimation from the ME principle should be more appropriate than any other else estimation approach. Also, given the constraints, the final estimation results should provide as much information as possible. From information theory, the relative entropy is used as a criteria to measure the uncertainty deduction when comparing different posterior probability distributions updated from a prior with different techniques. From the entropy principle, it is the only reasonable candidate to be such a measure for discrete probability distribution choosing. The relative entropy should be as large as possible from the prior probability distribution.

1 Introduction

A discrete attribute such as facies in reservoir modelling is very common in geoscience research. The outcome for one location is looked as the realization of a discrete random variable, the joint outcomes for a group of spatial locations is usually modelled as the realization of a multivariate random function. The uncertainty of the outcomes at a group of locations is characterized by a multivariate probability (MP) distribution.

The definition of a discrete MP distribution is described in the content of spatial facies or rock type observation. It is an extension from the traditional univariate discrete probability distribution. Inferring the probability distribution from limited sampled locations is always a challenge because there is no shape parameters for discrete MP distribution. The principle of Maximum Entropy(ME) provides a powerful tool to estimate an unknown probability distribution from limited constraints[1, 2]. The correctness of this inference principle in discrete MP is justified with the Entropy Concentration Theorem[3].

Finally,the relative entropy is used as a criteria to measure the uncertainty deduction when comparing different posterior probability distributions updated from a prior with different techniques. The relative entropy will measure the information difference when in favour of one probability instead of anther probability distribution[4].

2 Discrete probability definition

Discrete or categorical variables are very common in geostatistics research. For example, the rock type for one spatial location could be only one from the set $\{limestone, dolomite, anhydrite\}$. When dealing with such observations, the discrete random variable is used in classical probability theory. The outcomes for each observation will be looked as a random variable.

Upper case letters like S or Z denote random variables. Lower case letters like s or z denote the value or outcome of a random variable. If S is a discrete random variable, then it is usually defined in words. For

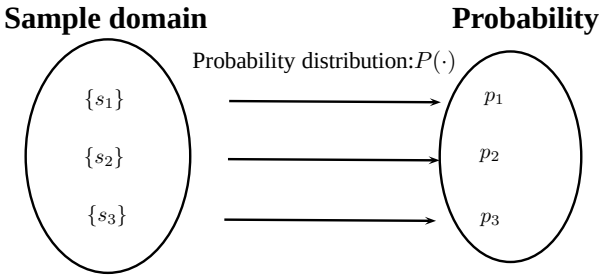

Figure 1: The univariate discrete probability distribution example for a sample domain

example, if it is defined as S : `the rock type of this location`, then its outcome s represents a specific rock type from set $\{limestone, dolomite, anhydrite\}$.

All the possible outcomes for one observation of a random process is called the domain of the random variable which is not necessarily a numerical set. As in this research, it usually is expressed in words and is denoted by a set $E : \{s_k, k = 1, 2, \cdots, K\}$. The main characters of this set are:

1. it is a set with limited items.
2. it could not be a numerical set.

In order to deal with these variables numerically, they are usually transformed into an binary variable as in the indicator approach. It could also be ordered and transformed into an integer set $\{k, k = 1, 2, \cdots, K\}$ directly. Basically, each random variable will also define a probability distribution as shown in Figure **??**. The function that gives the probability to the event that a random variable is exactly equal to some value is called a probability mass function(pmf). Suppose that S is a random variable defined on a domain E. Then the probability mass function $P(S) : R \rightarrow [0, 1]$ for S is defined as:

$$P(S) = Pr(S = s) = Pr(\{s_k \in E : S(s_k) = s\}) \tag{1}$$

It is a function that satisfy the properties: $P(S) \geq 0$ and $\sum P(S) = 1$.

Denote a spatial discrete variable as $S(\mathbf{u})$, the outcome of will be $s(\mathbf{u})$, where $\mathbf{u}$ could be any spatial coordinate. The set $E : \{s_1, \cdots, s_K\}$ are all the possible categories that can exist at any one of the locations in the domain. A location that is not perfectly known or needs to predicted is usually called "unknown", "unsampled" or "unclassified" and is denoted as $\mathbf{u}_0$. The outcomes at some locations are considered as sample locations and are referred to as $n-1$ sample locations denoted as $\{\mathbf{u}_1, \cdots, \mathbf{u}_{n-1}\}$. These n locations together could be in any spatial configuration.

A data event ω^n is a joint measurement of the outcomes at all the n locations, that is: $\omega^n : \{S(\mathbf{u}_1) = s_{k_1}, \cdots, S(\mathbf{u}_n) = s_{k_n}\}$. The data event space is defined as the set of all possible data events for the defined group of locations as: $\Omega^n : \{s(\mathbf{u}_1) \in E, \cdots, s(\mathbf{u}_n) \in E\}$. One example of all the possible data events composed by three locations and three outcomes for each location are plotted in Figure 2.

Each data event will have a probability p_ℓ to happen according some probability law that is defined as:

$$p_\ell = p(\omega_\ell^n) = Pr(\omega_\ell^n \in \Omega^n), \ell = 1, \cdots, N \tag{2}$$

where $N = K^n$ is the data event space dimension. The probabilities of all the possible data events will define a discrete multivariate probability mass function $P(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ which will characterize the probability of joint outcome for a group of locations $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$. It will satisfy $\sum P(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 1$ and $0 \leq P(\mathbf{u}_1, \ldots, \mathbf{u}_n) \leq 1$.

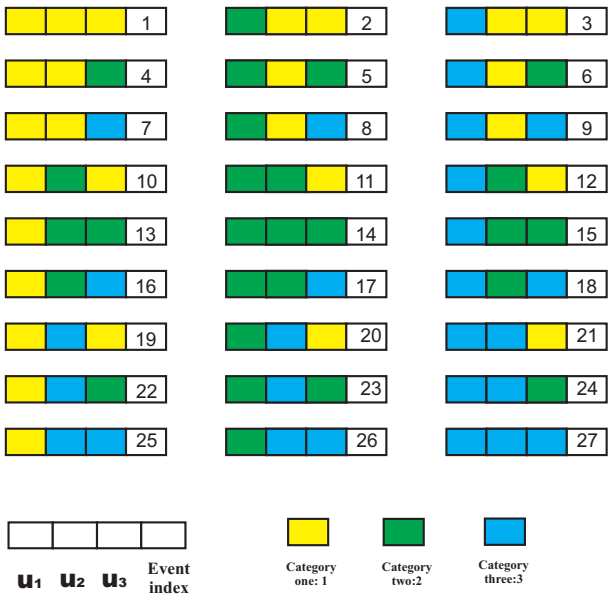

Figure 2: One example of all the possible data events composed by three locations and three possible categories

3 Probability estimation from information theory

After obtaining some samples from the stochastic process, the first step would be to estimate a probability distribution. The forecast uncertainty that we can make from the distribution depends on the estimated probability distribution. When the probability distribution shape is normal,the Fisher's variance-covariance based information measurement turns out to be a reasonable measure of uncertainty. When the probability is far from normal which is the case with discrete probability distributions, the information measurement defined by Shannon is an appropriate measure of uncertainty[5]. The entropy of a probability distribution $\mathbf{p} : \{p_\ell, \ell = 1, \cdots, N\}$ will be:

$$H(\mathbf{p}) = -\sum_{\ell=1}^{N} p_\ell \log p_\ell \tag{3}$$

Note: in the case of $p_\ell = 0$,it is defined that $H(p_\ell) = 0$.

The term entropy was first defined in thermodynamics as a measure of the change in randomness or disorder in a closed chemical system such as the result of a reaction[6, 7]. It was originally devised by Claude Shannon in 1948 to study the amount of information in a transmitted message and is expressed in terms of a discrete set of probabilities. Since then, Entropy has been widely accepted as a criteria to measure the uncertainty[8].

Although the concept of entropy is first proposed as an uncertainty evaluation for discrete probability distribution, it is also used in continuous ones[9]. It is defined as:

$$H(X) = -\int_a^b \log(f(x)) f(x) dx \tag{4}$$

Where random variable X with a range $(a, b)$and a pdf f. In the next part, it will be shown that the entropy has a good relation with the variance for a normal distribution. For a random variable X which has a normal

distribution $N(\mu, \sigma^2)$, the entropy will be:

$$\begin{aligned} H(X) &= -\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} exp(-\frac{1}{2}[\frac{x-\mu}{\sigma}]^2) \\ &\times \log\left\{\frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}[\frac{x-\mu}{\sigma}]^2)\right\} dx \\ &= \log(\sigma\sqrt{2\pi}) + \frac{\log e}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} y^2 dy \\ &= \log(\sigma\sqrt{2\pi}) + \frac{\log(e)}{\sqrt{\pi}} \frac{1}{2}\sqrt{\pi} \\ &= \log(\sigma\sqrt{2e\pi}) \end{aligned} \tag{5}$$

where the substitution $y = \frac{x-\mu}{\sqrt{2}\sigma}$ is used. It is interesting to observe that the entropy for a normal distribution is a function of σ but not of μ. This shows that the entropy and variance both measure uncertainty.

There could be all kinds of constraints to the unknown probability distribution. The most common one is the different order of moments to the desired probability distribution. Some other characteristic of the distributions that can be expressed as expected values. Both of them can be expressed as a linear operation to the desired probability distribution $\sum a_{m\ell} p_\ell = b_m; m = 1, 2, \cdots, M$. These quantities can be incorporated as constraints into the analysis.

An important goal is to construct a multivariate probability distribution from those easily obtained lower order statistics. That is inferring a multivariate probability $\{p_\ell, \ell = 1, \cdots, N\}$ which will satisfy some constraints. For the probability estimation under the constraints, the most extensively used technique is the Maximum Entropy (ME) principle. It is based on the premise that *when estimating the probability distribution, the best estimation results will keep the largest remaining uncertainty(the maximum entropy)consistent with all the known constraints. In that way, no more additional assumptions or biases were introduced in the estimation*[1, 2].

Generally, assuming any desired multivariate probability is $\mathbf{p} : \{p_\ell, \ell = 1, \cdots, N\}$, the objective function will be satisfied from some constraints:

$$\texttt{Maximum:} \qquad H(\mathbf{p}) = -\sum_{\ell=1}^{N} p_\ell Log(p_\ell) \tag{6}$$

subject to:

$$\sum p_\ell = 1 \tag{7}$$

$$\sum a_{m\ell} p_\ell = b_m; m = 1, 2, \cdots, M \tag{8}$$

The maximization of $H(\mathbf{p})$ subject to the above constraints results in a convex programming problem. The dual of this convex programming problem is actually unconstrained, so that there could always exist a solution to the above maximum question. The details of numerical solution can be found in another paper in this volume.

4 Entropy Concentration Theorem

As applying the ME principle to the MP estimation, the distributions that with higher entropy are favoured over others. Is that any solid theory underlying this modeling principle? Another natural question that may arise is that how far other possible distributions are from the maximum entropy in an entropy sense. That is

Granted that the distribution of maximum entropy has a favoured status, in exactly what sense, and how strongly, are distributions of lower entropy ruled out? Just what are we accomplishing when we maximize entropy? [3]

Janynes's Entropy Concentration Theorem(ECT) presents a possible justification of applicability of the ME principle to probabilistic forecasting. It is also called the uniqueness of entropy and a rigid proof is given in reference[10] (P.119). Informally speaking, it states that the most possible probability distribution $\{p_1...p_N\}$ are those that maximize the Shannon entropy given the probability of each events is the mathematical expectation of the frequency as the stochastic process repeated for T times as $T \to \infty$.

Consider the classic problem of a random experiment of the variable X, which can take the values x_ℓ,for $\ell = 1, \cdots, N$. After T trails, all the possible outcomes yielding a set of observed frequencies for each possible value of X given by

$$p_\ell = \frac{t_\ell}{T} \tag{9}$$

After this experiment is repeated many times, what set of frequencies $\{p_\ell\}$ can be realized in the greatest number of ways? A possible way to find this is by maximizing the multiplicity (subject to any set of linear constraints), which is given by

$$W(p_1, \cdots, p_\ell) = \frac{T!}{(Tp_1)! \cdots, (Tp_k)!} \tag{10}$$

Jaynes(1982) shows that when $T \to \infty, t_i \to \infty$ so that $p_\ell \to$ constant, the expression in (10) would obtain

$$T^{-1} \log W(p_1, \cdots, p_N) \to -\sum_{\ell=1}^{N} p_\ell \log p_\ell \tag{11}$$

which means that for large T, maximizing the entropy, the set of frequencies that can be realized in greatest number of ways are also found at the same time.

Furthermore, all runs of the experiments will yield distributions of frequencies with entropy in the range

$$H_{max} - \Delta H \leq H(p_1, \cdots, p_N) \leq H_{max} \tag{12}$$

where $2\Delta H$ is asymptotically χ^2 distribution with $\nu = N - M - 1$ degree of freedom and N is the number of probabilities, M is the number of linear constraints in the maximization problem and 1 results from the normalization constraint. Thus in terms of the upper tail area $(1 - F)$, ΔH is given by

$$2N\Delta H = \chi^2_\nu (1 - F) \tag{13}$$

Given incomplete information, the maximum entropy distribution is not only the one that can be realized in the greatest number of ways; in fact, for large T the overwhelming majority of all possible distribution compatible with the constraints have entropy very close to the maximum. As $T \to \infty$, the one of the maximum entropy becomes highly typical of those allowed by the constraints[3].

5 Evaluation of Estimation Results

From information theory, the information from the final probability should have minimum uncertainty for decision making. How can one measure the information difference when using one probability instead of another? How can one decide to accept one probability rather than other else? In this case, the relative entropy(KL divergence, KL distance, information discrimination)is one of the best measures [11, 4].

If there are two potential probability distributions, $\mathbf{p}$ and $\mathbf{q}$ for a random variable X, the relative entropy about X is defined as:

$$J(\mathbf{p}||\mathbf{q}) = \sum \mathbf{p} \log \frac{\mathbf{p}}{\mathbf{q}} \tag{14}$$

Conventionally, in the above definition, $0 \log \frac{0}{0} = 0$ and $0 \log \frac{0}{q} = 0$ also $p \log \frac{p}{0} = \infty$.

The relative entropy is always nonnegative and is zero if and only if $\mathbf{p} = \mathbf{q}$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Nonetheless, it is often usful to measure the "information distance" between distributions. Kullback and Leibler have shown that this measure $J(\mathbf{p}||\mathbf{q})$ can be used to develop a consistent statistical theory for measure expected amount of information given by a set of observations. In other words, the statistics $J(\mathbf{p}||\mathbf{q})$ gives the expected amount of information that an observation X yields in favour of the distribution $\mathbf{p}$ as opposed to the distribution $\mathbf{q}$.

Assuming that $\mathbf{u}_0$ is the location to be predicted. The outcome for location $\mathbf{u}_0$ will be characterized by its probability distribution $P(\mathbf{u}_0)$. Assuming we are in the case of complete lack of knowledge about how this location relates with others. In this case, one of the possible probability distributions is the uniform probability distribution:

$$p(\mathbf{u}_0; s_k) = \frac{1}{K}, \quad k = 1, \cdots, K \tag{15}$$

It is easy to show that this uniform probability distribution at this situation is the solution of maximum entropy approach.

Another situation is that only the global mean of each category $\{p_k, k = 1, \cdots, K\}$ are known. Without further spatial relationship, our best estimation for the event that $\mathbf{u}_0 = e_k$ can be written as:

$$p(\mathbf{u}_0; s_k) = p_k, \quad k = 1, \cdots, K \tag{16}$$

If the situation is that some sampled locations are obtained and it is known they are related to the unsampled location somehow. Generally, after obtaining the outcomes of the related surrounding neighbouring locations as $(\mathbf{u}_1 = s_{k_1}, \cdots, \mathbf{u}_n = s_{k_n})$ or simplify as (n), the prior probability distribution $P(\mathbf{u}_0)$ will be updated to a posterior probability distribution $p(\mathbf{u}_0|(n))$ which is a conditional probability distribution and will assign a probability value $p(\mathbf{u}_0; s_k|(n))$ to the event $\mathbf{u}_0 = s_k$ given the outcomes at the surrounding locations as:

$$p(\mathbf{u}_0; s_k|(n)) = Pr\{\mathbf{u}_0; s_k|(n)\} \tag{17}$$

There could be many methods to update the prior probability distribution as in (15) and (16) to posterior probability distribution model as in (17). The question we always need to answer is which posterior probability distribution should one take given that the estimated results all satisfy the known constraints. The relative entropy would give a clear answer to this question.

As already stated, the probability distribution with maximum distance from the prior distribution will bring minimum uncertainty. Thus, the larger relative entropy from the prior probability, the more information has obtained after updating to posterior probability distribution. For example, there are four possible categories existing in the domain. There are four estimated probability distributions $\mathbf{p^1}, \mathbf{p^2}, \mathbf{p^3}, \mathbf{p^4}$ estimated from different techniques which are listed together with the prior global proportion $\mathbf{p^0}$ in Table 1. From the relative entropy calculated from the global prior mean as shown in Table 1, it is clearly the estimated probability distribution $\mathbf{p^4}$ has the biggest relative entropy from the prior mean. Thus, the probability distribution $\mathbf{p^4}$ will give minimum uncertainty. It also makes sense intuitively as from probability distribution $\mathbf{p^4}$, only category 1 will exist with no uncertainty.

Probability Distribution	p_1	p_2	p_3	p_4	$J(\mathbf{p^i}\|\|\mathbf{p^0})$
$\mathbf{p^0}$	0.2	0.3	0.15	0.35	0.0000
$\mathbf{p^1}$	0.4	0.11	0.2	0.29	0.1699
$\mathbf{p^2}$	0.7	0.05	0.1	0.15	0.6197
$\mathbf{p^3}$	0.8	0.01	0.05	0.14	0.8918
$\mathbf{p^4}$	1	0	0	0	1.6094

Table 1: The relative entropy of four different probability distributions to the prior probability

6 Conclusion

Theoretically, the maximum entropy principle provide a unique probability distribution estimation approach given that the constraints do not fully define the distribution. One more example for this point is in spatial categorical variable modeling, usually different techniques are available to use. The relative entropy provide a valuable measurement to the estimated probability distribution evaluation. The larger the distance from the prior distribution, the more information gaining in favour of posterior instead of prior distribution.

References

[1] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(6):620–630, 1957.

[2] E. T. Jaynes. Information theory and statistical mechanics. ii. *Physical Review*, 108(2):171–190, 1957.

[3] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939, 1982.

[4] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[5] D. V. Gokhale. Approximating disrete distributions, with applications. *Journal of the American Statistical Association*, 68(344):1009–1012, 1973.

[6] J.W. Gibbs. A method of geometrical representation of thermodynamic properties of substances by means of surfaces. *Thermodynamics*, 1, 1873.

[7] Myron Tribus. *Thermodynamics and thermostatics*. D. Van Nostrand, New York, 1961.

[8] C.E. Shannon. A mathematical theory of communication. *Bell System Technical journal*, 27:379–423, 1948.

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

[10] David Applebaum. *Probability and Information: an integrated approach*. Cambridge University Press, 1996.

[11] H. Ku and S. Kullback. Approximating discrete probability distributions. *Information Theory, IEEE Transactions on*, 15(4):444–447, Jul 1969.