# Mutual Information and Its Application In Spatial Statistics

Yupeng Li and Clayton V. Deutsch

*Mutual information is a bivariate statistics that is widely used in information theory, but little used in geostatistics. This statistic can be used to summarize the spatial correlation from the bivariate probability between two locations. It can also be used as a criteria to select the most informative conditioning data when a limited number of conditioning data can be used.*

## 1 Introduction

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[1] and it is used in many research areas such as knowledge discovery and data mining[2]. The mutual information of two discrete random variables $S_i$ and $S_j$ is defined as:

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

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 $s_i$ and $s_j$ respectively. The basis of the log function will depend on the subject area. If the log function is based of 2, the measurement of the mutual information would be a bit. Mutual information measures how much of our uncertainty will be reduced after knowing one of the two variables.

**Calculation example**

A small example is used to illustrate its calculation. Table 1 is a simple bivariate probability. The univariate marginal probability of the table 1 is listed in table 2. The pointwise mutual information calculated(assuming base $e$) from this simple bivariate probability is listed in table 3. As the results show, although point wise mutual information could be negative, the final mutual information will be positive as it is the expectation.

| $s_1$ | $s_2$ | $P(s_1, s_2)$ |
|---|---|---|
| 0 | 0 | 0.1 |
| 0 | 1 | 0.7 |
| 1 | 0 | 0.15 |
| 1 | 1 | 0.05 |

Table 1: One bivariate probability for mutual information calculation example

## 2 Properties of Mutual Information

Mutual information quantifies the dependence between the joint distribution of $S_i$ and $S_j$ and what the joint distribution would be if $S_i$ and $S_j$ were independent. For example, if the outcomes at $s_i$ and $s_j$ are

| | $P(s_1)$ | $P(s_2)$ |
|---|---|---|
| 0 | 0.8 | 0.25 |
| 1 | 0.2 | 0.75 |

Table 2: The univariate marginal probability from Table 1

| | | $i(s_1, s_2)$ |
|---|---|---|
| 0 | 0 | $0.1 \times \log \frac{0.1}{0.8 \times 0.25} = -0.0693$ |
| 0 | 1 | $0.7 \times \log \frac{0.7}{0.8 \times 0.75} = 0.1079$ |
| 1 | 0 | $0.15 \times \log \frac{0.15}{0.2 \times 0.25} = 0.1648$ |
| 1 | 1 | $0.05 \times \log \frac{0.05}{0.2 \times 0.75} = -0.0549$ |
| | | $\sum = 0.1458$ |

Table 3: Mutual information calculation example

independent, then knowing the outcomes at $s_i$ does not give any information about the outcomes at location $s_j$. So their mutual information is zero. At the other extreme, if $s_i$ and $s_j$ are highly correlated, knowing the outcomes at one location will always inform the outcomes at the other location and that is their correlation would be 1. In this case, these two locations is actually bring no more information for the random variable. Thus, the information or uncertainty will be equal to any one of them.

The mutual information is always positive. This point can be proofed from the entropy function and conditional entropy function of random variable $S_i, S_j$ as given by T. M. Cover [1] . Given two random variable $S_i, S_j$, the mutual information would be:

$$\begin{aligned} R(S_i; S_j) &= \sum_{s_i \in I} \sum_{s_j \in J} P(s_i, s_j) \log \left( \frac{P(s_i, s_j)}{P(s_i)P(s_j)} \right) \\ &= \sum_{s_i \in I} \sum_{s_j \in J} P(s_i, s_j) \log \frac{P(s_i|s_j)}{P(s_i)} \\ &= -\sum_{s_i \in I} \sum_{s_j \in J} P(s_i, s_j) \log P(s_i) + \sum_{s_i \in I} \sum_{s_j \in J} P(s_i, s_j) \log P(s_i|s_j) \\ &= -\sum_{s_i \in I} P(s_i) \log P(s_i) - \Bigg( -\sum_{s_i \in I} \sum_{s_j \in J} P(s_i, s_j) \log P(s_i|s_j) \Bigg) \\ &= H(S_i) - H(S_i|S_j) \end{aligned} \tag{2}$$

In above equation (2), the $H(S_i|S_j)$ is the conditional entropy which is the expected value of the entropies of the conditional distribution averaged over the conditional random variables and can be defined as:

$$\begin{aligned} H(S_i|S_j) &= \sum_{s_j \in J} P(s_j) H(S_i|S_j = s_j) \\ &= \sum_{s_j} P(s_j) (\sum_{s_i} P(s_i|s_j) \log p(s_i|s_j) \\ &= \sum_{s_i \in I} \sum_{s_j \in J} P(s_i, s_j) \log P(s_i|s_j) \end{aligned} \tag{3}$$

The relation $H(S_i|S_j) \leq H(S_i)$ will always be satisfied Thus, the mutual information would always be positive. The mutual information $R(s_i; s_j)$ will equal to 0 if and only if $s_i$ and $s_j$ are independent random variables. In independent case, the joint probability distribution will be $P(s_i, s_j) = P(s_i)P(s_i)$ and therefore:

$$\log \left( \frac{P(s_i, s_j)}{P(s_i)P(s_j)} \right) = \log \left( \frac{P(s_i)P(s_j)}{P(s_i)P(s_j)} \right) = \log 1 = 0$$

Usually the entropy $H(S)$ is regarded as a measure of uncertainty about the outcomes at location $s$ which could be a discrete random variable. Then $H(S_i|S_j)$ is the amount of uncertainty remaining about

$S_i$ after $S_j$ is known. Thus $H(S_i) - H(S_i|S_j)$ will be the amount of uncertainty in $S_i$ which is removed by knowing the outcomes at location $S_j$. This corroborates the intuitive meaning of mutual information as the amount of information (that is, reduction in uncertainty) that knowing either variable provides about the other.

The lower limit of mutual information is seen to be zero. The upper limit will be obtained when these two random variable $S_i, S_j$ are perfectly correlated with each other. The upper limit would be the entropy calculated from any one of them. This point can be illustrated from the a bivariate normal distribution.

**Entropy of a normal distribution**
Define $S$ is a normal distribution, that is $S \sim N(\mu, \sigma^2)$. Its entropy will be calculated as:

$$\begin{aligned} H(S) &= -\frac{1}{\sigma(\sqrt{2\pi})}\int_{-\infty}^{+\infty} exp(-\frac{1}{2}[\frac{z-\mu}{\sigma}]) \times \log\left\{\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{1}{2}[\frac{z-\mu}{\sigma}])\right\}dx \\ &= \log(\sigma\sqrt{2\pi}) + \frac{\log e}{\sqrt{\pi}}\int_{-\infty}^{+\infty} exp(-y^2)y^2dy \\ &= \log(\sigma\sqrt{2e\pi}) \end{aligned} \tag{4}$$

In above equation (4), the entropy about a normal distribution will only be a function of variance. So, the entropy which is a measurement of information and uncertainty, have a strong relationship with variance which is also a measure measurement of uncertainty.

**Entropy of conditional probability distribution**
Given two variables that are both normally distributed $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, and assume their correlation coefficient is $\rho$. Then the conditional probability distribution of $Y$ given $X = x$ would also be a normal distribution $Y_x \sim N(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x-\mu_1),\ \sigma_2^2(1-\rho^2))$. Then from the normal probability entropy equation as defined in equation (4), the entropy of a conditional probability would be;

$$H_X(Y) = \log(\sigma_2^2(1-\rho^2)\sqrt{2e\pi}) \tag{5}$$

As before, the mutual information between two random variable $X, Y$ can be calculated from the entropy as:

$$R(X,Y) = H(Y) - H_X(Y) \tag{6}$$

Put the normal distribution entropy obtained in equation (4) and equation (5) into the mutual information calculation, then:

$$\begin{aligned} R(X,Y) &= H(Y) - H_X(Y) \qquad (7) \\ &= \log(\sigma_2\sqrt{2e\pi}) - \log(\sigma_2^2(1-\rho^2)\sqrt{2e\pi}) \\ &= -\frac{1}{2}\log(1-\rho^2) \end{aligned}$$

As shown in Equation (7), the mutual information can be interpreted from the correlation coefficient $\rho$. The stronger the correlation, the greater is the mutual information between $X$ and $Y$. So when $|\rho| = 1$, then the mutual information would close to infinity. But from the entropy view, for two totally related variables, knowing the distribution for variable $S_i$, the probability distribution of another variable $S_j$ will also be known. So, actually, only one variable exists. Now, the mutual information should be the uncertainty or the information measured by the entropy of itself $R(S_i, S_j) = H(S_i) = -\sum p(s_i)\log p(s_i)$. It is also one of the basic principles in information theory that a variable contains at least as much information about itself as any other variable can provide[1] .

# 3 Some examples from training images

Under the same second-order stationary assumption as that for the variogram calculation in geostatistics [3], the calculated mutual information function for any two locations at lag distance $\mathbf{h}$ as

$$R(\mathbf{h}) = \sum\sum P(\mathbf{u}, \mathbf{u}+\mathbf{h}) \log \left( \frac{P(\mathbf{u}, \mathbf{u}+\mathbf{h})}{P(\mathbf{u})P(\mathbf{u}+\mathbf{h})} \right) \tag{8}$$

Where $P(\mathbf{u}, \mathbf{u}+\mathbf{h})$ would be the bivariate probability statistics for a random variable $S(\mathbf{u})$ and it would be: $P(\mathbf{u}, \mathbf{u}+\mathbf{h}) = Pr(S(\mathbf{u}) = s(\mathbf{u}), S(\mathbf{u}+\mathbf{h}) = s(\mathbf{u}+\mathbf{h}))$. While $P(\mathbf{u})$ and $P(\mathbf{u}+\mathbf{h})$ are the univariate probability at this two locations.

Two training images are used to calculate the mutual information along different directions. The first example of the mutual information calculated from one training image is shown in Figure 1. The lag distance that mutual information reaches the first minimum value shows the continuity of the facies in the whole map. After that low *valley* period, the mutual information increasez again.

In this example, the yellowstone data set is used as shown in Figure 2. The calculated mutual information is shown in Figure 2. As the distance increased, the mutual information value reaches the minimum value without increasing again.There is no such cyclity as found in Figure 2 as in this second training image is more stochastic in the long range. While for training image one, in the long range, there is good cross contact pattern which is reflected from the mutual information diagram.

As it can reflect the spatial dependent of two locations, it could be used as a kind of spatial statistic. The experimental variograms and the respective mutual information diagram calculated from the training image one are shown in Figure 3. As shown in the comparison, in the short lag distance, before the indicator variogram reaching their sill and mutual information reaching the first valley, these two statistics are almost the same. While in the long range, the mutual information function reflect more information than indicator variogram. In Figure 3, the mutual information is calculated from the indicator transformed data for comparison purposes.

# 4 Conclusion and discussion

The mutual information has integrated the direct and cross indicator variogram information into one statistics. It can be used as a spatial statistics to choosing the conditioning data. In the new proposed Direct Multivariate Probability Estimation(DMPE) approach for facies modeling, the bivariate probability matrix

$$P(\mathbf{h}) = p(\mathbf{u} = s_k, \mathbf{u}+\mathbf{h} = s_k'), s_k, s_k' = 1, \cdots, K$$

is used instead of using indicator covariance. At each step, the mutual information between any two locations can be easily calculated from this bivariate statistics. Thus, from the mutual information view point, the DMPE would integrate all the information form the surrounding conditioning data in a non linear style.

Also, when the number of conditioning data are constrained by the probability space, finding the most informative conditioning data and integrating them together will be very important. It could be used a criteria to choose conditioning data when only limited conditioning data can be used.

# References

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

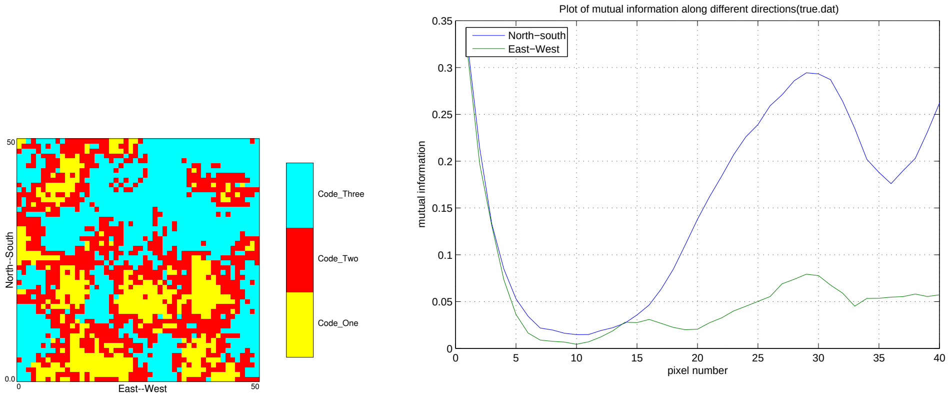

Figure 1: The mutual information(Right) calculated along two directions from the training image(Left)

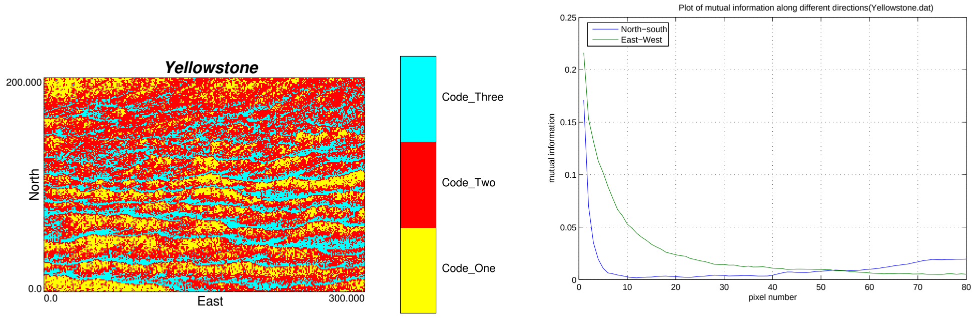

Figure 2: Training image two

[2] 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.

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

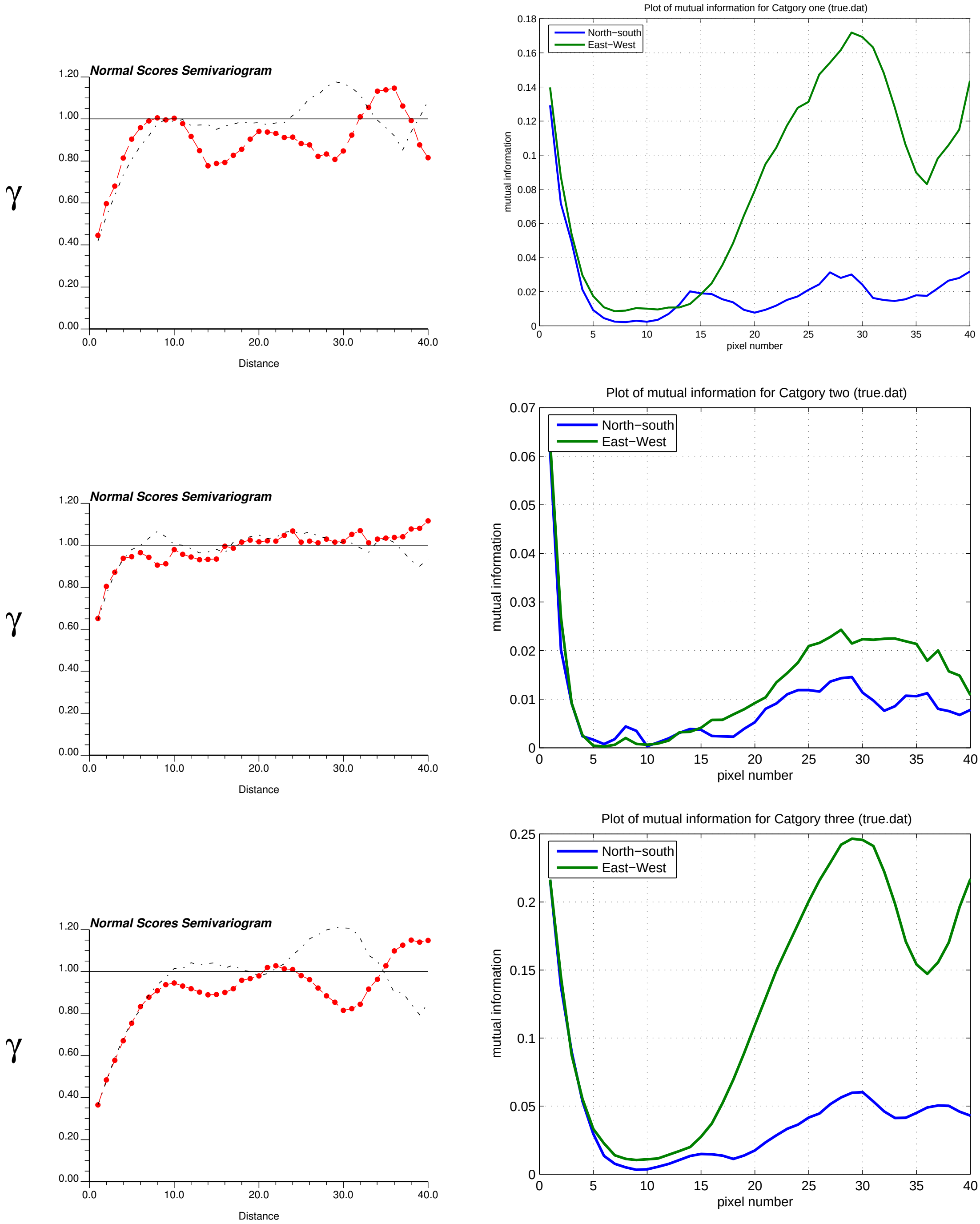

Figure 3: The comparison of experimental variogram and the mutual information