Application of the Sampling Distribution of the Correlation Coefficient

E.B. Niven and C. V. Deutsch

The correlation coefficient is uncertain when the sample size is small. A correlation coefficient calculated from a small data set may not be trustworthy. This research examines the uncertainty in the correlation coefficient based on the measured correlation and the number of data.

Introduction

Although obtaining data from wells is expensive, especially in offshore environments, seismic geophysical data can be acquired at a fraction of the cost and its vast areal coverage provides details on the reservoir that are unreachable by wells. As a result, seismic attributes are widely used as predictors of reservoir properties. When a physically justifiable relationship between a seismic attribute and a reservoir property is demonstrated, the uncertainty of interwell predictions of reservoir properties can be significantly reduced leading to better geostatistical models. With well data being so expensive to obtain, the integration of multivariate data in the presence of sparse data is of particular importance. The redundancy and the relationships between the secondary data and the variable being estimated are frequently quantified by the correlation coefficient. Often, correlations are estimated from a small number of observations. When the sample size is small, the uncertainty about the value of the true correlation can be very large, particularly when the estimated correlation is low (Kalkomey, 1997).

In order to assess the uncertainty in geostatistical models, which rely on the correlation between well and seismic data, this research examines the uncertainty in the correlation coefficient and reviews the sampling distribution of the correlation coefficient as a function of the available data. Two programs were developed for this research. The first is called NIND, which calculates the number of independent data in a dataset. The second is called SAMP_DIST_CORR, which calculates the distribution of the correlation coefficient based on the measured correlation and the number of data.Pearson's Correlation Coefficient

Let $(x_1, y_1),...,(x_n, y_n)$ be n observations from a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, where μ_x and σ_x^2 are the mean and variance of *x*, μ_y and σ_y^2 are the mean and variance of *y*, and ρ is the correlation coefficient between *x* and *y* given by $\rho = \beta\sigma_x / \sigma_y$ where β is the slope parameter of regression of *y* on *x*. The sample correlation coefficient commonly used for estimating ρ is the Pearson's product-moment correlation coefficient defined by:

$$r_p = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2\right]^{1/2}} \tag{1}$$

The Distribution of r

The distribution of *r* as given in Johnson, Kotz and Balakrishnan (1995) is as follows:

$$p_R(r) = \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\sqrt{\pi}\,\Gamma(\frac{1}{2}(n-1))\Gamma(\frac{1}{2}n-1)} \times \sum_{j=0}^{\infty} \frac{[\Gamma(\frac{1}{2}(n-1+j))]^2}{j!}(2\rho r)^j \tag{2}$$

Where $-1 \le r \le 1$.

Note that (2) also assumes that (X_i,Y_i) and (X_j,Y_j) are mutually independent if $i \neq j$. In formula (2), ρ is the measured correlation and n is the number of independent data points.

Calculating the Number of Independent Data

Say we have a number of observations X_i, where i=1,…,n.

Then:

$$Var\{X_i\} = Cov\{X_i, X_i\} = C_{ii} = \sigma^2_{data} \tag{3}$$

And we know that:

$$Var\{\overline{x}\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij} \tag{4}$$

But, we also know that:

$$Var\{\overline{x}\} = \frac{\sigma^2_{data}}{N_{Independent}} \tag{5}$$

Where $N_{Independent}$ is the number of independent data. And, from the variogram:

$$C_{ij} = \rho_{ij}\sigma^2_{data} \tag{6}$$

Therefore:

$$N_{Independent} = \frac{\sigma^2_{data}}{Var\{\overline{x}\}} = \frac{\sigma^2_{data}}{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{ij}\sigma^2_{data}} \tag{7}$$

Finally, we have:

$$N_{Independent} = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{ij}} \tag{8}$$

A program called NIND was developed which automatically calculates the number of independent data based on (8) and the variogram model.

Calculating the Sampling Distribution for the Correlation Coefficient

A program called, SAMP_DIST_CORR, was created to calculate the sampling distribution for the correlation coefficient. The program uses the formula in (2) with the measured correlation and number of independent data as inputs.

Practical Considerations

Although the summation in (2) is to infinity, it tends to converge rapidly except where the measured correlation is quite high (i.e. $\rho > 0.9$). Thus, the upper summation bound and a tolerance parameter are specified inputs into SAMP_DIST_CORR. The program calculates the percentage of instances where the summation parameter does not converge to a value smaller than the specified tolerance parameter. If the percentage of values not converging is too high, the summation parameter can be increased (or the tolerance can be increased).

Discussion and Application

One example of application of this research is in collocated co-simulation. In collocated co-simulation a Markov-type assumption is made where collocated secondary information is assumed to screen further away data of the same type. The collocated co-simulation relies on the measured correlation between the primary and secondary data. In some cases such as in off shore oil and gas reservoirs, there may be few wells or samples upon which a correlation may be calculated. In these cases, the uncertainty in the correlation coefficient may be quite large.

Figure 1 shows the distributions for the correlation coefficient for a measured correlation of 0 as a function of sample size. The figure shows that as *n* increases from 4 to 100, the distribution becomes more and more narrow. However, as is shown in the figure, even with 100 samples indicating a correlation of 0.0, there is still significant uncertainty in the true correlation. In fact, Figure 1 shows that there is 95% probability of the true correlation being between -0.2 and 0.2 (a range of 0.4).

Figure 2 is similar to Figure 1 except that the measured correlation coefficient is 0.5. The sampling distribution for the ρ=0.5 case is slightly more narrow than the ρ=0 distribution. However, as is shown in Figure 2, there is still significant uncertainty in the correlation even where the sample size is 100. Figure 2 shows that there is 95% probability of the true correlation being between 0.35 and 0.65 (a range of 0.35).

One could use the program ROBUSTCORRCO (as documented elsewhere in this report) to obtain a more robust correlation coefficient. Then the program NIND could be used to calculate the number of independent data. Finally the program SAMP_DIST_CORR could be used to obtain the sampling distribution for the correlation coefficient. Then collocated co-simulation could be completed with the P_{10}, P_{50} and P_{90} correlations to observe the impact of the uncertainty in the correlation on the measured reserves. Or, a Monte Carlo simulation approach could be used to randomly sample the distribution of *r* as an input into the collocated co-simulation.

Conclusions

The relationship between primary and secondary data in geostatistics is frequently quantified using the correlation coefficient. In cases where there are a small number of samples the uncertainty in the correlation coefficient can be quite large. This research presents a way to calculate the number of independent data points. The sampling distribution for the correlation coefficient can then be calculated based upon the measured correlation and the number of independent data.

References

Johnson, Richard A., and Dean W. Wichern. 2007. *Applied multivariate statistical analysis*. 6th ed. Upper Saddle River, NJ: Pearson prentice Hall.

Kalkomey, C. T. 1997. Potential risks when using seismic attributes as predictors of reservoir properties. *The Leading Edge* (March 1997): 247-51.

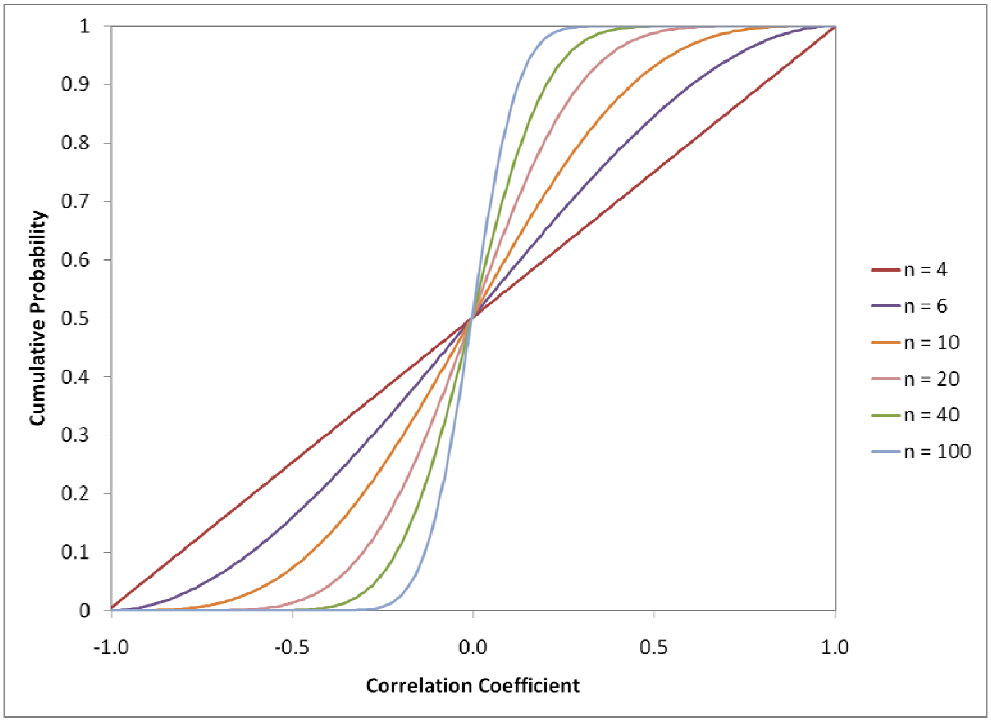

Figure 1: The distribution of the correlation coefficient for ρ=0 and for increasing n.

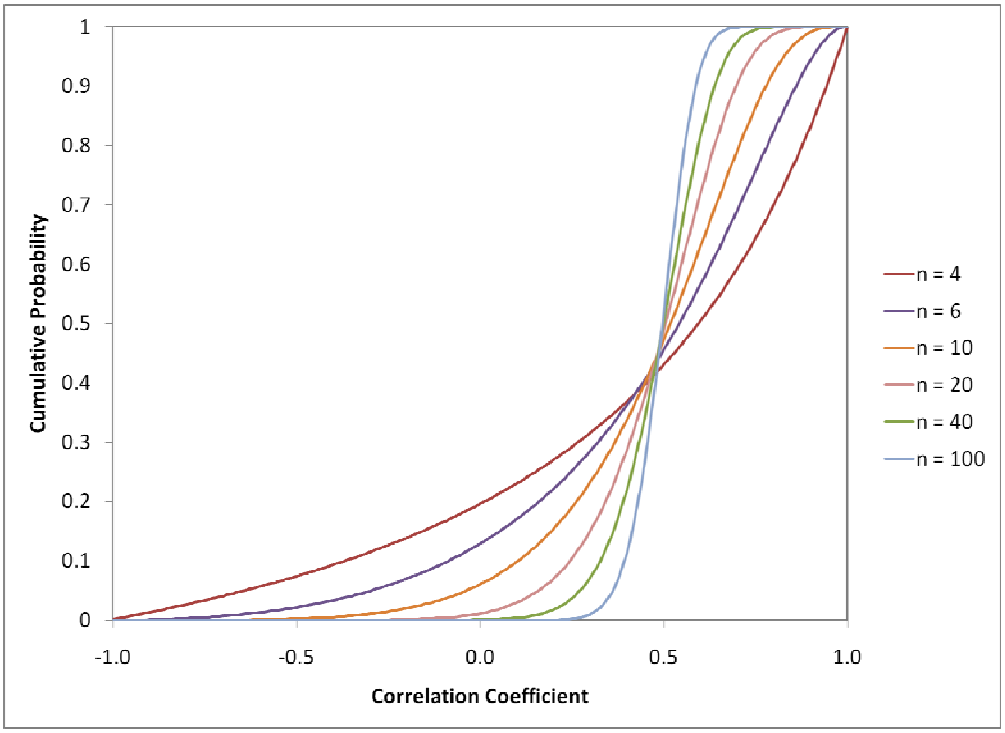

Figure 2: The distribution of the correlation coefficient for ρ=0.5 and for increasing n.