A Robust Method for Calculating the Correlation Coefficient

E.B. Niven and C. V. Deutsch

Relationships between primary and secondary data are frequently quantified using the correlation coefficient; however, the traditional Pearson correlation coefficient is known to be heavily influenced by outlier data points. This paper presents a method for calculating a more robust, updated correlation coefficient. The updated correlation coefficient is a weighted average of correlations calculated by leaving various combinations of data out. The proposed updated correlation coefficient is less sensitive to outlier data than either the Pearson or Spearman correlation coefficients.

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 the 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. The traditional Pearson correlation coefficient is known to be heavily influenced by outlier data (Isaaks and Srivastava, 1989; Abdullah, 1990; Shevlyakov, 1997). This research presents a method to calculate a correlation coefficient that is more robust and less sensitive to outliers.

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:

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

The sample means for x and y are known to be sensitive to outliers. As a result, the estimate in (1) is also sensitive to outliers in either x, y, or both variables (Abdullah, 1990 and Shevlyakov, 1997).

Spearman's Rank Correlation Coefficient

As an alternative to Pearson's correlation coefficient, the non-parametric Spearman's rank correlation coefficient, r_s, can be calculated as follows:

$$r_s = 1 - \frac{6D^2}{n(n^2 - 1)} \tag{2}$$

Where $D^2 = \sum_i^n (r_{y_i} - r_{x_i})^2$

And r_{x_i} and r_{y_i} are the ranks of x_i and y_i, respectively. Spearman's correlation coefficient does not require the assumption that there is a linear relationship between the variables and is generally more resistant to outliers than Pearson's coefficient. However, as is shown later in this paper, Spearman's rank correlation coefficient is still unacceptably sensitive to outliers, particularly in the presence of sparse data.

Correlation in the Presence of Outliers

Figure 1 shows a scatterplot of the bivariate relationship between two variables, x and y. The scatterplot shows what appears to be a strong correlation between the two variables marred by one potential outlier data point at a location of (7, 1). The Pearson and Spearman correlation coefficients for the data shown in Figure 1 are 0.291 and 0.214, respectively. If we were sure the point at (7, 1) was an outlier, we could simply remove it from the plot and the Pearson and Spearman correlations would increase to 0.910 and 0.943, respectively. However, if we are not sure that the point is an outlier we should probably not remove it. Note that, while the Spearman correlation coefficient is generally more resistant to the effects of outliers, in this case it is more strongly affected by the potential outlier data point.

A More Robust Correlation Coefficient

We could conduct a "leave one out test" (LOOT), whereby a data point is removed from the dataset and the correlation is recalculated. This procedure can be repeated *n* times for a dataset with *n* points, leaving a different data point out each time. The result is *n* calculated correlation coefficients. A LOOT was conducted for the data shown in Figure 1 and the results are shown in Table 1. For the first 6 leave one out tests, the resulting correlations are very low and unrepresentative of the obvious correlation in the data. However, the last test calculates a correlation of 0.910.

Table 1: Resulting correlations from a leave one out test (actual correlation is 0.23)

Coordinates of Left Out Point (x,y)	Resulting Correlation
(1.00, 1.98)	0.081
(2.00, 3.20)	0.235
(3.00, 3.53)	0.269
(4.00, 7.25)	0.318
(5.00, 5.44)	0.272
(6.00, 9.31)	0.002
(7.00, 1.00)	0.910

The more robust correlation coefficient is based upon the idea of weighted average of the correlations calculated in the LOOT. The idea is to weight the correlations according to their difference from the actual correlation as follows:

$$w_i = \left| \rho_{Actual} - \rho_{i,LOOT} \right|^{\alpha} \tag{3}$$

Where:

- w_i is the resulting weight assigned to each correlation calculated in the leave one out test
- ρ_{Actual} is the Pearson's correlation coefficient calculated from the original data
- $\rho_{i,LOOT}$ is the i^{th} Pearson's correlation coefficient calculated from leaving out the i^{th} data point in the leave one out test
- α is a weighting coefficient and is a function of the number of data ($\alpha = 1 + n/20$)

In the above example the correlation obtained from removing the point at (7, 1) is the most different from the actual correlation of 0.291 and thus receives the most weight.

Then, the more robust correlation coefficient calculated from the LOOT for sparse datasets is defined as:

$$r_{Robust,LOOT} = \frac{\sum_{i=1}^{n} w_i \rho_{i,LOOT}}{\sum_{i=1}^{n} w_i} \tag{4}$$

The updated correlation coefficient is essentially a weighted average correlation of the correlations calculated in the LOOT, where the weights are defined in (3). The weighting scheme in (3) assigns the greatest weights to the correlations that are the most different from the actual Pearson's correlation coefficient. The idea is that the data points that have the biggest impact on the correlation are the ones that are most likely to be outliers. For the data shown in Figure 1, the LOOT correlation weights and updated correlation coefficient is as follows in Figure 1. As is shown in the table, the updated correlation coefficient from the LOOT is calculated to be 0.569, seems reasonable given that we may not want to totally exclude the potential outlier.

Table 2: Resulting correlations from a leave one out test (actual correlation is 0.23)

Coordinates of Left Out Point (x,y)	Resulting Correlation	Weight (w_i)
(1.00, 1.98)	0.081	0.12
(2.00, 3.20)	0.235	0.02
(3.00, 3.53)	0.269	0.01
(4.00, 7.25)	0.318	0.08
(5.00, 5.44)	0.272	0.05
(6.00, 9.31)	0.002	0.19
(7.00, 1.00)	0.910	0.52
Updated Correlation from Leave One Out Test = 0.569		

Consider another example data set shown to the right of Figure 2. The figure shows a dataset with 9 data points. With the exception of two potential outlier data points near a (x, y) location of (12, 2), the figure appears to show a strong correlation between x and y. However, due to the influence of the two potential outlier points, the resulting Pearson's and Spearman's correlation coefficients are -0.050 and -0.133, respectively. If we were certain that the two points near (12, 2) were outliers, we could remove them from the dataset and the Pearson correlation coefficient would increase to 0.853. Note that the data shown to the right of Figure 2 represent another case where the rank correlation coefficient performs worse than the Pearson's correlation coefficient.

If we calculate an updated correlation based upon the LOOT, the updated correlation improves to 0.083. The reason for the meager improvement in the updated correlation calculated from the LOOT is that there are two outliers affecting the otherwise strong correlation in the data. So, when only one of those points is left out, the correlation is still adversely affected by the other remaining outlier.

Thus, we could perform a "leave *two* out test", where every combination of two data are left out with a correlation calculated each time. In the leave two out test, the resulting updated correlation is 0.176, which is a slight improvement over the LOOT updated correlation.

We could also perform a leave three out test and a leave four out test and so on, up to a leave n-3 out test. For the leave n-3 out test, correlations are calculated for each combination of three point subset of the original data.

If we performed each of these tests, leaving out a different amount of data each time, we would be left with n-3 updated correlations. We could then weight each of those correlations as follows:

$$w_{X,LXOT} = \left| \rho_{Actual} - \rho_{X,LXOT} \right|^{\alpha} \tag{5}$$

Where:

- ρ_{Actual} is the original data correlation
- $\rho_{X,LXOT}$ is the updated correlation calculated in each leave "x-data" out test
- $w_{X,LXOT}$ are the weights calculated for each updated correlation from the leave "x-data" out test
- α is the same weighting exponent as in (3) (i.e. $\alpha = 1 + n/20$)

Then the overall updated correlation is calculated as follows:

$$r_{OverallUpdated} = \frac{\sum_{X=1}^{\frac{3n}{4}-3} w_{X,LXOT} \rho_{X,LXOT}}{\sum_{X=1}^{\frac{3n}{4}-3} w_{X,LXOT}} \tag{6}$$

It was noted that as the number of data increases the updated correlations calculated using only small amounts of data (for example, say we have 20 data and we want to calculate a correlation by using 4 data points and leaving out 16) tend to adversely affect the overall updated correlation. So, rather than considering correlations for every possible subset of data, the overall updated correlation only considers correlations calculated by leaving out 1 through $\frac{3n}{4} - 3$ (where the ratio is always rounded down) data points. Although this is a rather arbitrary decision, it seems to have improved the results considerably.

When the overall updated correlation coefficient is calculated for the data in Figure 1, the algorithm calculates updated correlations leaving out 1 through 4 data points. The overall updated correlation is calculated to be 0.229. The updated correlations and the overall updated value is shown in Table 3.

Table 3: Updated correlations from the leave x-data out test and the overall updated correlation for the data in Figure 1 (right)

Leaving out the following number of data	Updated Correlation
1 Data Point	0.280
2 Data Points	0.251
3 Data Points	0.176
4 Data Points	0.083
Overall Updated Correlation Coefficient = 0.229	

Breakdown Properties of the Overall Updated Correlation

To illustrate the breakdown properties of the overall updated correlation coefficient in (6) compared with the Pearson and Spearman correlation coefficients, a simulation study was carried out as presented in Abdullah (1990).

First, 100 'good' observations are generated according to the linear relation $y_i = 2 + x_i + u_i$ where x_i is drawn randomly from a normal distribution with a mean of 5.0 and a variance of 1.0. u_i is drawn from a normal distribution with a mean of 0 and a standard deviation of 0.2. The results were as follows: $r_{Pearson}$ =0.984, $r_{Spearman}$=0.976 and $r_{OverallUpdated}$=0.941.

Next, the data was slowly contaminated. In increments of 10 data points, the 'good' data was replaced with 'bad' data points. The conataminated data points were generated according to the linear relation where x_i is uniformly distributed on [5, 10] and y_i is drawn from a normal distribution with a mean of 2 and a standard deviation of 0.2.

This was repeated until only 50 'good' observations remained. Table 4 shows the values of Pearson's, Spearman's and the overall updated correlation when 'good' observations are replaced by increasing fractions of outliers. A breakdown plot that illustrates the value fo the correlation coefficients as a function of the percentage of outliers is shown in Figure 2.

Table 4: Updated correlations from the leave x-data out test and the overall updated correlation for the data in Error! Reference source not found.

Contamination (%)	Pearson ρ	Spearman ρ	Overall Updated ρ
0	0.984	0.976	0.941
10	-0.070	0.503	0.963
20	-0.317	0.195	0.92
30	-0.451	-0.091	0.947
40	-0.603	-0.380	0.391
50	-0.601	-0.489	0.266

Discussion

A Fortran 90 program called ROBUSTCORRCO was created which automatically calculates the updated correlation coefficients for each LXOT as well as an overall updated correlation. In cases where there are more than approximately 20 data points, the time to calculate the number of combinations of data in the LXOT becomes prohibitively large. As a result, a monte-carlo-like simulation approach was implemented in which 1000 data combinations for each LXOT are randomly sampled rather than calculating every possible data combination.

One example of an application of the more robust updated correlation coefficient 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 correlation coefficient is used to specify the relationship between primary and secondary data.

However, 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 this case, the overall updated correlation coefficient could be calculated and compared to the traditional Pearson correlation.

However, even using the updated correlation coefficient, there is still significant uncertainty in the true underlying correlation. The overall updated correlation and the number of data could be used to define a distribution of uncertainty in the correlation coefficient as outlined in Niven and Deutsch (2008).

Conclusions

The relationship between primary and secondary data in geostatistics is frequently estimated using the correlation coefficient. However, the traditional Pearson and Spearman correlation coefficients are very sensitive to outlier data points. A method for calculating an updated correlation coefficient is presented in this paper. The updated correlation coefficient is shown to be more robust than either Pearson's or

Spearman's correlation coefficient. Although the method is somewhat ad-hoc, it appears to work well for a large range of examples.

References

Abdullah, M. B. 1990. On a robust correlation coefficient. *The Statistician* 39, : 455-60.

Isaaks, E. H., and R. M. Srivastava. 1989. *An introduction to applied geostatistics*. New York: Oxford University Press, Inc.

Niven, E. B., and C. V. Deutsch. 2008. Application of the sampling distribution for the correlation coefficient. Centre for Computational Geostatistics: Report 10.

Shevlyakov, G. L. 1997. On robust estimation of a correlation coefficient. *Journal of Mathematical Sciences* 83, (3): 434-8.

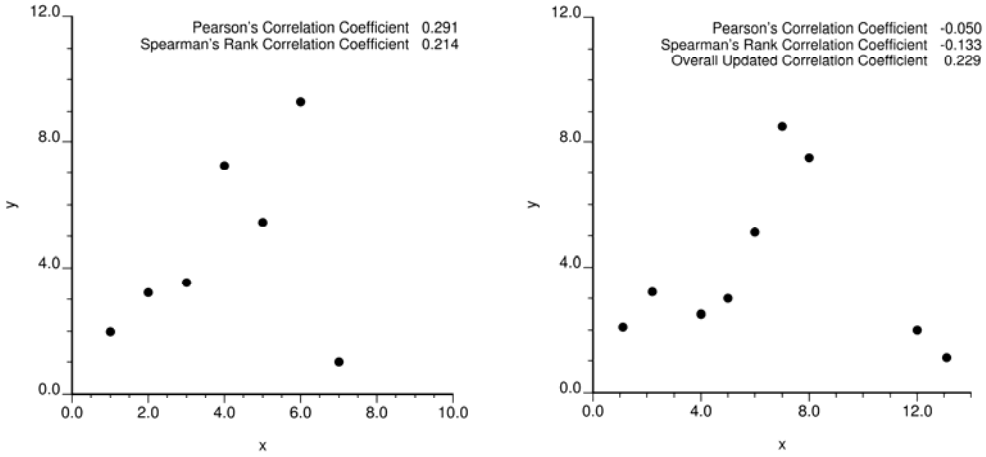

Figure 1: An example dataset with an otherwise high correlation between x and y marred by one apparent outlier data point and a second example dataset with two potential outliers.

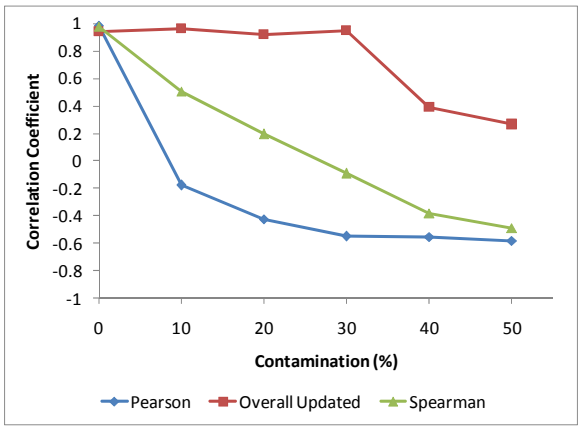

Figure 2: Simulation study comparing effect of contaminated data on the Pearson, Spearman and proposed overall updated correlation coefficient.