A Short Note on Fixing the Correlation Matrix

Chad Neufeld and Clayton V. Deutsch

Centre for Computational Geostatistics (CCG)
University of Alberta

Bayesian updating and other techniques require matrices of correlation coefficients between multiple variables. Often, there are different numbers of paired data available for each entry. This almost always leads to a correlation matrix that is not positive definite. Sometimes, the correlation matrix is positive definite, but it leads to unstable weights applied to highly redundant data. This paper proposes an iterative scheme for correcting the correlation matrix. The changes are constrained by the number of data used to calculate the correlation and how sensitive the matrix is to changing a specific correlation. The result is a stable matrix that can be used for combining the secondary data.

Introduction

Many geostatisticians are faced with the problem of how to use multiple secondary variables for building their models. Most algorithms are coded to use one secondary variable. A method was developed here at the CCG for combining multiple secondary variables into one "super-secondary" variable [1].

The merging is done under a multivariate Gaussian framework. All of the variables are normal score transformed and then correlation matrix is calculated on the normal score values. The correlation matrix, between the primary and secondary variables, is used as the basis for the merging process. The advantage of this method is that the secondary variables are merged using a full model of redundancy. This redundancy model accounts for the relationship between the primary and secondary variables, as well as the redundancy between the secondary variables.

Merging the secondary variable relies completely on the correlation matrix. It is analogous to kriging's dependence on the covariance matrix. The correlation matrix must be positive definite for the merging process to complete successfully. When all of the data are homotopically sampled, i.e. every attribute is available at every location, there will be no problems. The correlation matrix is guaranteed to be positive definite. When the data are not homotopically sampled, the correlation matrix may or may not be positive definite.

There are many reasons that the sampling could not be homotopic. Cost of the samples, changes in the sampling protocol or bad samples are just a few of the reasons. When faced with limited samples for some variables, there are two choices that can be made: (1) use the small homotopic subset, or (2) use as many samples as possible between variable pairs. Consider a small three variable example with 100 locations. Variables 1 and 2 have been assayed at all 100 locations and variable 3 has only been sampled at 10 locations. Here is where we have to make the decision to use the subset of 10 samples and get a correlation matrix that is positive definite, or use a different number of samples for each variable. Using a different number of samples increases the

reliability of the calculated correlations, but it may also cause the correlation matrix to not be positive definite.

Most projects have enough samples that they can reliably use the subset of homotopic samples. However, some projects have very few samples for some of their critical variables. In these cases it is necessary to use a different number of samples for each variable pair. In these cases the correlation matrix may not be positive definite.

This paper proposes an iterative method for changing the correlation matrix to make it positive definite. The goal is to change as few of the terms as possible to make the matrix positive definite.

Methodology

We used an iterative scheme for correcting the correlation matrix. The goal of the correction is to make the matrix positive definite without making unnecessary changes to the correlation coefficients. Some aspects that will be considered are: (1) choosing which correlation in the matrix to change, (2) the number of samples used for calculating the correlation, (3) the sensitivity that a specific correlation has on the matrix, (4) the amount to change the correlation, and (5) the maximum weights that are calculated using the matrix. The above points will be discussed below.

Point 1: All of the nodes in the correlation matrix have an equal probability of being selected. The only nodes that are not allowed to change are the direct correlations; they are fixed to a value of one.

Point 2: The number of samples is the only indication available for the reliability of the correlation coefficient. As the number of samples increases, the correlation becomes more and more stable. This is accounted for in the matrix correction through a calibration factor.

The calibration factor allows all of the correlations to be changed. It just limits the amount of the change based on the number of samples. A simple function is used for calculating the maximum change allowed given the number of pairs. The function is:

$$CorrFact(ix,iy) = \begin{cases} 1.0 - 0.9 \cdot \dfrac{nsamp}{100}, & \text{if } nsamp \le 100 \\ 0.1, & \text{if } nsamp > 100 \end{cases} \tag{1}$$

This results in correlations with a small number of pairs having larger changes then correlations that have a large number of pairs. The threshold of 100 was chosen subjectively.

Point 3: Each element of the correlation matrix has a different impact on the matrix. Some elements will indefinitely be more important for the matrix correction. A simple iterative scheme is used to calculate the matrix sensitivities.

The initial Eigen values are calculated with the input matrix. Then, each element in the matrix is changed by a small amount. After changing the correlation, the Eigen values are recalculated. The difference in the minimum Eigen value, before and after the change, is calculated. This is repeated twice at each node. Once when the correlation is increased by 0.01 and then decreased by -0.01. The average absolute difference is used as the sensitivity value.

$$sens(ix,iy) = \frac{1}{2}\left[abs\left(eig_i - eig_{\rho(ix,iy)+0.01} \right) + abs\left(eig_i - eig_{\rho(ix,iy)-0.01} \right) \right] \tag{2}$$

Where eig_i is the minimum initial Eigen value and *eig* is the minimum Eigen value for a small change in the correlation.

Point 4: The iterative change made to the correlation matrix is a function of a random number, the number of samples used to calculate the correlation, and the sensitivity for that particular correlation. The change to the correlation is calculated as:

$$\Delta\rho = (\text{R-0.5}) \cdot sens(ix,iy) \cdot CorrFact(ix,iy) \cdot 0.005 \quad (3)$$

where R is a uniform random number between zero and one. Therefore, the maximum delta that could be applied for one iteration is 0.005.

Point 5: The last consideration is the weights that are calculated from the fixed matrix. Recall that a simple kriging system is used to determine the weights for combining the secondary data. Large weights, although theoretically correct, can cause major problems when combining the secondary data. The weights are not accounted for in the optimization process. The final weights are calculated after the matrix has been corrected. It is up to the user to ensure that the resulting weights are reasonable. A simple check is done in the program. A warning is written to the screen if the maximum weight is great than 0.9.

A maximum descent approach is used for fixing the correlation matrix. This was chosen over a more complex optimization method. A more complex method will not produce a better matrix. But it would increase the amount of time it takes the program to run. The outline of the program workflow is:

1. Calculate the initial Eigen values for the input correlation matrix,
2. If the minimum Eigen value is greater than specified, go to 11,
3. Randomly select a correlation coefficient to change,
4. Draw a random number R,
5. Calculate the change to the correlation coefficient for the selected location,
6. Update the correlation matrix,
7. Calculate the Eigen values for the updated matrix,
8. If the minimum Eigen value is greater than specified, go to 11,
9. If the minimum Eigen value has increased, keep the change and go back to 3,
10. If the minimum Eigen value has decreased, reverse the change and go back to 3,
11. Calculate the weights for combining the secondary variables,
12. Write a warning if the maximum weight is greater than 1.0,
13. Exit the program.

The output file appends three columns of information to the input file: (1) the fixed correlation matrix, (2) the delta between the initial and the corrected correlation matrices, and (3) the correlation sensitivity.

Example

An example was taken from a coal-bed methane data set. There are 14 variables in total. Of the 14 variables, 4 are primary variables that will be predicted using the remaining 10 variables as secondary data. Table 1 lists the number of samples for each data variable. The first four variables are the primary variables.

Table 1: Number of samples for the different variables.

Variable	Number of Samples
Spinner Flow	85
Surface Pressure	74
21 day Flow	66
21 day Pressure	66
Top	101
Base	105
Gross	101
Net Thickness	105
Number of Intersects	105
Average Thickness	105
Initial Gas Content	21
Initial Density	21
Langmuir Pressure	3
Reservoir Pressure	35

Since there are a variable number of samples, we will not use a homotopic subset for calculating the correlation matrix. This may result in a correlation matrix that is not positive-definite. Figure 1 shows the initial correlation matrix and the number of data used for the calculation.

The initial matrix is not positive-definite. This was checked by calculating the Eigen values for the matrix. The minimum Eigen value is -1.10. If all of the Eigen values are positive, the matrix has a unique solution. If any Eigen value is negative, the matrix needs to be corrected. In some cases, the matrix will need to be corrected when all of the Eigen values are positive to correct for large kriging weights.

The next step was to fix the correlation matrix. The parameters for the correlation matrix fixing program, `fixcorrmat`, are below:

```
                  Parameters for FIXCORRMAT
                  *************************

START OF PARAMETERS:
corrmat.out                    -file with correlation matrix
14                             -   number of variables
1  3                           -   columns for rho, ndata
0.2                            -minimum Eigenvalue
4    1 2 3 4                   -prediction variables
10   5 6 7 8 9 10 11 12 13 14  -data variables
fixcorrmat.out                 -file for output
```

The minimum Eigen value is the only subjective parameter that needs to be chosen. If set too low, the resulting weights will not be reliable. If set too high, the matrix correction will take too

long to run. Usually, it is set high to ensure a good matrix. We will see later that there is a maximum value for the minimum Eigen value.

The data variables and prediction variables are not used for the optimization. They are used only used for calculating the weights. The final weights are written to the output screen. If any weight is greater than 0.9, a warning is written as well.

Figure 2 shows the sensitivity for each variable in the correlation matrix. The sensitivity is a combination of the number of data and how that specific correlation impacts the minimum Eigen value. Figure 3 shows the corrected matrix and the delta between the initial matrix and the corrected matrix. Note that most of the changes are small, $<\pm 0.3$. However, there are a few large changes that were made to the correlation matrix. We want to ensure that the changes were made to appropriate correlations in the matrix; i.e. to correlations that had a small number of data or correlations that are sensitive (have a large impact on the Eigen values).

Scatterplots of the correlation delta versus the sensitivity and the number of data are shown in Figure 4. The relationships are where expected. Correlations that are highly sensitive had a large delta and correlations with a large number of pairs have a small delta.

Interesting plots are shown in Figure 5 and Figure 6. Figure 5 shows the minimum Eigen value for the matrix versus the number of iterations. The matrix is quickly corrected to have only positive Eigen values, within 20000 iterations, but 180000 iterations are required before a stable matrix is reached. Figure 6 shows the maximum weight versus the minimum Eigen value. The maximum weight does not stabilize until around 160000 iterations, at the same time that the minimum Eigen value stabilizes.

The last check done was to see how fixing the correlation matrix impacted the correlation between the primary variable and the combined secondary variables. This correlation is important when modeling the primary variable with the combined secondary variables in collocated cokriging or co-simulation. Ideally, the correlation will not change as the matrix is corrected. Figure 7 shows the correlations between the primary variables and the combined secondary variable that corresponds to the primary variable versus the number of iterations and the minimum Eigen value. As with the weights, the correlations do not stabilize until the minimum Eigen value has stabilized. It is interesting to note that the final correlations between the primary and the combined secondary are similar to what is expected from the initial correlation matrix. For example, in the initial matrix, spinner flow has correlations around 0.3 with the secondary variables. We would expect the correlation of the combined secondary to be greater than 0.3, but not significantly bigger. The resulting correlation is 0.42.

Conclusions

Merging multiple secondary variables is becoming a common practice in geostatistics; however, it may be impossible to get a positive-definite matrix in a sparse data setting. Matrices that are not positive definite can be fixed with the methodology presented in this paper. Correcting the matrix has two benefits: (1) it can fix matrices that are not positive-definite and (2) it stabilizes the weights from the matrix. The correlation fixing program iteratively changes correlations within the matrix to maximize the minimum Eigen value. As the minimum Eigen value increases, the matrix becomes more stable. Most matrices will asymptotically reach a maximum minimum Eigen value. When this has been reached, the matrix can be used for combining the secondary variables.

The changes made to the correlation matrix are constrained. Correlations that have a low number of data are preferentially changes over correlations that have a large number of data. In addition to this, correlations that have a higher impact on the minimum Eigen value are also changed more often than correlations that do not have an impact on the minimum Eigen value. The example presented showed how the program corrects the correlation matrix. It fixed the matrix and stabilized the weights for combining the secondary variable.

Reference

[1] S. Zanon and C. V. Deutsch. Direct Prediction of Reservoir Performance with Bayesian Updating Under a Multivariate Gaussian Model. In *Petroleum Society's 5th Canadian International Petroleum Conference (55th Annual Technical Meeting)*, Calgary, Alberta, Canada, June 8, 2004.

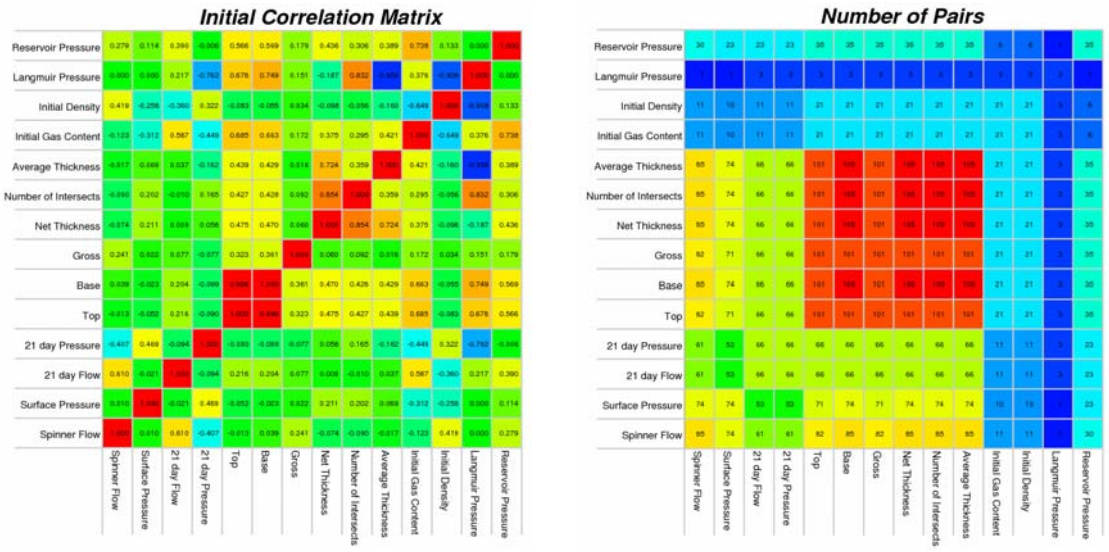

Figure 1: The initial correlation matrix and the number of data used for the calculation. The correlations are on the left and the number of data is shown on the right.

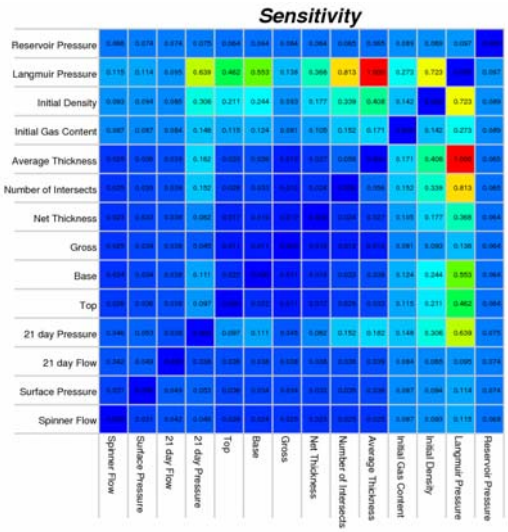

Figure 2: Sensitivity for each correlation coefficient. Correlations with a high sensitivity have a large impact on the minimum Eigen value.

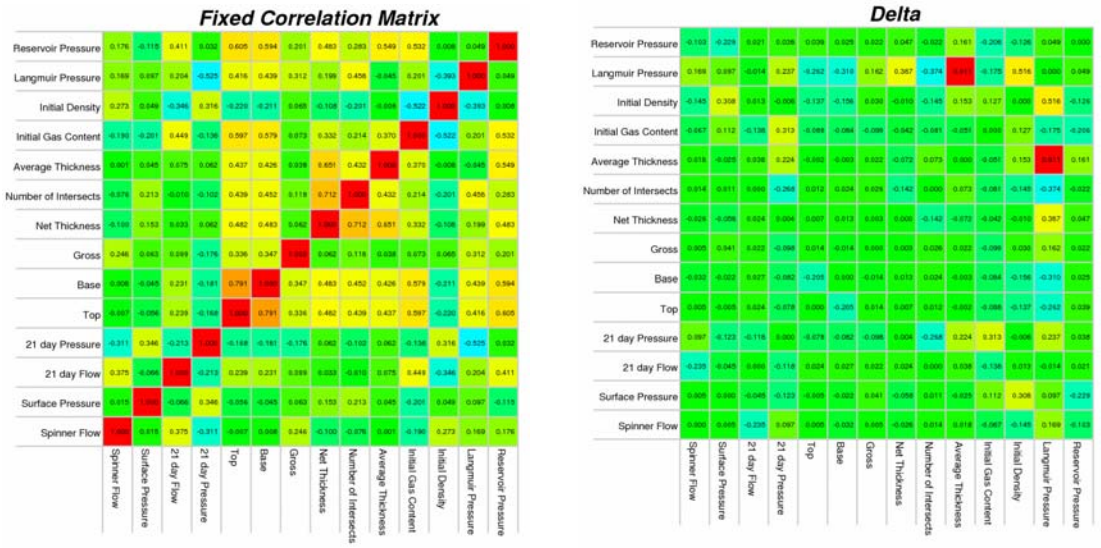

Figure 3: The corrected correlation matrix is shown on the left and the delta from the initial correlation is shown on the right. The color scale for the delta matrix is from -0.9 → 0.9.

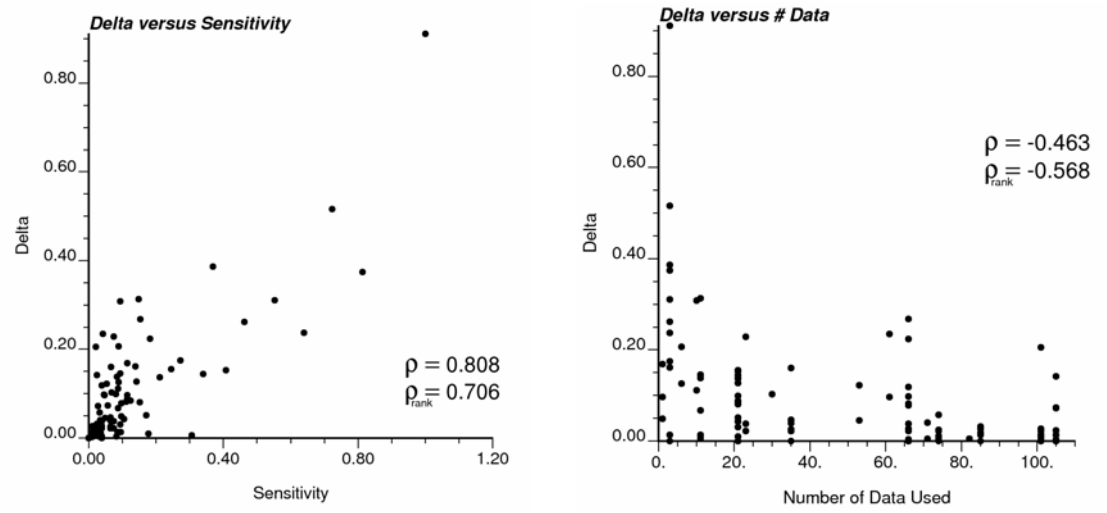

Figure 4: The scatter plots show the change in the correlation coefficient versus the matrix sensitivity and the number of data used for calculating the correlation.

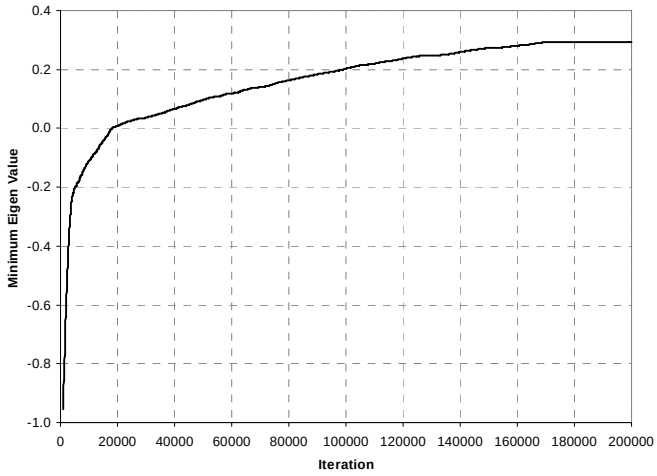

Figure 5: Minimum Eigen value versus the number of iterations.

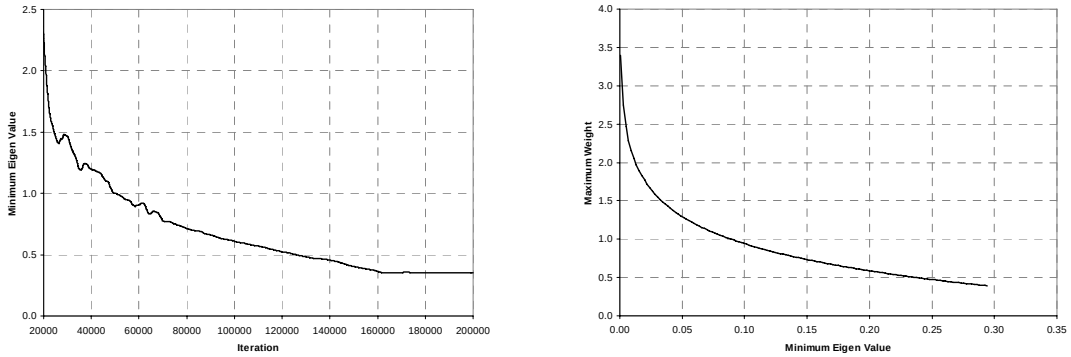

Figure 6: Maximum kriging weight versus number of iterations and minimum Eigen value.

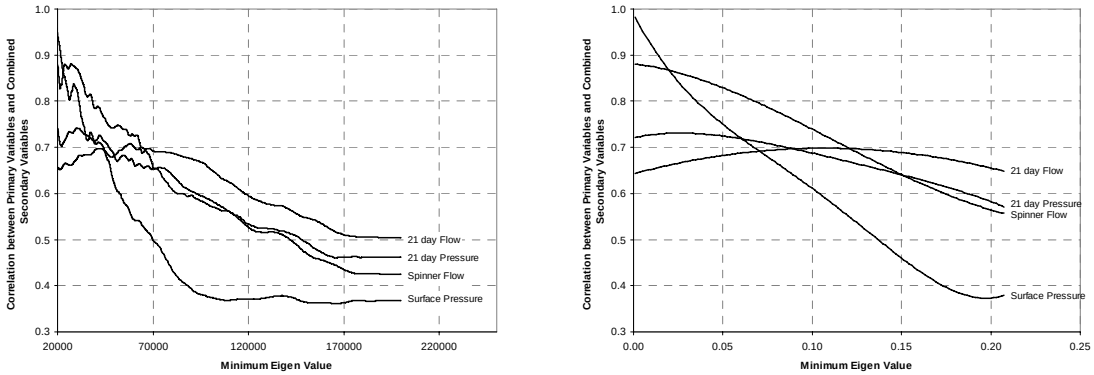

Figure 7: Correlation between the primary variable and the combined secondary variable that corresponds to the primary variable versus the number of iterations and the minimum Eigen value.