Overview of Multivariate Statistical Techniques for Geostatistical Applications

Oy Leuangthong (oy@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Abstract

The basis of all geostatistical modeling is statistical inference. Such inference is based on the data and the relations between them. To create sound geostatistical models of regionalized variables, the underlying relationship between the variables must be understood in order to apply the appropriate analytic tools. The assumption that the multivariate distribution is well behaved (i.e. homoscedastic and linear) is implicit when carrying out conventional geostatistics techniques; however, geologic data rarely conform to such well behaved distributions. In these instances, multivariate statistical techniques must be adapted to the data so that conventional geostatistical simulation can proceed.

Although multivariate statistical methods are well documented, few resources exist that documents their application to geostatistical problems. This paper provides an overview of multivariate statistical techniques for use in a geostatistical framework. Several groups of statistical techniques are presented: dimension-reducing, transformation, and classification techniques. During the exploratory data analysis phase of a study, the group of classification techniques could be effective in distinguishing between different populations. Once the decision to use only a certain number of samples is made, we may wish to try to reduce the number of different variables for simulation. In this instance, the dimension reducing techniques would be most effective. In the case where dimension reduction is not an issue, but non-linearity and/or complex multivariate constraints limit the applicability of typical geostatistical tools, the data transformation methods may prove particularly useful.

Introduction

Geostatistics is a relatively new and rapidly growing area in the geosciences and applied mathematics. The field is devoted to the application of statistical techniques in the study of spatially variable phenomena. Although geostatistics was first developed to improve ore reserve estimation in a mining context, it has grown to encompass other areas of the earth sciences.

Today the use of geostatistics is no longer hampered by the computational restrictions of its early days in the 1960s and 1970s. Technological advances in the past few decades make it possible to build complex 3-dimensional numerical models of geologic phenomena that were once limited by computational time and effort.

Modeling of geologic data typically involves consideration of multiple variables. For example, the modeling of a petroleum reservoir typically involves modeling porosity and permeability. In a mining context, several different minerals and/or metals may be present at the mine site.

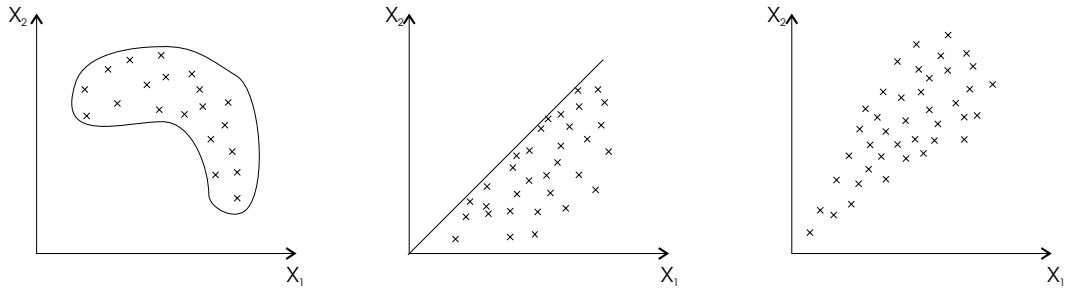

Figure 1: Schematic Illustration of Different Bivariate Distributions: Non-linear(left), Constraint(centre) and Heteroscedastic (right)

Problem Definition

The basis of all geostatistical modeling is statistical inference. Such inference is based on the data and the relations between them. To create geostatistical models of the regionalized variables, the underlying relationship between the variables must be understood in order to apply the appropriate analytic tools.

Conventional geostatistical approaches dealing with multiple variables include simulation using either collocated or full co-kriging. Both approaches require that the primary variable be defined at every location within the grid. The difference in the two approaches lies in the amount of secondary data used in simulation: the former technique uses only the collocated primary data in simulation while the latter technique uses secondary data within the range of correlation.

An important assumption inherent in conventional techniques is that the multivariate distribution is homoscedastic and linear; however, geologic data rarely satisfy such assumptions. Instead, the multivariate distributions may show signs of non-linearity, mineralogical constraints, and heteroscedasticity (See Figure 1). In these instances, our concern lies in the available statistical techniques that can be applied to transform the data so that conventional geostatistical simulation can proceed.

Many techniques exist to analyse multivariate data; however, the application of these techniques in geostatistics have been limited. Multivariate statistical techniques are well documented; however, few resources exist that documents these different techniques with specific geostatistical applications. The objective of this report is to provide an overview of multivariate statistical techniques for use in a geostatistical framework.

The following sections look at different groups of techniques. The first class of techniques are the dimension reducing methods, which include principal components analysis (PCA) and factor analysis. These seek to simplify the multivariate problem by reducing the number or dimension of the data. Another distinct set of approaches is data transformation. In particular, two not-so-common yet very powerful techniques are discussed: stepwise conditional transformation and alternating conditional expectation. The next group of techniques are the classification techniques which include both discriminant analysis and cluster analysis. As the name suggests, both approaches focus on the classification of data into groups - one requires that groups are pre-established while the other seeks to define the different groups.

Throughout this paper, a multivariate data set from a nickel laterite deposit is used to demonstrate and compare the results of most of the multivariate techniques presented. The data is comprised of 4 mineral variables: nickel, iron, silicate oxide, and magnesium oxide.

Dimension Reducing Methods

As the title suggests, the techniques presented in this chapter have a common aim: to reduce the dimension of the data, thus simplifying the multivariate problem. The two primary approaches that fall within this group of methods are principal components analysis and factor analysis. The development of the former technique is credited to the work of Pearson in 1901 and Hotelling in 1933, while the concept of factor analysis first originated with work by Spearman in 1904 and 1926. The following sections describes the basic theory and methodology of principal components and factor analysis.

Principal Components Analysis

The basis for principal components analysis (PCA) is the transformation of correlated variables into uncorrelated variables called principal components. We first begin by looking at the n centered samples (i.e. residuals) of the p random variables Z_i, where $i = 1, \ldots, p$, and denoted as Z in matrix notation.

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix} = Z_{(n \times p)}$$

The covariance matrix of the data is then given by:

$$Var(Z_{(n \times p)}) = \frac{1}{n} Z^T Z = V_{(p \times p)}$$

Define variables Y_i, $i = 1, \ldots, p$, such that they are uncorrelated with each other with zero mean, i.e. $E\{Y_i\} = 0$. Furthermore, these variables are linear combinations of the original variables:

$$Y_{(n \times p)} \quad = \quad Z_{(n \times p)} \times A_{(p \times p)} \qquad with \quad A^T A = I$$

where I is the identity matrix and A is an $p \times p$ orthogonal matrix of coefficients, referred to as a transformation matrix. The covariance matrix of the Y variables is:

$$Cov(Y) = \frac{1}{n} Y^T Y = \begin{bmatrix} c_{11} & & 0 \\ & \ddots & \\ 0 & & c_{pp} \end{bmatrix} = C_{(p \times p)}$$

where c_{ii} is the variance of the Y_i variable. Solving for the matrix A becomes a spectral decomposition problem [18]:

$$
\begin{aligned}
Y &= ZA \\
\frac{1}{n}Y^TY &= \frac{1}{n}Y^TZA \\
C &= \frac{1}{n}(ZA)^TZA \\
C &= \frac{1}{n}A^TZ^TZA \\
C &= A^T\left(\frac{1}{n}Z^TZ\right)A \\
C &= A^TVA \\
VA &= AC
\end{aligned}
$$

Therefore the transformation matrix A is simply a matrix of orthonormal eigenvectors of V. Since the covariance matrix is a positive definite matrix, then all the eigenvalues are positive and are interpreted as the variance of the Y_i variables.

These Y_i variables are the principal components of Z. The importance of a principal component is derived directly from the rank of the eigenvalue, i.e. the largest eigenvalue corresponds to the principal component that contributes maximally to the variance of the Z data. If the variability of the data can be adequately captured by consideration of only the first few principal components, then the dimension of the original multivariate problem is reduced.

One consequence of finding uncorrelated variables that maximizes the variance is the sensitivity to outliers. Outliers inflate the variance of the data, and as a result the principal components may not account for the variance of the true, representative data (i.e. excluding outliers).

Steps in PCA

1. Calculate the residual value of the data samples for each variable by determining the mean of variable X_i and then subtracting this mean from each sample to obtain Z_i, $i = 1, \ldots, p$ where p is the number of variables.

2. Determine the covariance matrix, V between the centered data, Z_i and Z_j, $i, j = 1, \ldots, p$.

3. Find the eigenvalues and corresponding eigenvectors. The coefficients of the first principal component are given by the eigenvector, a_{1i}, $i = 1, \ldots, p$, and its variance is given by its eigenvalue, λ_1.

4. Decide whether to discard the lowest variance-contributing principal component(s). Conduct geostatistical analysis using the remaining principal components.

5. After geostatistical analysis and simulation, back transform principal components into the original centered variables Z_i by matrix multiplication using the inverted rotation matrix, A.

The use of standardized variables translates to finding the principal components of the correlation matrix.

Application of PCA

A principal components analysis was performed on the nickel laterite data. Figure 2 shows cross plots between the four principal components, confirming their independence. Figure 3 shows the relative importance of each component based on their variance contribution. As expected, the first component accounts for the maximal variance of approximately 91%, followed by the second component with a variance contribution of 6.7%. If we were satisfied with simulating at least 95% of the deposit's variability, then simulation is greatly simplified by considering only two independent variables. This simplification of geostatistical simulation is another attractive consequence of PCA.

Factor Analysis

The goals of factor analysis are similar to those of PCA: to reduce the dimension of the data by finding uncorrelated variables. One of the most obvious differences between the two techniques is the definition of the uncorrelated variables, which are referred to as *factors* in this approach. The n standardized data of the p-dimensional random variable Z are individually defined as a linear combination of the m factors ($m < p$), f_k, plus an independent "error" term, ε_i, specific to Z_i, $i = 1, \ldots, p$. So the j^{th} sample of Z_i is defined as:

$$z_{ji} = \sum_{k=1}^{m} a_{ik} f_{kj} + \varepsilon_{ji} \qquad i = 1, \ldots, p \quad j = 1, \ldots, n$$

where f_{kj} is the k^{th} common factor for the j^{th} sample, ε_{ij} is the specific factor for the i^{th} variable of the j^{th} sample, and a_{ik} is the factor loading on the k^{th} factor for the i^{th} variable. Each variable then has different factor loadings. For obvious reasons, the f_{kj} factors are often referred to as the common terms, while the ε_{ij} term is referred to as the specific factor. In matrix notation, the data variable Z is expressed as:

$$\begin{aligned} Z_{(n\times p)} &= F_{(n\times m)} \times A_{(m\times p)} + \varepsilon_{(n\times p)} \\ Z &= AF + \varepsilon \end{aligned} \tag{1}$$

This technique is dependent on several key assumptions:

1. The sample data X are standardized to obtain Z with zero mean and unit variance.
2. The common factors, F, are assumed to be uncorrelated with zero mean and unit variance.
3. The specific "error" term ε has a mean of zero and is assumed to be independent of the common factors and each other . However, no assumption is made about the variance of the specific term, the value of which is denoted by ψ.

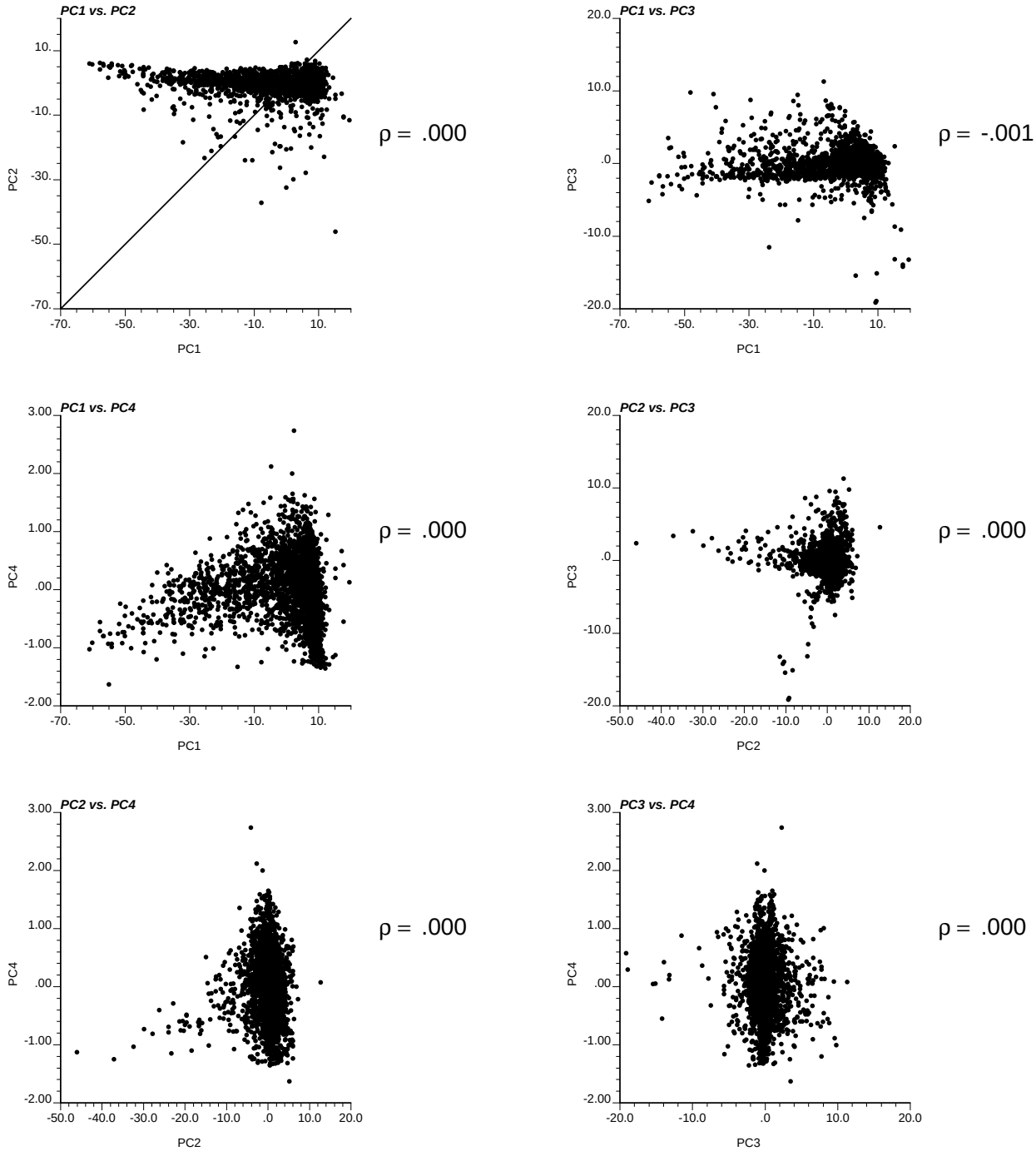

Figure 2: Cross Plot of Principal Components using PCA on Nickel Laterite Deposit with 4 Variables

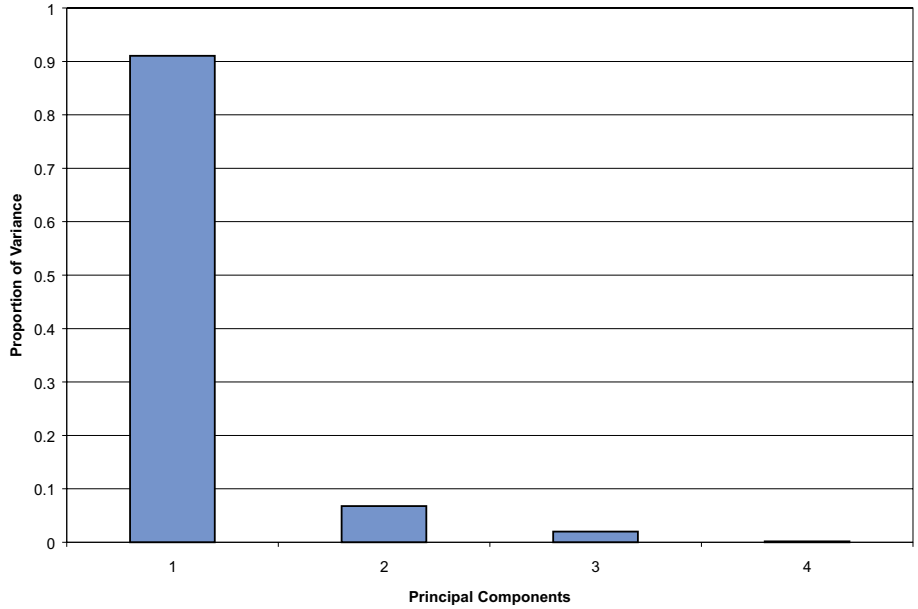

Figure 3: Variance Contribution of Each Principal Component

Arising from these assumptions is another fundamental difference between the two methods: factor analysis is based on a specific statistical model while PCA is not based on any statistical model [13].

From equation 1, it follows that the variance-covariance matrix of Z is given by:

$$\begin{aligned} Var(Z) &= Var(AF + \varepsilon) \\ Var(Z) &= A^T A + \psi \end{aligned} \tag{2}$$

Since the specific factors are uncorrelated with each other, then the off-diagonal terms of ψ are zeros. Equation 2 shows that the off-diagonal terms of $Var(Z)$ (i.e. covariance between the observations of Z_i) are explained solely by the factor loadings. Furthermore, since Z has unit variance, the correlation between two observations z_i and z_j is given by :

$$r_{ij} = \sum_{k=1}^{m} a_{ik} a_{jk} \tag{3}$$

From equation 3, two observations are highly correlated if the factor loadings for both observations are high for the same factors [4, 13]. For this reason, factor analysis is associated with finding the factors that contribute to maximal *covariance* while PCA is concerned with finding components with maximal *variance*.

Another important measure is the *communality* of factors on Z_i, which quantitifies how much of the variance of Z_i is accounted for by the m common factors. This is defined as the sum of the square of the factor loadings for that variable:

$$Comm(Z_i) = \sum_{k=1}^{m} a_{ik}^2 \tag{4}$$

Steps in Factor Analysis

1. Standardize the original data X_i to obtain Z_i, $i = 1, \ldots, p$.

2. Determine factor loadings, A. The most common way of doing this is to perform PCA on the data, and use the first m principal components as the m factors.

$$\begin{aligned} Y_{(n\times p)} &= Z_{(n\times p} \times A_{E(p\times p)} \\ Z_{(n\times p)} &= Y_{(n\times p)} \times A_{E(p\times p)}^T \end{aligned}$$

   where Y is the matrix of principal component scores from PCA and A_E is the rotation matrix from PCA and is referred to as the extraction matrix in factor analysis. Choose the first m principal components as initial factors for factor analysis:

$$Z_{(n\times p)} = Y_{(n\times m)} \times A_{E(m\times p)}^T + \varepsilon_{(n\times p)}$$

   Note that although the m factors are independent of (1) each other and (2) the specific factor, the specific factors themselves are *not* uncorrelated with each other. This essentially violates the third assumption of this technique; however, since the first m principal components account for the maximal variance, the correlation between the specific factors should be relatively insignificant.

3. Perform a factor rotation to find new factors that describes the data equally well. Essentially, this involves finding the rotation matrix to the $Y_{(n\times m)}$ factors obtained from PCA in the previous step, so that the new factors are linear combinations of the PCA factors. A standard method to finding an orthogonal factor rotation is called varimax rotation, which basically maximizes the sum of the factor loadings, and hence the covariance of the data.

$$Z_{(n\times p)} = F_{(n\times m)} \times A_{R(m\times p)}^T + \varepsilon_{(n\times p)}$$

   where A_R is the rotation matrix by optimizing some predefined function (based on an optimization criterion such as varimax), and F_{ni} is equivalent to the scaled PCA factors after dividing by the standard deviation of Y_i, or the square root of the i^{th} eigenvalue. Note that the factor rotation need not be uncorrelated (orthogonal), it can also be correlated (oblique).

4. Calculate the factor scores for each observation [9]:

$$F = (A_R^T A_R)^{-1} A_R^T Z \tag{5}$$

Perform geostatistical modeling using these uncorrelated factor scores.

5. After geostatistical simulation, back transform the simulated factor scores by using the relation in equation 5 in step 4. Back transform twice more by using the inverted factor rotation matrix in step 3 *and* the PCA factor extraction matrix in step 2 to get results in terms of the original standardized data.

Application of FA

Factor analysis was performed on the nickel laterite data. Principal components analysis was used to extract the initial loadings from the standardized data. A table of the eigenvalues and the percentage of the variance explained by each component is given in Table 1.

Principal Component	Eigenvalue	% of Variance	Cumulative
PC1	2.755	68.87	68.87
PC2	0.958	23.96	92.83
PC3	0.232	5.79	98.62
PC4	0.055	1.38	100.00

Table 1: Initial Loadings to be used for Factor Analysis

The initial eigenvalues show that in order to account for at least 95% of the deposit's variability, 3 common factors are required (i.e. $m = 3$). The corresponding factor extraction matrix (A_E) obtained from PCA and the communalities for each variable are shown in Table 2.

Variable	Factor 1	Factor 2	Factor 3	Communality
Ni	-0.253	0.967	-0.003	1.000
Fe	-0.977	-0.083	0.052	0.965
SiO_2	0.923	0.091	0.367	0.994
MgO	0.940	0.085	-0.307	0.985
Variance	2.7548	0.9583	0.2317	3.9448
% Variance	0.689	0.240	0.058	0.986

Table 2: Factor Extraction Matrix

All the communalities are very high, so most of the variance for each variable is accounted for by using 3 factors. In fact, 98.6% of the total variance is represented by the common factors. An examination of the magnitude of the factor loadings on each factor (ignoring the signs), we see that Factor 2 explains almost all of the variance of Ni, and that Factor 1 explains most of the variance of Fe, SiO_2 and MgO. The fact that 3 of 4 variables are explained by only 1 of the factors suggests that a factor rotation may simplify the factors (and hence interpretation of the results). Varimax rotation was performed and the resulting factor rotation matrix and corresponding communalities are given:

Variable	Factor 1	Factor 2	Factor 3	Communality
Ni	-0.068	-0.053	0.996	1.000
Fe	-0.805	-0.557	0.082	0.965
SiO_2	0.507	0.856	-0.067	0.994
MgO	0.932	0.333	-0.074	0.985
Variance	1.7781	1.1574	1.0093	3.9448
% Variance	0.445	0.289	0.252	0.986

Table 3: Factor Rotation Matrix Obtained from Varimax Rotation

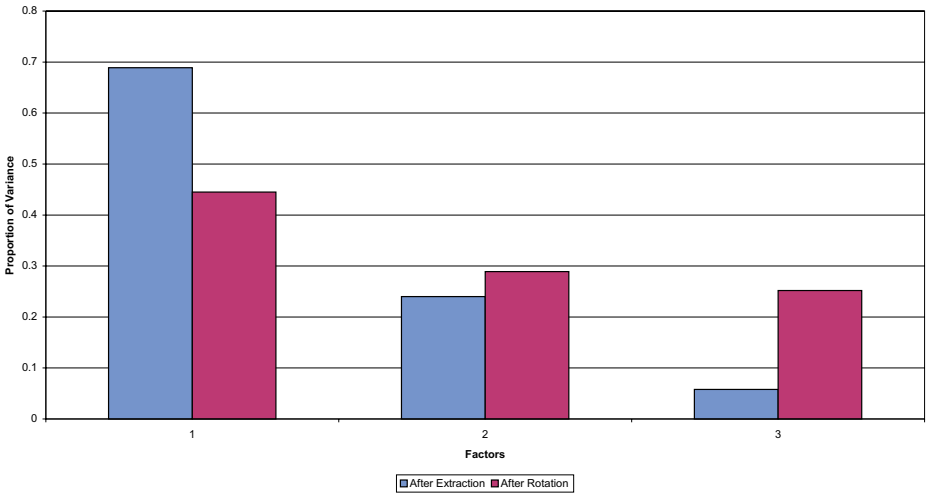

Figure 4: Variance contribution of each factor before and after varimax rotation

From Table 3, we see that the same variance proportion is explained after factor rotation, but now interpretation of the factors is simpler: Factor 3 is strongly related to Ni, while Factors 1 and 2 explain the variability of Fe and SiO_2, and Factor 1 explains most of variance of MgO.

Figure 4 summarizes the variance contribution of each factor before and after rotation. Overall, the importance of the factors are more balanced after rotation. Cross plots of the three factors are shown in Figure 5 to illustrate the independence of the rotated factors after varimax rotation. Note that an oblique transformation would have produced correlated factors. However, the use of varimax rotation in this example and the resulting uncorrelated factors will greatly simplify the geostatistical simulation of this deposit.

General Comments on Dimension Reducing Techniques

From the previous sections, we note that PCA is a variance-oriented technique while FA is covariance-oriented. There exists a very specific statistical model within the FA approach, while PCA is not dependent on any particular model. The solution obtained from PCA is unique and exact; while FA produces several possible solutions, owing to the available options in factor extraction and rotation methods. This makes PCA a more attractive

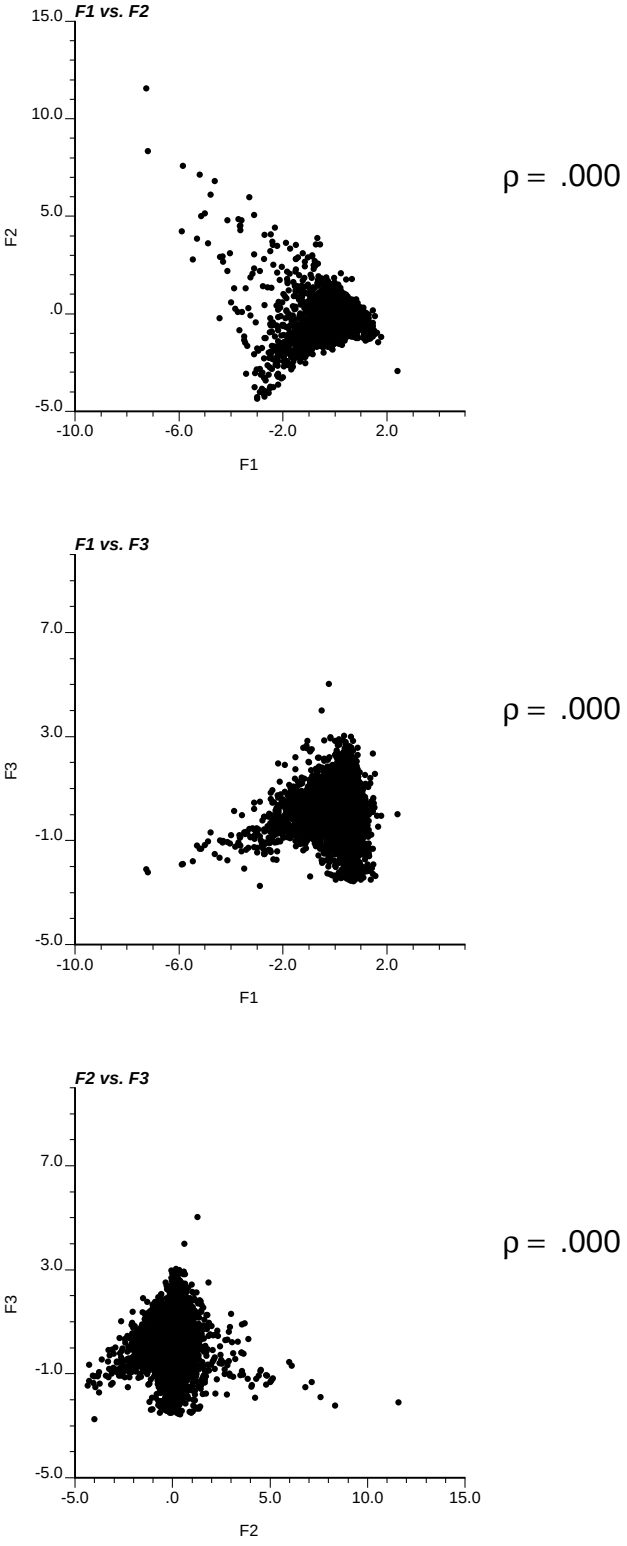

Figure 5: Cross plots of rotated factors

technique to most statisticians. Furthermore, Seber (1984) notes that factor analysis is not suitable for categorical data [15].

Attempts at non-linear FA has resulted in some theoretical and practical difficulties [10]. Due to these slight differences, there is considerable confusion between the techniques in the literature. Overall though, PCA is more commonly applied while factor analysis seems to remain popular in its founding discipline of psychology. Regardless of the current popular areas of application, considerable potential exists for the application of these dimension-reducing techniques in geostatistics.

Data Transformation Methods

This chapter focuses on the transformation of one set of variables to another set that simplifies both analysis and simulation. Although the techniques discussed in the previous chapter are technically data transformation techniques, the distinction is made in that this group of methods does not strive to reduce the dimension of the problem. Two methods - stepwise conditional transformation and alternating conditional expectation - are discussed. Neither are common to geostatistics, but the following sections will show why they are attractive for future applications in multivariate geostatistical analysis.

Stepwise Conditional Transformation

This technique, first introduced by Rosenblatt in 1952, bears resemblance to the normal transformation technique. In the univariate case, the stepwise-conditional technique is identical to the normal score transform, that is, the variable is transformed using the Gaussian distribution.

In a bivariate problem, the normal transformation of the second variable is conditional to the probability class of the first variable. Correspondingly, for k-variate problems, the k^{th} variable is conditionally transformed based on the $(k-1)^{th}$ variable [14].

$$\begin{aligned} Y_1 &= G^{-1}[Prob(Z_1 \leq z_1)] \\ Y_{2|1} &= G^{-1}[Prob(Z_2 \leq z_2 \mid Y_1 = y_1)] \\ Y_{3|21} &= G^{-1}[Prob(Z_3 \leq z_3 \mid Y_2 = y_2, Y_1 = y_1)] \end{aligned}$$

The result of this transformation are uncorrelated transformed variables. Since each class of Y_2 data is independently transformed to a normal distribution, any correlation between Y_2 and Y_1 is essentially removed (i.e. $\rho = 0$). Consequently, the simulation of a multivariate problem will not require cosimulation due to the independence of the transformed variables. This is the primary motivation for transforming multiple variables in a step-wise conditional fashion.

Limitations of the Stepwise Conditional Transformation

There are several limitations to this transformation method. The modeler should recognize that the variable Y_2 (and all other conditionally transformed variables) are not "real"

variables. Direct back transformation of the Y_2 variable using an inverse Gaussian transformation will not yield correct results. Each variable must be back transformed using the appropriate conditional distribution used in the forward transformation.

The main limitation of the stepwise conditional transformation lies in the need for a large data set. In order to classify data and transform each class, there must be sufficient data to identify a conditional distribution. If there is sparse data in one class, the conditional distribution will not be representative of the "true" distribution for that class. Transformation may not remove the correlation between Y_1 and Y_2 data.

In addition, transformation of classified data may lead to artifacts in the classes. The classification of data is based on partitioning the standard normal distribution according to the number of classes specified. The calculated probability thresholds correspond to equal probability intervals. Two identical secondary data values should have the same transformed values (Y_2) if their paired primary data values are the same; however, if the corresponding primary variables just happen to fall into different probability intervals, then transformation may produce significant differences in the secondary variable Y_2.

In the presence of non-paired data, two particular scenarios are of interest: (1) different sampling density, and (2) partially overlapping samples. In the first case, if one variable is more highly sampled than all others, then the more densely sampled variable should be chosen as the primary sample. All other variables should be conditionally transformed in an order corresponding to the number of available data (highest to lowest). Unlike the first case where a possible ordering sequence may be applicable, the second scenario presents a bigger limitation to the transformation algorithm. If only a fraction of the primary and secondary data are paired (i.e. overlapping sampling areas), then transformation of the remaining non-paired data cannot be obtained using this technique.

Application of Stepwise Conditional Transformation

Without loss of generality, data transformation was performed on only two of the four data variables from the nickel laterite data. Nickel was chosen as the primary variable and iron was arbitrarily chosen as the second variable.

Cross plots are shown for the original data and the stepwise conditionally transformed variables in Figure 6. The cross plot of the original data clearly shows the non-linear relationship between nickel and iron. After transformation, not only is non-linearity removed but the transformed data are virtually uncorrelated, allowing for simplified simulation of the model variables.

Alternating Conditional Expectation (ACE)

The alternating conditional expectation (ACE) algorithm was first introduced by Brieman and Friedman (1985) [2], as a flexible and powerful non-parametric transformation that requires no assumption to be made about the functional form of the multivariate distribution.

We begin by defining random variables (RVs), $Y, X_1, \ldots, X_p$, where Y is a response variable and $X_1, \ldots, X_p$ are the predictor variables. Arbitrary functions $\theta(Y), \phi(X_1), \ldots, \phi(X_p)$ with zero mean corresponding to these variables are also defined. The theoretical basis of the algorithm assumes the distributions for RVs $Y, X_1, \ldots, X_p$ are known, and $E\theta^2(Y) = 1$.

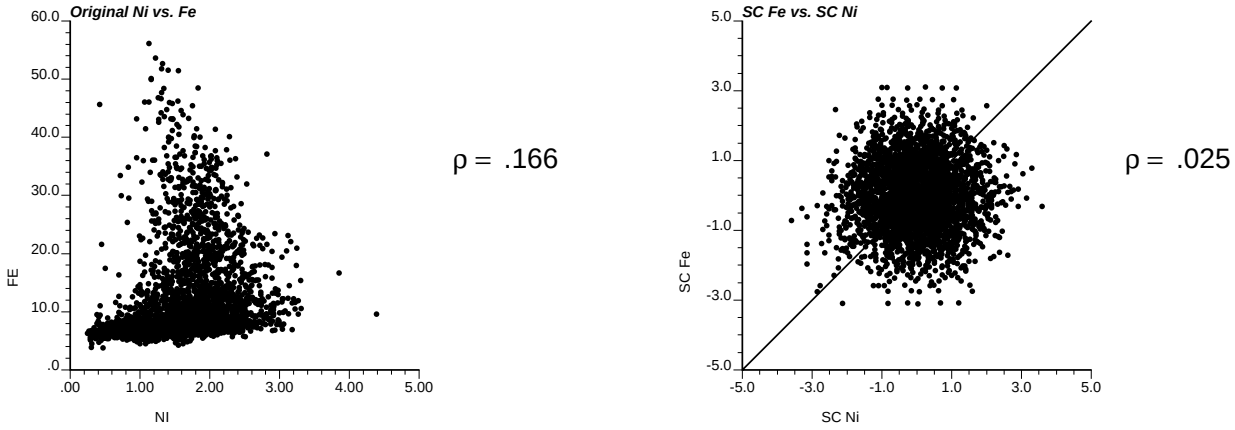

Figure 6: Cross plots of the original data and the stepwise conditionally transformed data for Nickel and Iron from the Nickel Laterite data.

Regression of $\theta(Y)$ is performed using $\sum_{i=1}^{p} \phi_i(X_i)$. The fraction of the variance not explained by regression is quantified as:

$$e^2(\theta(Y), \phi(X_1), \ldots, \phi(X_p)) = \frac{E\{[\theta(Y) - \sum_{i=1}^{p} \phi_i(X_i)]^2\}}{E\theta^2(Y)} \tag{6}$$

Optimal transformations are chosen as those which minimize 6 with respect to *all* the random functions $\theta(Y), \phi(X_1), \ldots, \phi(X_p)$.

Basic Algorithm

The ACE algorithm is an iterative procedure that is used to find the optimal transformations $\theta^*, \phi_1^*, \ldots, \phi_p^*$. Without loss of generality, consider the bivariate case where p=1. The optimal transformations, θ^* and ϕ_1^*, minimize:

$$e^2(\theta(Y), \phi(X)) = E[\theta(Y) - \phi(X)]^2 \tag{7}$$

The algorithm is carried out in the following steps:

1. Set $\theta(Y) = Y \div \|Y\|$ where $\|\cdot\| = \sqrt{(\sum(\cdot)^2)}$
2. Repeat until equation 7 is a minimum:

   Calculate $\phi_1(X)$:

   $$\phi_1(X) = E[\theta(Y)|X]$$

   Set $\phi(X) = \phi_1(X)$

   Calculate $\theta_1(X)$:

   $$\theta_1(Y) = \frac{E[\phi(X)|Y]}{\|E[\phi(X)|Y]\|}$$

   Set $\theta(X) = \theta_1(X)$

   Calculate Δe^2.

3. θ and ϕ are solutions to θ^* and ϕ^*

In the bivariate case, the optimal transformations not only minimizes e^2, but they also satisfy:

$$\rho^*(X,Y) = \rho(\theta^*, \phi^*) = max\rho[\theta(Y), \phi(X)] \tag{8}$$

where $\rho^*(X,Y)$ is the maximal correlation between X and Y. Essentially, the method is aimed at finding the optimal transformations for the functions that make the relationship between $\theta(Y), \phi(X_1), \ldots, \phi(X_p)$ as linear as possible. This allows for the application of conventional geostatistical techniques (which assume a linear relationship between the model variables).

One of the most important assumptions implicit in ACE is that the RVs $Y, X_1, \ldots, X_p$ have a multivariate distribution [3]. As well, unlike its theoretical basis the distributions of the RVs $Y, X_1, \ldots, X_p$ are not known in practice. Instead, we typically have limited samples from which a distribution is usually assumed or fitted. In this respect, the goals of ACE are modified so that the optimal transformations $\theta^*, \phi_1^*, \ldots, \phi_p^*$ are now estimated from the data themselves, rather than based on assumed distributions [2]. ACE, in its practical form, is an iterative algorithm. It uses a smoothing algorithm to estimate the conditional expectation. There are many smoothing techniques that can be used, however, Friedman uses his own developed algorithm - the *supersmoother*.

Application of ACE

For illustrative purposes, the first example is based on 200 data values generated using the model $y = exp(1 + 2x) + \varepsilon$, where x is U(0,2) and ε is N(0,10). Figure 7 shows the cross plots between the original data, the predictor variable and its transform, the response and its transform, and the transformed response vs. the transformed predictor.
Clearly, the bivariate distribution of the transformed variables is more linear than that of the original variables. Non-linearity between the original variables has been removed without assuming the form of the data distribution.

ACE was then performed on the nickel laterite data, again looking only at two variables: nickel and iron. The bivariate relation between the original and transformed variables are shown in Figure 8.

Again, the cross plot of the transformed variables shows an increase in the correlation coefficient between the transformed variables. In this instance, ACE has indeed made the transformed variables more correlated, however the resulting distribution shows evidence of some constraint in the transformed Fe (all values lie below 1.0). Some other transformation should be applied in order to remove this constraint, for example a data re-expression algorithm could be applied [11, 12]. Overall, linearity of the transformed bivariate distribution is not apparent, and so the assumptions of linear correlation inherent in conventional geostatistical analysis may still not be adequately satisfied after applying the ACE transform.

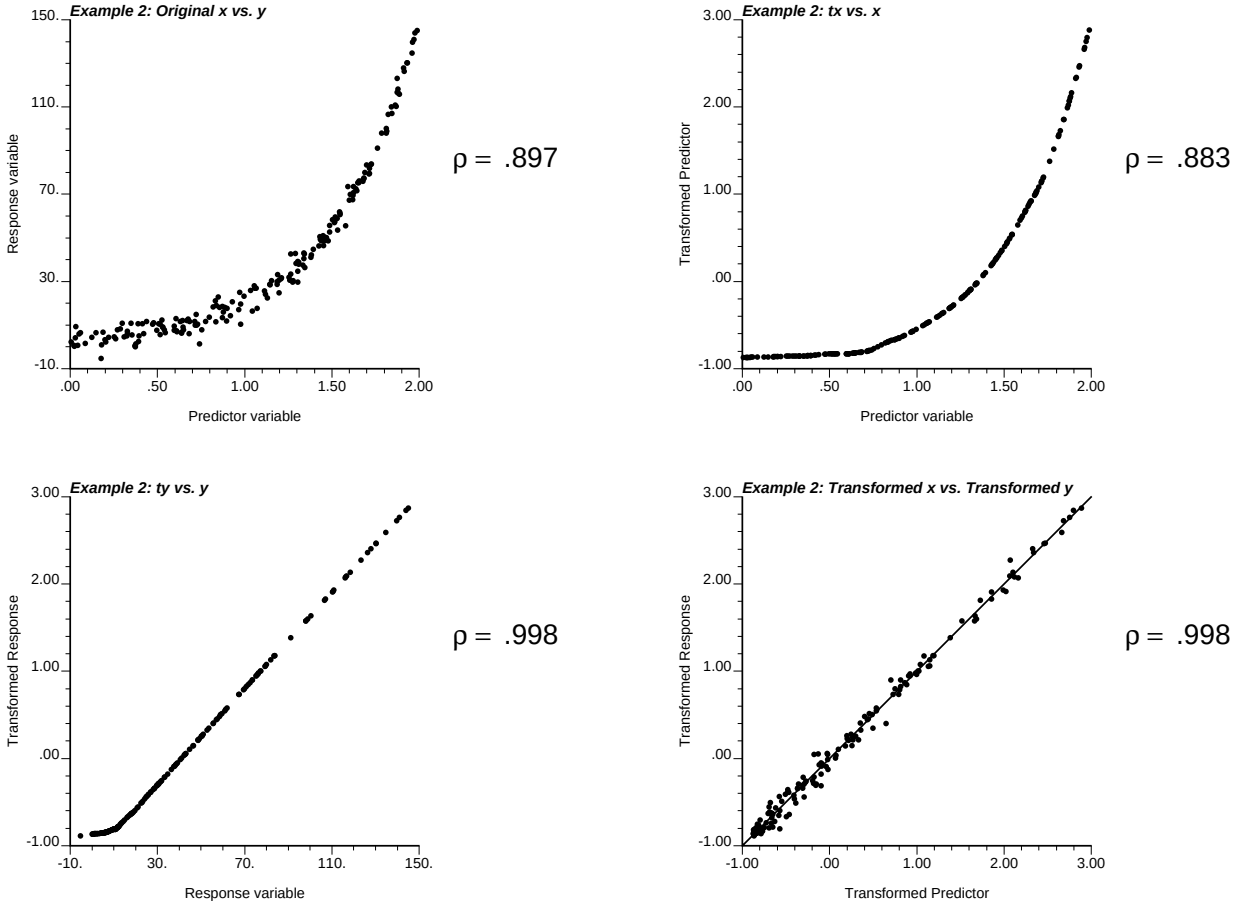

Figure 7: Tranformation of $y = exp(1 + 2x) + \varepsilon$, where x is U(0,2) and ε is N(0,10)

Classification Techniques

Up to now all the techniques have been concerned with finding an alternate set of variables that can be used to simplify geostatistical modeling. In this section, we shift to methods that may be useful from an exploratory data analysis perspective.

Discriminant Analysis

Discriminant analysis involves two types of multivariate data: 1) a set of groups with known distribution, and 2) a set of data with unavailable *a priori* information on the group that it belongs [7]. The objective of discriminant analysis is to reconcile the second type of data with the first type based on the different observations on each sample. In a geological context, this technique may be useful to determine the properties that characterize the different facies types. Based on these properties, samples taken from unidentified facies can be classified based on different observations on that sample to determine the facies group to which it likely belongs.

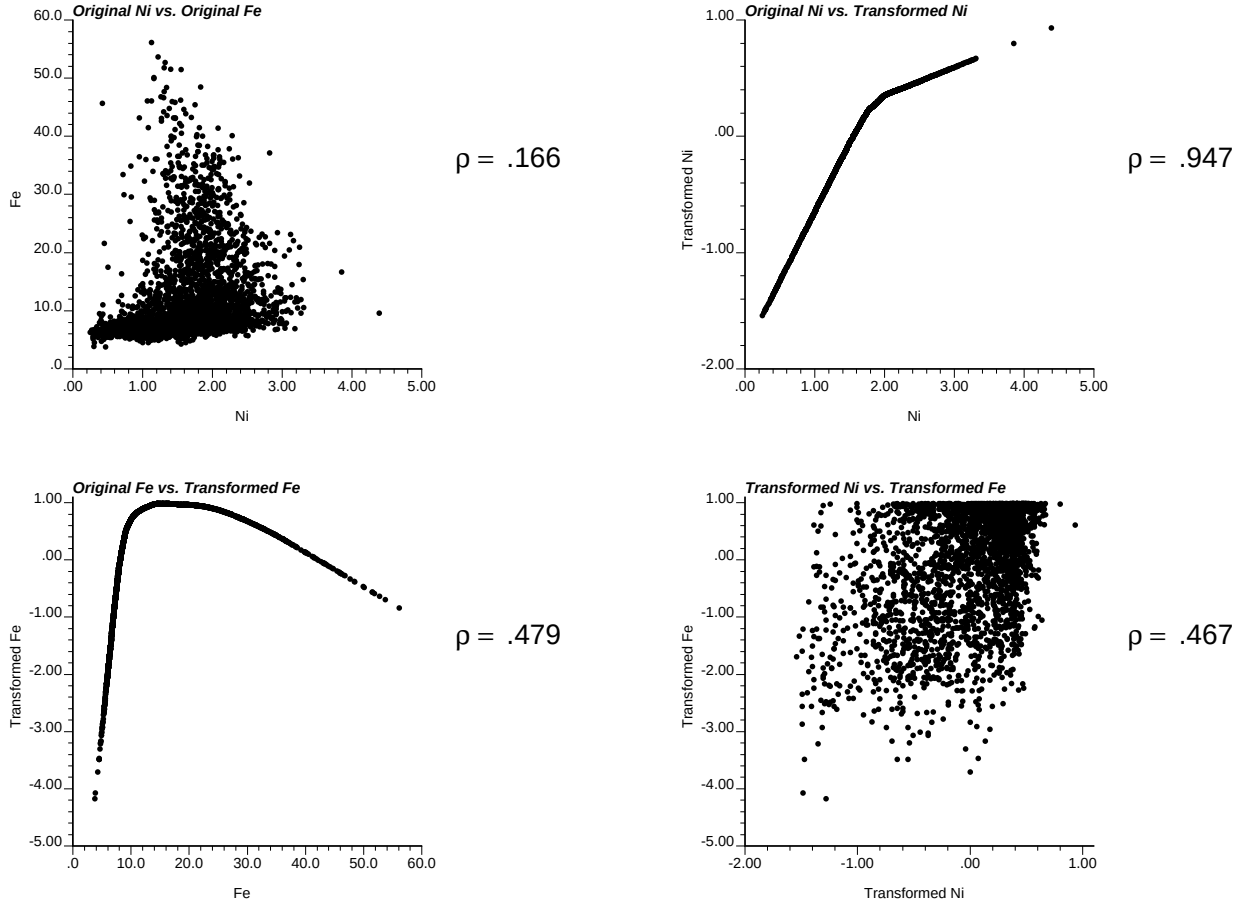

Figure 8: Cross plots for ACE transformation of Nickel Laterite Data

The first step in discriminant analysis is to represent the observations that clearly fall within the different G groups. Once done, this set of data then becomes the measure by which the groups are characterized. For example, data known to be sampled within a certain sedimentary layer, say sandstone, would belong to the sandstone group. Likewise, samples taken within shale would be grouped and identified as the shale facies group. Based on the grouped samples, a representation of the different measurements would show the separation between the groups. This is usually done in a spatial context and is typically referred to as the discriminant space [7].

The second step is to determine the variables that best discriminates between the two or more groups. The focus is now shifted to finding the *right* measure to classify the unknown data. Many techniques exist that classify the data based on different criteria; these include two-group linear discriminant analysis, heterogeneous covariance matrices, classification by nearest neighbor methods, classification into one of several groups and classification using Mahalanobis distances. Seber [1984] provides a schematic illustration of the linear and quadratic discriminant analysis techniques, which is reproduced here in Figure 9 [15].

The cost of misclassifying the data is another fundamentally important concept central

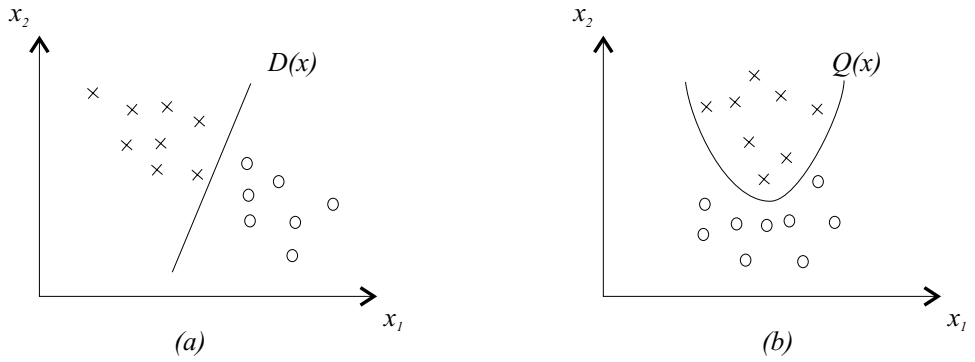

Figure 9: Use of (a) linear and (b) quadratic discriminant analysis to separate groups of data. Source: Seber (1984)

to discriminant analysis. The determination of the *right* classification measure is an optimization problem wherein the cost associated to classifying the i^{th} sample to the wrong group must be minimized. The result of this optimization is the following misallocation rule where x is assigned to group G_1 if :

$$\frac{f_1(x)}{f_2(x)} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)} \tag{9}$$

where f_i is the probability density function of the i^{th} group, π_i is the prior probability that x belongs to the i^{th} group, and $C(i|j)$ is the cost of misclassifying a sample from group j to group i. If equation 9 is not satisfied, then assign x to group G_2 [1, 4, 15].

Linear Discriminant Functions

Due to its simplicity and optimal properties, linear discriminant analysis is the most common method used in practice [17]. The concept of a linear discriminant function was first introduced by Fisher in 1936. The idea is based on the definition of a variable Y that maximally separates the G groups and is a linear combination of the variables X_i, $i = 1, \ldots, p$:

$$\begin{aligned} Y &= a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \\ Y &= \sum_{i=1}^{p} a_i X_i \end{aligned}$$

In order to maximize the separation between the groups, the coefficients, a_i, are determined by solving the following matrix system:

$$\begin{aligned} C \times A &= M \\ A &= C^{-1} M \end{aligned} \tag{10}$$

where:

M = column matrix of difference in means between groups, e.g. $\overline{G_1} - \overline{G_2}$

C = covariance matrix between the different groups, and

A = matrix of coefficients

If the two populations are normally distributed with equal covariance matrices, then the best classification rule is to classify sample x into the group G_1 if

$$\upsilon \quad = \quad M^T \times C^{-1} \left(X - \frac{1}{2}(\overline{G}_1 + \overline{G}_2) \right) \geq c \tag{11}$$

where $c = ln(\pi_2/\pi_1)$. Otherwise, classify sample x into group G_2 [7].

There are many other classification rules, the suitability of which depends on the data distributions, the cost associated to misclassification, and the goals of the classification rules (ranging from minimizing the number of samples misclassified to reducing the actual error rate for classifying future samples).

Perhaps the most important thing to keep in mind is that discriminant analysis is concerned with assigning samples to *pre-established* groups [4]. Unfortunately, in the geosciences these groups are rarely pre-determined and analysis is mainly concerned with identifying the different groups.

Cluster Analysis

At the start of any study, little is known about the data much less the population(s) from which they came. The main objective of cluster analysis is to identify groups within a set of data with n samples on which there are p-variate observations. Essentially the groupings will identify samples that are similar based on the p-variate observations and distinguish them from those that are dissimilar. Each sample then can only be assigned to one group only and is considered dissimilar to samples belonging to other groups.

Cluster analysis is not limited to the separation of data samples, we could also cluster variables so that highly correlated variables are grouped together so that some average of the variables could be used in analysis [4]. We have already seen this type of clustering in the dimension reducing methods shown above, and so the following discussion will focus on the clustering of samples.

Clustering of data samples can be achieved via several different approaches: process of agglomeration or division [13]. The process of agglomeration requires that at the start of analysis, each sample forms its own group of one. Groups that are close to each other are then combined to form a group, this is done until all the individual samples are placed into the m groups (where $m < n$). Alternatively, we could begin with only one group, to which all samples belong. Samples that are far apart are then divided, this continues until the required number of groups are formed and all samples have been accounted for. Chatfield & Collins (1980) provides schematic illustrations of clustering of samples based on two variables, x_1 and x_2, and is shown here as Figure 10.

If we continued to perform cluster analysis within the first set of groups and identify sub-groups within the groups and so on, then the result is a hierarchical clustering scheme. This scheme essentially breaks down the primary groups into secondary groupings, until

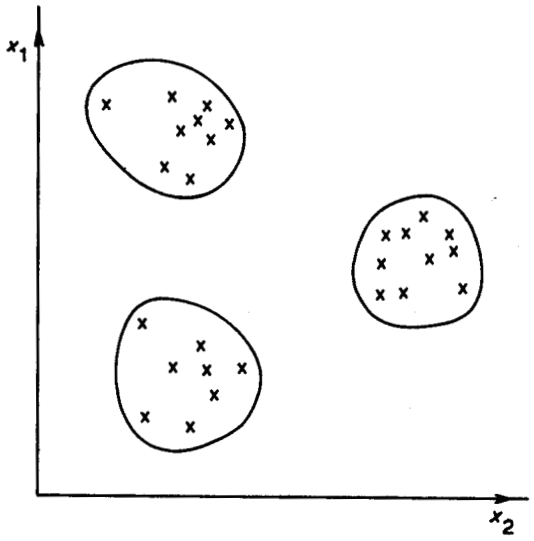

Figure 10: Clustering Data into Groups. Source: Chatfield & Collins(1980)

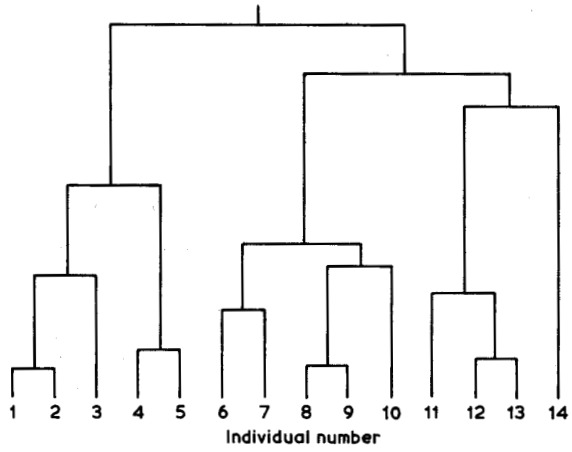

Figure 11: Hierarchical Clustering: Dendrogram or Hierarchical Tree. Source: Chatfield & Collins(1980)

eventually each sample is its own group which essentially amounts to clustering by division. Figure 11 shows a schematic illustration of a hierarchical tree (commonly referred to as a dendrogram), also taken from Chatfield & Collins (1980).

The decision to combine or divide groups is based on distance measures between the data and the group centres, which may be dependent on the means, variances and covariances of the two different populations. Metric distance measures are those which are strictly positive. The Euclidean metric distance function is a common measure and is given by:

$$d_{ij} = \sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \tag{12}$$

where d_{ij} is the distance between group i and group j, p is the number of variables, X_{ik} is the value of variable k for group i.

Another common distance measure is the Mahalanobis distance which accounts for correlations between the p variables, written in matrix form below:

$$d_{ij}^2 = (m_i - m_j)^T C^{-1} (m_i - m_j) \tag{13}$$

where m_i is the $p \times 1$ column vector of means for group i, and C is the $p \times p$ covariance matrix between the groups. The Mahalanobis distance has often been referred to as the generalized Euclidean distance [5]. Other distance measures include squared Euclidean, Chebyshev, Penrose, nearest neighbour and Minkowski distance [5, 6, 13, 16].

The main differences in the large number of cluster analysis algorithms lies in the choice of 1) the clustering process and 2) the distance measure applied.

General Comments on Classification Algorithms

Use of both discriminant and cluster analysis in geostatistics are likely limited to exploratory analysis tools applied prior to geostatistical simulation. They should provide insight into the potentially different populations found within the data set. Decisions of stationarity may be applicable to only the clusters or classes identified using these techniques.

Final Comments

In geostatistics, we frequently work with multivariate data sets. Conventional geostatistical tools make assumptions regarding data distributions, stationarity of population statistics, and linear correlation between variables. Unfortunately, the sample data rarely conform to all these assumptions and must be transformed in such a manner to allow for use of conventional geostatistics.

There are a multitude of multivariate statistical tools that can be useful in geostatistics, however in the past their application has been fairly limited. During the exploratory data analysis phase of a study, the group of classification techniques could be effective in distinguishing between different populations. Once the decision to use only a certain number of samples is made, we may wish to try to reduce the number of different variables for simulation. In this instance, the dimension reducing techniques would be most effective. However,

in the case where dimension reduction is not an issue, but non-linearity and/or mineralogical constraints limit the applicability of typical geostatistical tools, the data transformation methods may prove particularly useful.

The objective of this paper was to provide an overview of some of the available techniques, and is by no means meant to be an exhaustive list. The effective use of any of these statistical tools (and others not mentioned here) is completely dependent on the understanding of the objectives and the correct application of each method.