CHAPTER 2

THEORY ON PRINCIPAL COMPONENT

Most of our statistical techniques involve operations on a single response variable such as weight, pH, temperature, specific gravity, concentration, and the like. This is natural because one is usually interested in a problem involving a single response. However, there are a number of occasions where more than one response variable is of importance in a problem, and these variables should be studied collectively in order to take advantage of the information about the relationship among them. This is field of *multivariate analysis.* Most multivariate techniques are merely extensions of univariate techniques such as *t* tests or the analysis of variance.

2.1 Introduction

The first use of principal component analysis (PCA) in quality control was due to Jackson and Morris (1957). Basically, PCA consists of transform p correlated variables, x, into a net set of uncorrelated variables, y, called *principal component (pc's)* [6].

The method of principal components is due primarily to Hotelling [1933] although the original concept goes back to Karl Pearson [1901]. In way industrial applications, the pc's do have physical interpretation and can be used as control variables in their own right. The same generalized T^2- statistics may still be employed and in the case of an indication of an out-of-control situation, the diagnosis of this condition may be enhanced by virtue of the fact that the pc's are uncorrelated.

For this study, we use PCA method. PCA is concerned with explaining the variance-covariance structure of a set of variables through a few *linear combinations* of these variables. Its general objectives are (1) data reduction and (2) interpretation.

2.2 Principal Component Analysis

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

As we shall see, principal components depend solely on the covariance matrix Σ (or correlation matrix ρ) of $X_1, X_2, \ldots, X_p$. Their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the simple components when the population is multivariate normal.

Suppose $X = [X_1, X_2 \ldots X_p]$ is a p-dimensional random variable X as the vector. When row(m) of a data matrix X correspond to samples while columns(n) correspond to variables.

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & \cdots & X_{pp} \end{bmatrix}_{mxn}$$

The marginal means μ_i and variances σ^2_i are define as mean, expectation or expected value $\mu_i = E(X_i)$ and $\sigma^2_i = E(X_i - \mu_i)^2$, i = 1, 2, ... , respectively. Specifically,

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i)dx_i & \text{If } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \sum_{allx_i} x_i p_i(x_i) & \text{If } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases}$$

$$\sigma_i^2 = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i)dx_i & \text{If } X_i \text{ is a continuous random function } f_i(x_i) \\ \sum_{allx_i} (x_i - \mu_i)^2 p_i(x_i) & \text{If } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases}$$

The behavior of any pair of random variables, such as X_i, X_j , is described by their joint probability function, and a measure of the linear association between them is provided by the covariance

$$\sigma_{ik} = E[(X_i - \mu_i)(X_j - \mu_j)] \qquad (2\text{-}1)$$

$$= \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f_{ij}(x_i, x_j)dx_i dx_j & \text{if } X_i, X_j \text{ are continuous random variables with the joint density function } f_{ij}(x_i, x_j) \\ \sum_{allx_i} \sum_{allx_{jj}} (x_i - \mu_i)(x_j - \mu_j) p_{ij}(x_i, x_j) & \text{if } X_j, X_j \text{ are discrete random variables with the joint probability function } p_{ij}(x_j, x_j) \end{cases}$$

Similarly the expectation of a random matrix is the matrix of expected value of the random elements. For the generalization of the variance to multidimensional variates let define the covariance of the elements X_i and X_j of X as the *product moment* of those variates about their respective means;

$$
\begin{aligned}
Cov(X_i, X_j) &= E\{[X_i - E(X_i)][X_j - E(X_j)\} \\
&= E(X_iX_j) - [E(X_i)][E(X_j)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_i x_j f_{ij}(x_i, x_j)dx_i dx_j - [E(X_i)][E(X_j)] \\
&= \sigma_{ij} \\
&= \left[n\sum_{k=1}^{n} x_{ik}x_{jk} - \sum_{k=1}^{n} x_{ik} \bullet \sum_{k=1}^{n} x_{jk}\right] / [n(n-1)]
\end{aligned}
$$

where $f_{ij}(x_i, x_j)$ is the joint density of X_i and X_j. If $i=j$, the covariance is the variance of X_i, and we shall customarily write $\sigma_{ij} = \sigma^2_i$. The extension of the variance notion to the p-component random vector X is the matrix of variances and covariances

$$
\Sigma = Cov(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma_{pp} \end{bmatrix} \quad (2\text{-}2)
$$

We shall call this symmetric matrix the *covariance matrix* (Σ) of X. Where $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}$ are respective variances of the p variates and the remaining elements are covariances among them. When all of the diagonal elements of this matrix are unity, it is simply a matrix of correlation variables; otherwise, the correlation between the i^{th} and j^{th} characteristics is define to be

$$
\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \quad (2\text{-}3)
$$

for values of i and j ranging between 1 and p. Where ρ is called *correlation coefficient.*

The key idea from matrix algebra related to the method of principal components is that p x p symmetric, nonsingular matrix, such as the covariance

matrix Σ [3], may be reduced to a diagonal matrix λ by premultiplying and postmultiplying by a particular orthonormal matrix U such that

$$U'\Sigma U = \lambda \qquad (2\text{-}4)$$

The diagonal elements of λ, λ_1, λ_2, ..., λ_p, are called the *characteristic roots, latent roots or eigenvalues* of Σ. The columns of U, u_1, u_2, ..., u_p, are *called characteristic vectors, eigenvectors or loading vector* of Σ. The characteristic roots may be obtained from the following determinantal equation, called the *characteristic equation*

$$|\Sigma - \lambda I| = 0 \qquad (2\text{-}5)$$

where I is the identity matrix. This produces a p^{th} degree polynomial in λ from which the values λ_1, λ_2, ..., λ_p are obtained.

Consider the linear combinations [1]

$$\begin{aligned} Y_1 = a_1'X &= a_{11}X1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 = a_2'X &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ Y_p = a_p'X &= a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned} \qquad (2\text{-}6)$$

$$\begin{aligned} &Var(Y_i) = a_i'\Sigma a_i && i = 1, 2, \ldots, p \\ &Cov(Y_i, Y_k) = a_i'\Sigma a_k && i,k = 1, 2, \ldots, p \end{aligned}$$

The principal components are those correlated linear combination $Y_1, Y_2, \ldots, Y_p$ whose variances $Var(Yi) = a_i'\Sigma a_I$ are as large as possible. It is clear that $Var(Yi) = a_i'\Sigma a_i$ can be increased by multiplying any a_1 by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. We therefore define

First principal component = linear combination a_i'X that maximizes Var(a_I'X) subject to a_i'a_I = 1

Second principal component = linear combination a_2'X that maximizes Var(a_2'X) subject to a_2'a_2 = 1 and Cov (a_I'X, a_2'X) = 0

At the i^{th} step.

i^{th} principal component = linear combination a_i'X that maximizes Var(a_i'X) subject to a_i'a_i = 1 and Cov (a_i'X, a_k'X) = 0 for k < i

A principal component axis transformation will transform p correlated variables $x_1 \ldots x_p$ into p new uncorrelated variables $y_1, \ldots, y_p$, the coordinate axes of these new variables being described by the vectors u_i which make up the matrix U of direction cosines used in the following general transformation

$$Y = U'(x - \bar{x}) \qquad (2\text{-}7)$$

Here $\bar{x}$ and x are p x 1 vectors of the original variables and their means, respectively. The transformed variables are called the *principal component of* x or *score*. The i^{th} principal component would be

$$y_i = u_i'(x - \bar{x}) \qquad (2\text{-}8)$$

and will have mean zero and variance λ_i.

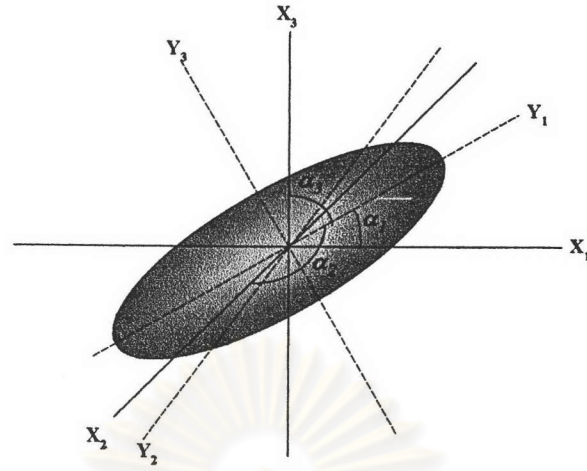

Figure 2.1 Principal axes of trivariate observation

The geometrical interpretation of components as the varieties corresponding to the principal axes of the scatter of the observations in space. Imagine that a sample of N trivariate observations has scatter plot show in Fig 2.1 [2] where the origin of the response axes has been taken at the sample means. The swarm of points seems to have a generally ellipsoidal shape, with a major axis Y_1 and less well defined minor axes Y_2 and Y_3. Let us confine out attention for the moment to major axis and denote its angles with the original response axes as α_1, α_2 and α_3. If Y_1 passes through the sample mean point, its orientation is completely determined by the direction cosines

$$u_{11} = \cos \alpha_1 \quad u_{21} = \cos \alpha_2 \quad u_{31} = \cos \alpha_3 \qquad (2\text{-}9)$$

where $u_{11}^2 + u_{21}^2 + u_{31}^2 = 1$.

It is known from analytic geometry that the value of the observation $[x_{i1}, x_{i2}, x_{i3}]$ on the new coordinate axis Y_1 will be

$$y_{i1} = u_{11}(x_{i1} - \bar{x}_1) + u_{21}(x_{i2} - \bar{x}_2) + u_{31}(x_{i3} - \bar{x}_3) \qquad (2\text{-}10)$$

2.3 Standardizing the Sample Principal Component

In general, sample principal components are not invariant with respect to changes in scale. As we mentioned in the treatment of population components, variables measured on different scales or on a common scale with widely differing ranges are often standardized. For the sample, standardization is accomplished by constructing the n x p data matrix of standardized observations

$$R = \begin{bmatrix} z_{11} & z_{12} & \cdots & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & \cdots & z_{pp} \end{bmatrix} = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sigma_{11}}} & \frac{x_{12}-\bar{x}_1}{\sqrt{\sigma_{11}}} & \cdots & \cdots & \frac{x_{1p}-\bar{x}_1}{\sqrt{\sigma_{11}}} \\ \frac{x_{21}-\bar{x}_2}{\sqrt{\sigma_{22}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{\sigma_{22}}} & \cdots & \cdots & \frac{x_{2p}-\bar{x}_2}{\sqrt{\sigma_{22}}} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_p}{\sqrt{\sigma_{pp}}} & \frac{x_{n2}-\bar{x}_p}{\sqrt{\sigma_{pp}}} & \cdots & \cdots & \frac{x_{np}-\bar{x}_p}{\sqrt{\sigma_{pp}}} \end{bmatrix} \quad (2\text{-}11)$$

yields the sample mean vector

$$\bar{R} = \frac{1}{n}\begin{bmatrix} \sum_{j=1}^{n} \frac{x_{j1}-\bar{x}_1}{\sqrt{\sigma_{11}}} \\ \sum_{j=1}^{n} \frac{x_{j2}-\bar{x}_2}{\sqrt{\sigma_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp}-\bar{x}_p}{\sqrt{\sigma_{pp}}} \end{bmatrix} = 0 \quad (2\text{-}12)$$

and sample covariance matrix

$$S = \frac{1}{n-1}Z'Z = \frac{1}{n-1}\begin{bmatrix} \frac{(n-1)\sigma_{11}}{\sigma_{11}} & \frac{(n-1)\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \cdots & \frac{(n-1)\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{(n-1)\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{(n-1)\sigma_{22}}{\sigma_{22}} & \cdots & \cdots & \frac{(n-1)\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{(n-1)\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{(n-1)\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \cdots & \frac{(n-1)\sigma_{pp}}{\sigma_{pp}} \end{bmatrix} \quad (2\text{-}13)$$

In the other hand, principal component can be calculated from *correlation matrix* from above definition. Generally, extracting components from S rather than R remains closer to the spirit and intent of principal component analysis, especially if the components are to be used in further computations. However, in some cases, the principal components will be more interpretable if R is used. For example, if the variances differ widely or if the measurement units are not commensurate, the components of S will be dominated by the variables with large variances. The other variables will contribute very little. For a more balanced representation in such cases, components of R may be used.

As with any changed of scale, when the variables are standardized in transforming to R, the shape of swarm of points will change. (Note, however, forming to R, any further changes of scale on the variables would not affect the components because changes of scale do not change R.)

To illustrate how the eigenvalues and eigenvectors change when converting from S to R, we use a simple bivariate example in which one variance is substantially larger than the other [7]. Suppose that S and the corresponding R have the values

$$S = \begin{bmatrix} 1 & 4 \\ 4 & 25 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

The eigenvalues and eigenvectors from S are

$$\lambda_1 = 25.65 \qquad a_1' = (0.160, 0.987)$$
$$\lambda_2 = 0.35 \qquad a_2' = (0.987, -0.160)$$

The pattern we see in λ_1, λ_2, a_1, a_2 are quite predictable. The symmetry in a_1 and a_2 is due to their orthogonality, $a_1'a_2 = 0$. The large variance of x_2 in S ensures that the first principal component $y_1 = 0.160x_1 + 0.987x_2$ weights x_2 heavily. Thus the first principal component y_1 essentially duplicates x_2 and does not reflect the mutual

effect of x_1 and x_2. It is expected also that y_1 would account for virtually all of the total variance:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{25.65}{26} = 0.9865$$

The eigenvalues and eigenvector of R are

$\lambda_1 = 1.8$ $\quad a_1' = (0.707, 0.707)$

$\lambda_2 = 0.2$ $\quad a_2' = (0.707, -0.707)$

The first principal component of R,

$$y_1 = 0.707\frac{x_1 - \bar{x}_1}{1} + 0.707\frac{x_2 - \bar{x}_2}{51}$$

accounts for a high proportion of variances,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.8}{2} = 0.9,$$

Because the variables are fairly highly correlated(r=0.8). But the standardized variables $(x_1 - \bar{x}_1)/1$ and $(x_2 - \bar{x}_2)/5$ are equally weighted in y_1, due to the equality of the diagonal elements ("variances") of R.

We now list some general comparisons of principal components from R with those from S:

1. The present of variance accounted for by the components of R will differ from the percent for S.
2. The coefficients of the principal components from R differ from those obtained from S.

3. If we express the components from R in terms of the original variables, they still will not agree with the components from S. By transforming the standardized variables back to the original variables, the component R become

$$y_1 = 0.707\frac{x_1 - \bar{x}_1}{1} + 0.707\frac{x_2 - \bar{x}_2}{51}$$

$$= 0.707x1 + 0.141x2 + \text{const.}$$

$$y_1 = 0.707\frac{x_1 - \bar{x}_1}{1} - 0.707\frac{x_2 - \bar{x}_2}{51}$$

$$= 0.707x1\text{-}0.141x2 + \text{const.}$$

As expected, these are very different from the components extracted directly from S. This problem arises, of course, because of the lack of scale invariance of the components of S.

4. The components from a given matrix R are not unique to that R. For example, in the bivariate case, the eigenvalues of

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

are given by

$$\lambda_1 = 1 + r \qquad \lambda_2 = 1 - r,$$

but the components remain the same for all value of r;

$$y_1 = 0.707\frac{x_1 - \bar{x}_1}{\sigma_1} + 0.707\frac{x_2 - \bar{x}_2}{\sigma_2}$$

$$y_1 = 0.707\frac{x_1 - \bar{x}_1}{\sigma_1} - 0.707\frac{x_2 - \bar{x}_2}{\sigma_2}$$

In general form of principal component is

$$y_{i1} = a_{11}\frac{(x_{i1} - \bar{x}_1)}{\sigma_1} + a_{21}\frac{(x_{i2} - \bar{x}_2)}{\sigma_2} + a_{31}\frac{(x_{i3} - \bar{x}_3)}{\sigma_3} \qquad (2.14)$$

The component above does not depend on r. For example, they serve equally well for r = 0.01 and for r = 0.99. For r = 0.01, the proportion of variance explained by y1 is $\lambda 1/(\lambda 1+\lambda 2)$ = (1+0.01)/(1+0.01+1-0.01)=1.01/2=0.505. For r = 0.99, the ratio is 1.99/2 = 0.995. Thus the statement that the first component from a correlation matrix accounts for, say, 90% of the variance is not meaningful. In general, for p>2, the components from R deepened only on the ratios (relative values) of the correlations, not on their actual values, and components of a given R matrix will serve for other R matrices.

2.4 Principal Component Scores

To use the principal component variables in ensuing statistical analyses, it is necessary to compare principal component scores (values of the principal component variables) for each experimental unit in the data set. These scores provide the locations of the observations in a data set with respect to its principal component axes.

Let x_r be the vector of measured variables for the rth experimental units. Then the value (score) of the rth principal component variable for the rth experimental unit is $y_{rj} = a'_j(x_r - \bar{x})$, for $j = 1, 2, \ldots, p$ and $r = 1, 2, \ldots, N$.

2.5 Component Loading Vectors

Note that the eigenvectors of variance-covariance matrix that are being used to define the principal components are normalized to have length 1, that is, $a'_j a_j = 1$, for j =1, 2, …, p. This can sometimes be confusing when we are trying to interpret the principal components by examining the elements in the eigenvectors that define the principal components. A large element in one eigenvector may or may not be large in another eigenvector. That is, elements within an eigenvectors are not comparable.

This is because the eigenvectors are normalized to have length 1, which requires that the sum of the squares of the elements in each vector must equal 1. Thus the more elements in a single vector that are actually different from 0, the smaller each element must be. For example, if eight elements in a vector were nonzero and if all were of similar magnitude, they would each have a value equal to $\pm 1/\sqrt{8} = \pm 0.3536$; but if only two elements in a vector were nonzero and were of the same magnitude, they would each have a value equal to $\pm 1/\sqrt{2} = \pm 0.7071$.

To make comparisons between eigenvectors, many researchers scale the eigenvectors by multiplying the elements in each vector by the square root of the corresponding eigenvalue. Let $c_j = \lambda_j^{1/2} a_j$, for j = 1, 2, ..., p. These new vectors are called component loading vectors. The elements in the vector c_j are called component loadings and are scaled so that the yare generally larger than those of the less important components. The c_j's are still eigenvectors of Σ, but they have lengths equal to $\sqrt{\lambda_j}$ rather than length 1. All of the elements in all of the c_j's are comparable to one another. The ith element in c_j gives the covariance between the ith original variable and the jth principal component.

2.6 The Number of Principal Components

There is always the question of how many components retain. There is no definitive answer to this question. Things to consider include the amount of total sample variance explained, the relative sizes of the eigenvalues (the variances of the sample components), and the subject-matter interpretations of the components. In addition, a component associated with an eigenvalue near zero and, hence, deemed unimportant, may indicate an unsuspected linear dependency in the data.

A useful visual aid to determining an appropriate number of principal components is a *scree plot.* With the eigenvalues ordered from largest to smallest, a scree plot is a plot of l_l versus p –the magnitude of an eigenvalues versus its number of variables. To

determine the appropriate number of components, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size. Fig 2.2 shows a scree plot for the situation with six components. An elbow occurs in the plot in Fig 2.2 at about $p = 3$. That is, the eigenvalues after l_2 are all relatively small and about the same size. In this case, it appears without any other evidence, that two (or perhaps three) sample principal component effectively summarizes the total sample variance.

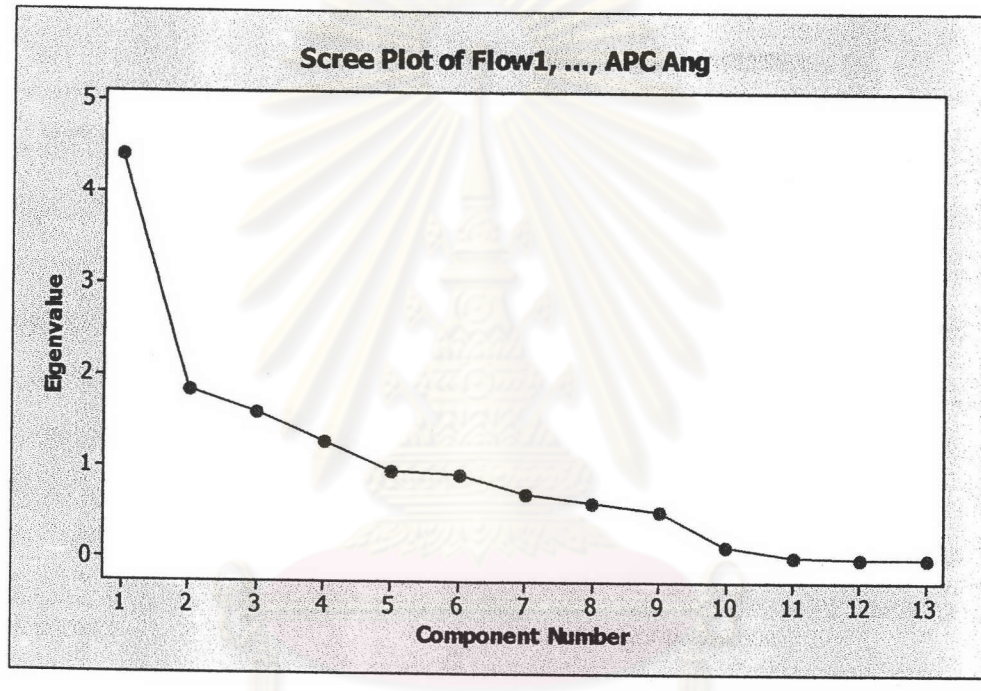

Figure 2.2 Scree Plot

2.7 Residual Analysis

How do we know our model fit? Assuming that the model is "correct", we have used the estimated regression function to make inferences. Of course, it is imperative to examine the adequacy of the model before the estimated function becomes a permanent part of the decision-making apparatus.

All the sample information on lack of fit is contained in the residuals

$$\varepsilon = Y - Z\beta \qquad (2\text{-}15)$$

or $$y_i = u_i' (x - \bar{x}) + \varepsilon \quad (2\text{-}16)$$

where

ε = Error

Y = Response

Z = Predictor variable

β = Constant

Properties of error term are assume as following

a) $E(\varepsilon_j) = 0$;

b) $Var(\varepsilon_j) = \sigma^2$ (constant)

c) $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$

Or we can calculate a lack of model fit statistic, Q, for each sample [4]. Q is simply the sum of sqaure of each row (sample) of ε for example, for the ith sample in X, x_i :

$$\varepsilon_i = x_i^*(I - PP') \quad (2\text{-}17)$$

$$Q_i = \varepsilon_i \varepsilon_i' \quad (2\text{-}18)$$

Where

I = Identity matrix (variable by variable)

P_p = the matrix of the first p eigen vectors retained in the PCA model.

The Q statistic indicates how well each sample confirms to PCA model. It is a measure of the amount of variation in each sample not captured by the p principal components retained in the model.

The critical value for residuals is

$$Q_\alpha = \theta_1\left[\frac{c_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2} + 1\right] \tag{2.19}$$

where

c_α = Normal deviative cutting off an area of α under the upper tail of distribution if h_0 is positive and under lower tail if h_0 is negative.

$$= \frac{\left[\left(\frac{Q}{\theta_1}\right)^{h_0} - \frac{\theta_2 h_0(h_0-1)}{\theta_1^2} - 1\right]}{\sqrt{2\theta_2 h_0^2}} \tag{2.20}$$

$$\theta_1 = \sum_{i=k+1}^{p}\lambda_i$$

$$\theta_2 = \sum_{i=k+1}^{p}\lambda_i^2$$

$$\theta_3 = \sum_{i=k+1}^{p}\lambda_i^3$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$$

2.8 Monitoring Quality with Principal Component

To improve the quality of goods and services, data need to be examined for causes of variation. When a manufacturing process is continuously producing items or when we are monitoring activity of a service, data should be collected to evaluate the capabilities and stability of the process. When a process is stable, the variation is produced by common causes that are always present, and no one cause is a major source of variation.

The purpose of any control chart is to identify occurrences of *special causes* of variation that come from outside of usual process. These causes of variation often indicate a need for a timely repair, but they can also suggest improvements to the process. Control charts make the variation visible and allow one to distinguish common from special causes of variation.

A control chart typically consists of data plotted in time order and horizontal lines [Fig 2.3], called *control limits* that indicate the amount of variation due to common cause.

2.8.1. *X*-bar chart is a useful chart. To create an $\overline{X}$ chart :

i. Plot the individual observation or sample means in time order.

ii. Create and plot the centerline $\overline{\overline{x}}$, the sample mean of all the observations.

iii. Calculate and plot the control limits given by

Upper control limit (UCL) = $\overline{\overline{x}}$ +3(standard deviation)

Lower control limit (LCL) = $\overline{\overline{x}}$ -3(standard deviation)

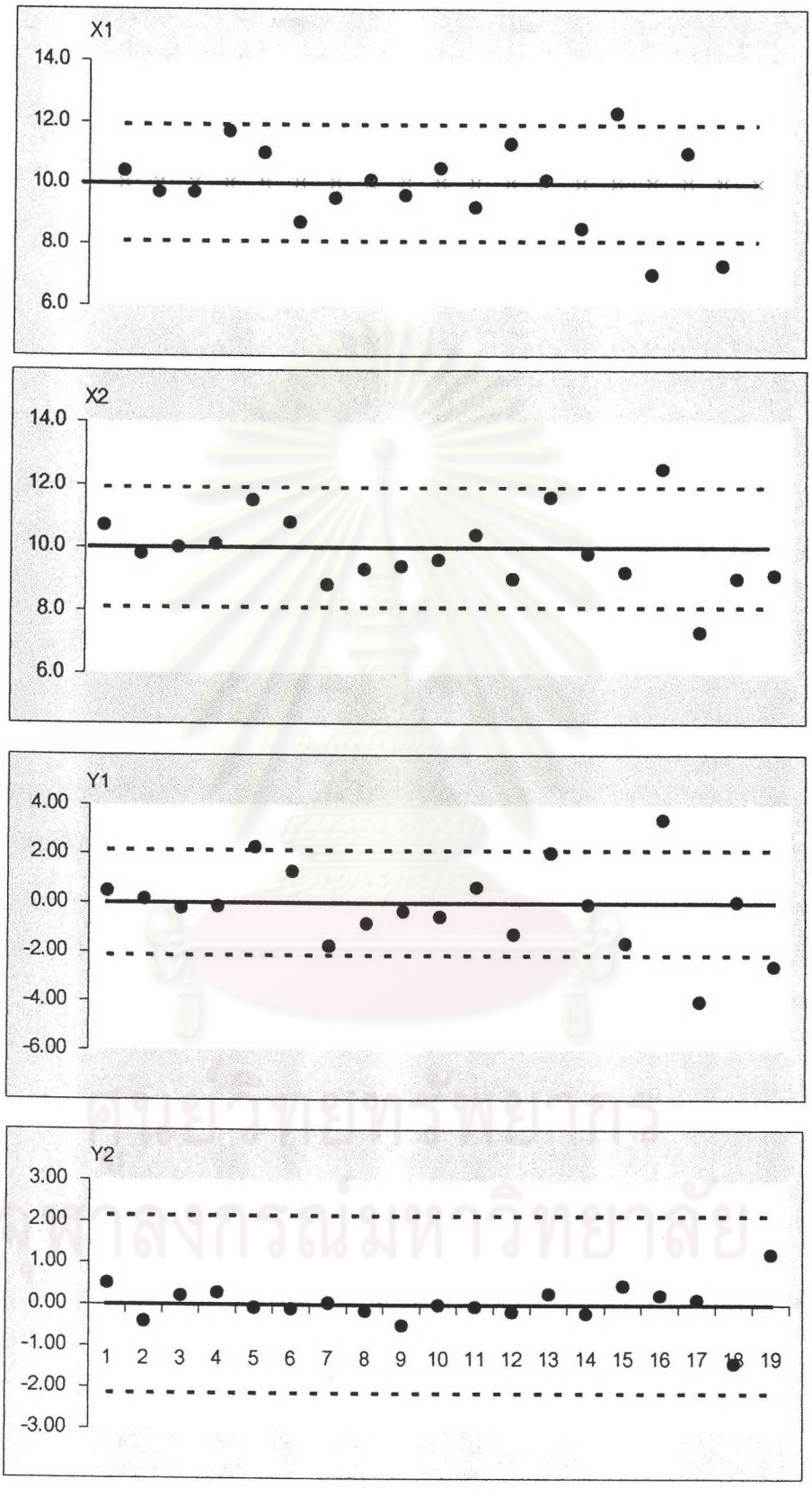

Figure 2.3 95% Control Charts: Original (X1, X2) and Principal Component (Y1, Y2)

2.8.2. Ellipse format chart – for a bivariate control region is the more intuitive of the charts, but its approach is limited to two variables. The two characteristics on the jth unit are plotted as a pair (x_{i1}, x_{i2}). The 95% ellipse consists of all x that satisfy

$$(x-\bar{x})' S^{-1}(x-\bar{x}) \le \chi_2^2(.05) \qquad (2\text{-}21)$$

$$\frac{s_{11}s_{22}}{s_{11}s_{22}-s_{12}^2}\left[\frac{(x_1-\bar{x}_1)^2}{s_{11}} - 2s_{12}\frac{(x_1-\bar{x}_1)(x_2-\bar{x}_2)}{s_{11}s_{22}} + \frac{(x_2-\bar{x}_2)^2}{s_{22}}\right] = T_{2,n,\alpha}^2 \qquad (2\text{-}22)$$

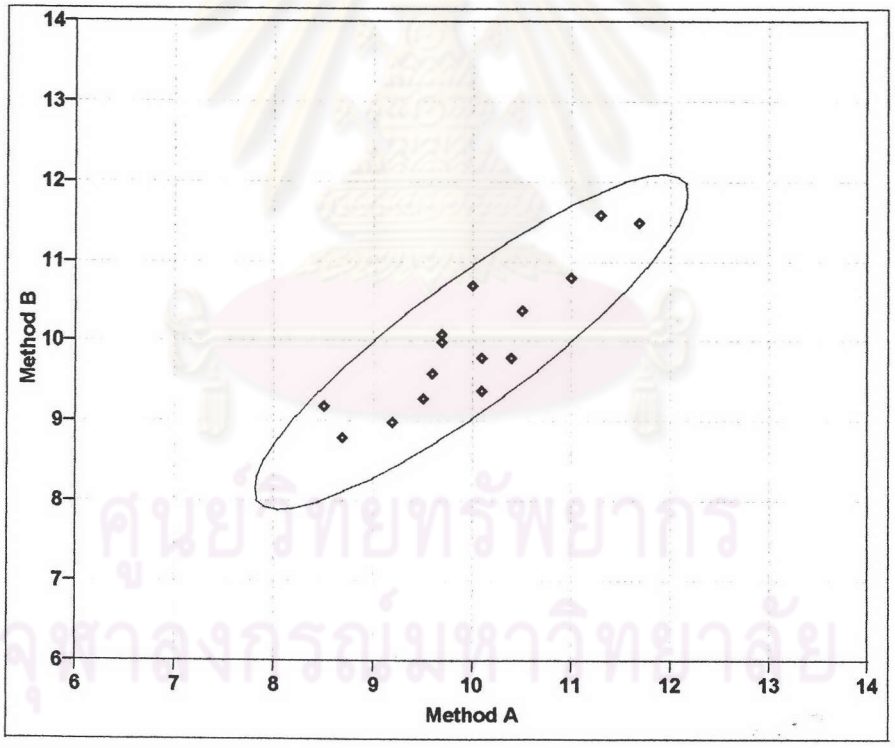

Figure 2.4 the quality control ellipse chart

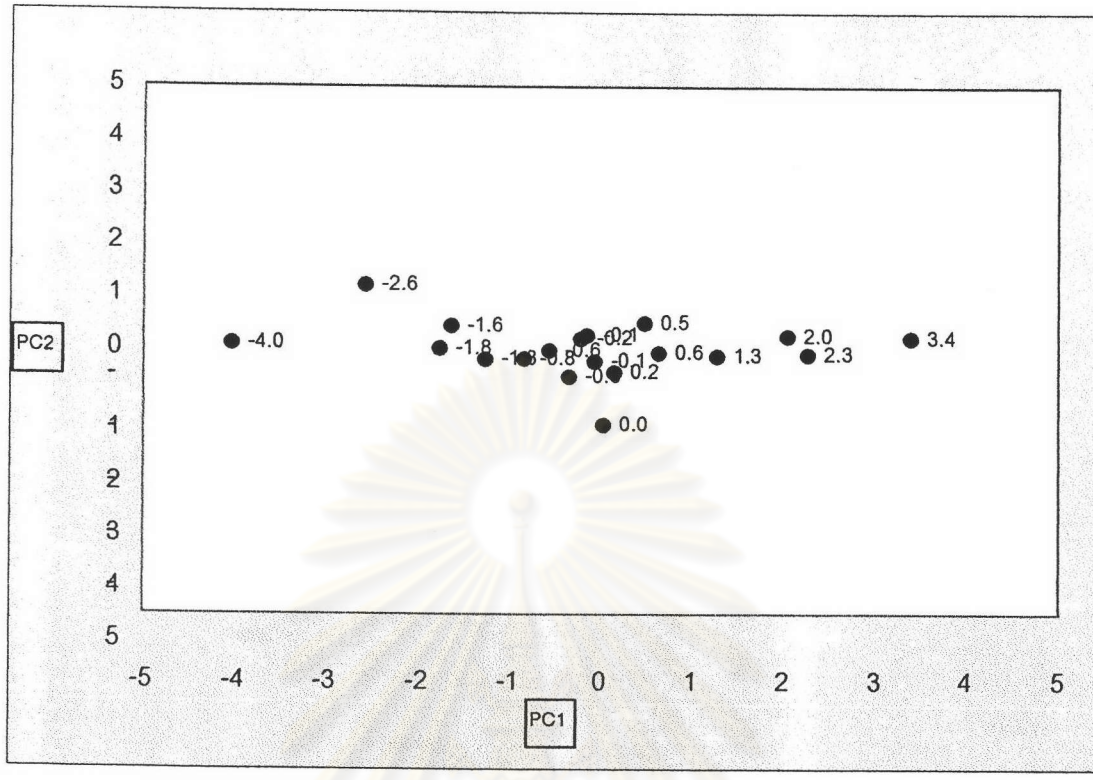

Figure 2.5 Principal Component Model Plot

2.8.3. T^2-Hotelling – Hotelling, in his paper "Multivariate Quality Control," was the first to consider the problem of analyzing a correlated set of variables in his analysis of bombsite data. He developed a control procedure based on a concept referred to as statistical distance, a generalization of the T statistic. The statistic was later named Hotelling's T^2 in his honor. About the same time, Mahalanobis developed a similar statistical distance that is referred to as Mahalanobis' distance. The two statistical distances differ by a constant. The T² statistic has emerged as an extremely useful metric for multivariate process control. The T^2 is the procedure used to construct multivariate control charts in most software packages such as QualStat. The following sections provide a basic primer on the fundamental theory behind multivariate SPC.

Another method for examining information provided in a multivariate observation is to re-express the vector as a single univariate

statistic. There are numerous procedures for achieving this result. Below, two different procedures are discussed. Regardless of how what statistical method is used, the statistic must contain all the information provided by the all the variables (assuming it uses all variables) and in some cases, can be interpreted and used in making decisions as to the status of a process.

Consider a process that generates an uncorrelated bivariate observation (x1 and x2). To represent them graphically, it is common to construct a two-dimensional scatter plot of the points. In addition, suppose there is interest in calculating the distance a particular point is from the mean point (or any other point). The distance between two points is always measured as a single number or value. This is true regardless of how many dimensions (variables) are involved in the problem.

The usual straight-line or Euclidean distance between two points is measured by the number of units that separate them. The squared straight-line distance between a point (x_1, x_2) and the mean point (μ_1, μ_2) is given by:

$$(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 = D^2 \qquad (2.23)$$

Note, we have taken the bivariate observation (x1, x2), and converted it to a single number D, the distance the observation is from the mean point.

It is of interest to note that if the distance from the mean vector is fixed, then all points that are the same distance from the mean can be represented by a circle with the center at the mean vector and a radius of D. In addition, any point located inside the circle has a distance to the mean point less than D. See the figure below.

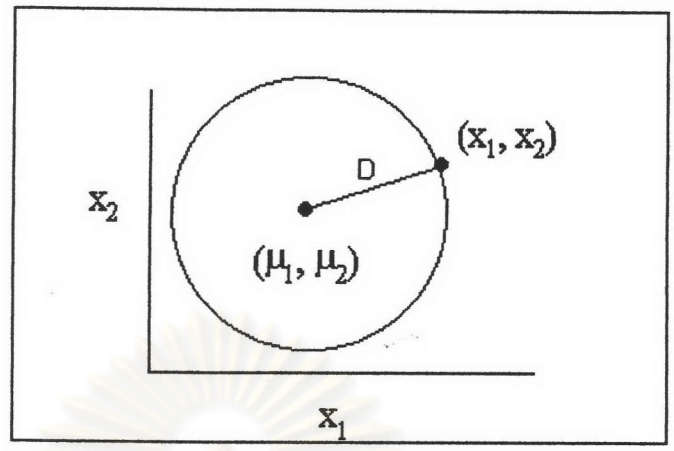

Fig 2.6 Region of Same Distance

This particular measure of straight-line distance is unsatisfactory for most statistical work because it assumes that each coordinate contributes equally in calculating the distance from the centroid with no consideration given to the variation or the scale differences between the variables. To correct this, consider the formula for the standardized values for each of the variables:

$$\frac{(x_1 - \mu_1)}{\sigma_1} \text{ and } \frac{(x - \mu_2)}{\sigma_2} \qquad (2.24)$$

and all points satisfying the relationship:

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_1^2} = SD^2 \qquad (2.25)$$

This particular measure is known as statistical distance. All points satisfying 2.23 are said to have the same statistical distance from the mean point. The Fig2.7 is an ellipse and is presented below.

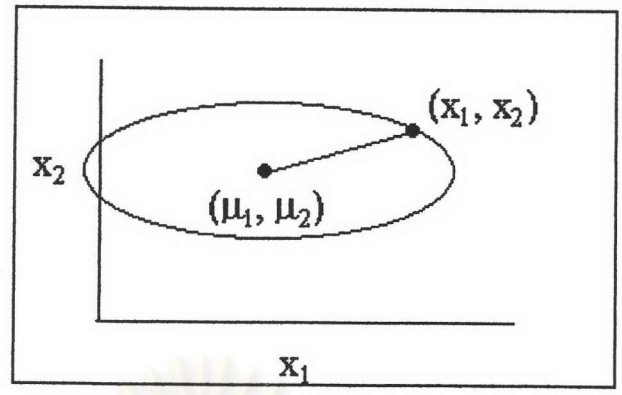

Fig 2.7 Region of Same Statistical Distance

Again, note any point inside the ellipse will have a statistical distance less than SD. Likewise, any point located outside the ellipse will have a statistical distance greater than SD.

Upon comparing the statistical distance to the straight-line distance there are some major differences. First, the statistical distance is standardized and therefore, there are no scales involved. Since the variables in a multivariate process may be measured in many different units, this removes the effects of varying scales and units. Second, from the previous figure, it is obvious that two points can have the same statistical distance but different Euclidean or straight-line distances from the mean vector. If the variances of the two variables are equal, then the statistical distance and Euclidean distance are the same.

The major difference between statistical and Euclidean distance lies in the fact that the two variables used in calculating the statistical distance are weighted inversely by their standard deviations, while both variables are equally weighted in the straight-line Euclidean distance. Thus, a variable with small variation will contribute more to the statistical distance than a variable with large variation. In other words, statistical distance is a weighted straight-line distance where

more importance is placed on the variable with less variation to compensate for its size relative to its mean.

Consider a scatter plot of two positively correlated variables as represented in the figure presented below.

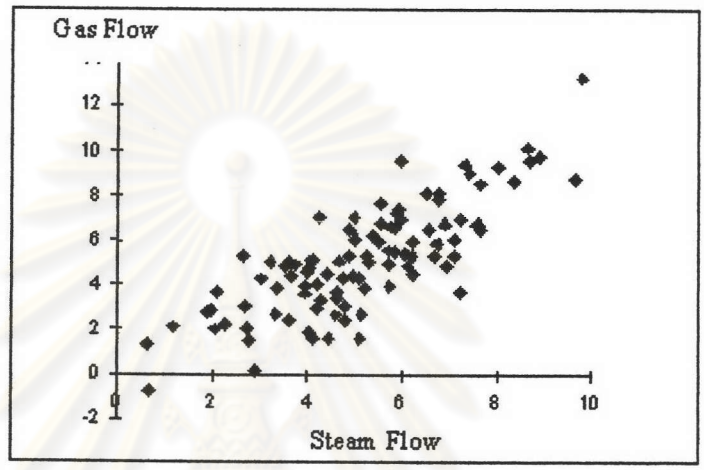

Fig 2.8 Scatter Plot of Correlated Variables

From analytical geometry, the equation of the ellipse encompassing all the points is of the form:

$$a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = SD^2 \qquad (2.26)$$

In this equation, the major and minor axes of the ellipse do not have to be parallel to the axis (x1, x2).

The statistical distance from the mean point that encompasses all the points can be represented as an ellipse. This concept is presented in the figure below.

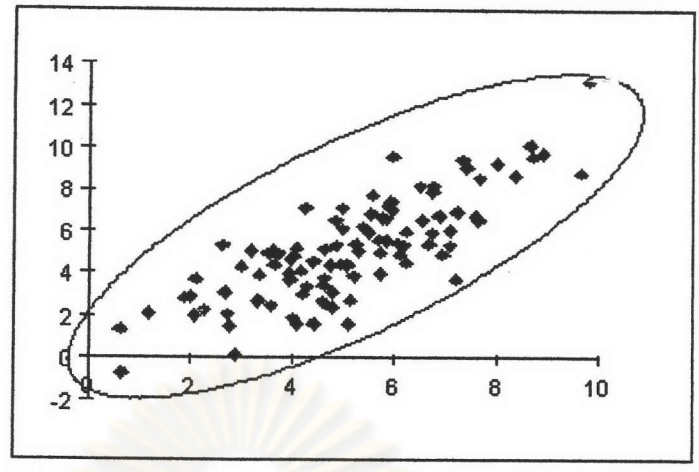

Fig 2.9 Encompassing Statistical Distance

One assumption fundamental to using Hotelling's T^2 to describe the behavior of statistical distance is the observation vector must follow a multivariate normal distribution. Under this assumption, i.e., (x1, x2) can be described jointly as a bivariate normal. The explicit form of 2.24 then becomes:

$$\frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1}-2\rho\left(\frac{(x_2-\mu_1)}{\sigma_2}\right)\left(\frac{(x_2-\mu_3)}{\sigma_3}\right)+\frac{(x_2-\mu_2)}{\sigma_2^2}\right]=SD^2 \qquad (2.27)$$

where r represents the correlation between the two variables (x1,x2). The product term between x1 and x2 accounts for the fact that the variables vary together and do not behave independently of one another. Also, note the absence of as product term in 2.23. When x1 and x2 are correlated, the major and minor axes of the ellipse differ from that of the variable space (x1, x2). If the correlation is positive, the ellipse will tilt upward as is evident in the figure above, whereas a negative correlation will tilt the ellipse in a downward direction. Using matrix notation, 2.25 can be expressed as follows:

$$(X-\mu)'\Sigma^{-1}(X-\mu)=SD^2 \qquad (2.28)$$

where X' = (x1,x2), μ' = (μ1, μ2), and is the inverse of the covariance matrix :

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \qquad (2.29)$$

where 12 = 21 and represents the covariance between x1 and x2. A similar expression can be written for many variables in this form. This is the quadratic form of the vector (X - μ) that represents the statistical distance. Therefore, 2.28 is known as Hotelling's T^2 statistic:

$$T^2 = (X - \mu)' \Sigma^{-1} (X - \mu) = SD^2 \qquad (2.30)$$

2.9 Interpreting Principal Components

Since principal components are linear combinations of the original variables, it is often necessary to interpret or provide a meaning to the linear combination. One can use the loadings for interpreting the principal components. The higher the loading of a variable, the more influence it has in the formation of the principal component score and vice versa. Therefore, one can use the loadings to determine which variables are influential in the formation of principal component. But, what do we mean influential? How high should the loading be before we can say that a given variable is influential in the formation of a principal component score? Unfortunately, there are no guidelines to help us in establishing how high is high. Traditionally, researchers have used a loading of 0.5 or above as the cutoff point.