Multivariate Normal Distribution and Confidence Ellipses

Multivariate statistics is largely built upon a straightforward extension of the Normal Distribution seen in Introductory Biostatistics. The classic formula for the Normal Distribution looks like this:

\[ f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

where \( f(x) \) refers to the probability density function (as accessed by \( \text{dnorm()} \) in R), \( \mu \) is the parameter for population mean, and \( \sigma^2 \) is the population variance. In this equation, the multiplied term: \( \frac{1}{\sqrt{2\pi \sigma^2}} \) may be viewed as a scaling factor, and the term in the exponent of \( e \): \( \left( \frac{x - \mu}{\sigma} \right)^2 \) doing much of the work in shaping the Normal Distribution's familiar "bell curve". The latter term can be interpreted as a description of squared distance \( (x - \mu) \) between some value of \( x \) who's probability is being assessed (along the \( x \) axis), and \( \mu \) the center of the probability density distribution, "standardized" by the distribution's known variance \( \sigma^2 \). It should be noted that \( f(x) \) only depends on this single scalar range variable \( x \), and as such, is "one-dimensional".

The Multivariate Normal Distribution now extends this idea of a probability density function into a number \( p \) of multiple directions \( x_1, x_2, \ldots, x_p \). To do so, the single mean of the distribution \( \mu \) becomes a vector \( \mu \) specifying means for each of \( x_1, x_2, \ldots, x_p \), and interpreted as a point in \( p \)-dimensional space. In addition, \( \sigma^2 \) is replaced by covariance matrix \( \Sigma \). This square matrix, by definition, contains variance for each of the \( x_1, x_2, \ldots, x_p \) variables along the main diagonal along with covariances between each pair of \( x \) variables interpreted as indicating degree of "collinearity" of these variables in \( p \)-dimensional space. As a result of these modifications, the probability density function \( p(x) \), still a scalar value, now becomes:

\[ f(x) = \frac{1}{\sqrt{p \cdot (2\pi)^2 \cdot \text{det}(\Sigma)}} e^{-\frac{1}{2} \left( x - \mu \right) \Sigma^{-1} \left( x - \mu \right)^T} \]

In this formulation, the exponent part \( \left( x - \mu \right) \Sigma^{-1} \left( x - \mu \right)^T \) is squared Mahalanobis distance, the extension of the concept of squared standardized distance \( \left( \frac{x - \mu}{\sigma} \right)^2 \) seen in the univariate case. As described in MTB 040, the matrix \( \Sigma^{-1} \) can be viewed as "correcting" the multivariate \( p \)-dimensional space for all covariances between the \( x \) variables, yielding a space of statistically equivalent distances.

In interpreting the multivariate density function, it is important to remember that \( f(x) \) is still a scalar value of probability that may be assigned to a position in \( p \)-dimensional space indicated by column vector \( x = (x_1, x_2, \ldots, x_p)^T \). If \( p = 2 \) dimensions, the familiar "bell curve" is now a peak in the \( z \) direction along a plane defined by \( x_1 \) and \( x_2 \), and any slice in the \( z \) direction along a diameter of the peak yields the "bell curve". For \( p > 2 \) the situation becomes harder to visualize, but the geometric concepts and mathematics remain identical.
As in univariate statistics, the Multivariate Normal Distribution, designated $N_p(\mu, \Sigma)$, has wonderfully useful properties, and is often invoked as an assumption in multivariate statistical tests. However, assessing the validity of this assumption in real data, is typically much more difficult. For one thing, a collection of data of useful size, typically yields a "sparse" distribution of points in $p$-dimensional space that is typically hard to visualize. Moreover, covariance (collinearity) between variables in these $p$ dimensions greatly complicates interpretation of this non-orthogonal space itself. However, in defense of multivariate statistics, the wealth of information embedded especially in the off diagonal elements in $\Sigma$ allows for greatly expanded field of inquiry with typically greater precision.

For extended discussion of properties assignable to the Multivariate Normal Distribution, see RA Johnson & DW Wichern (JW) *Applied Multivariate Statistical Analysis 4th Edition* 1998, and especially AC. Rencher (AR) *Methods of Multivariate Analysis* 1995. Following both, useful properties may be summarized as follows:

Linear combinations of variables with Multivariate Normal Distribution are Normally distributed.

All subsets of Multivariate Normally distributed variables are (Multivariate) Normally distributed.

However, the reverse is not necessarily true: individual or sets of variables may be (Multivariate) Normally distributed, but this doesn't ensure that the whole ensemble of variables is Multivariate Normally distributed.

Variables $x_1$, $x_2$, ... $x_p$ are independent if their covariances in $\Sigma$ are 0.

Conditional distributions of Multivariate Normally distributed variables are (Multivariate) Normally distributed.

If data are Multivariate Normally distributed then:

Sample mean vector $\bar{X}$ is a sufficient statistic for population mean $\mu$ and is distributed $N_p(\mu, (1/n)\Sigma)$.

Sample covariance matrix $S$ is a sufficient statistic for population covariance $\Sigma$ with $(n-1)S$ is a distributed according to the Wishart random matrix distribution with n-1 degrees of freedom.

$\bar{X}$ and $S$ are independent.

Squared Mahalanobis distance $D_M = (x - \mu)^T \Sigma^{-1} (x - \mu)$ is distributed according to the chi-square ($\chi^2$) distribution with $p$ degrees of freedom.

Even if the data is NOT Multivariate Normally distributed, an extension of the Central Limit Theorem applies: With large enough sample size, sample statistics $\bar{X}$, $S$ and $D_M$ still work as reasonable approximations.

What follows here is an examination of simulated bivariate data ($p=2$) to get an sense of what the Multivariate Normal Distribution looks like in reality, and the use of confidence ellipses based on the $\chi^2$ statistical distribution for $D_M$ in characterizing the Multivariate Normal Distribution.

**Generating Simulated Data in R:**

```
#MULTIVARIATE NORMAL DISTRIBUTION AND CONFIDENCE ELLIPSE:
# CREATE DATA USING RANDOM NUMBER GENERATOR:
X1=rnorm(1000,0,2)
X2=rnorm(1000,0,1)
X=cbind(X1,X2)
#WRITE DATASET:
write.table(X,file="c:/DATA/Multivariate/randombivariate.txt")
```
Reading Data:

Data is written to disk and read by MathCAD here.

\[
M := \text{READPRN}("c:/DATA/Multivariate/randombivariate.txt")
\]

\[
n := \text{rows}(M) = (1000)
\]

\[
p := \text{cols}(M) = (2)
\]

\[
i := 1 \ldots n \quad j := 1 \ldots p \quad \text{<index variables>}
\]

\[
lvec_i := 1
\]

\[
I := \text{identity}(n)
\]

Mean Vector:

\[
M_{bar} := \frac{1}{n} \cdot M^T \cdot lvec
\]

\[
M_{bar} = \begin{pmatrix}
-0.0327573 \\
0.0268316
\end{pmatrix} < \text{close to (0,0)}
\]

Covariance Matrix:

\[
S := \frac{1}{n-1} \cdot M^T \cdot \left(I - \frac{1}{n} \cdot lvec_lvec^T\right) \cdot M
\]

\[
S = \begin{pmatrix}
4.0040869 & 0.0113922 \\
0.0113922 & 1.0830404
\end{pmatrix}
\]

\[
\text{sd} := \begin{pmatrix}
\sqrt{S_{1,1}} \\
\sqrt{S_{2,2}}
\end{pmatrix}
\]

\[
\text{sd} = \begin{pmatrix}
2.0010215 \\
1.0406923
\end{pmatrix} \quad \text{<close to (2,1) as originally specified>
\]

\[\text{covariance between X1 & X2 is essentially 0. Variances for X1 & X2 occur along the main diagonal}\]
Confidence ellipses based on (JW eq. 4-8):

Making a set of ellipse points:
\[ \alpha := 0.05 \quad \text{df} := 2 \quad < \text{Set } \alpha \text{ as desired} \]
\[ c := \sqrt{\chi^2(1 - \alpha, \text{df})} \quad c = 2.448 \quad < \text{radius of circle} \]

Constructing the points:
\[ j := 1 \ldots 100 \quad \theta_j := \frac{j}{100} \cdot 2\pi \]
\[ x_{1j} := c \cdot \cos(\theta_j) \quad x_{2j} := c \cdot \sin(\theta_j) \]
\[ x := \text{augment}(x_{1}, x_{2}) \quad < \text{points on a circle of radius } c \]

Constructing Confidence Ellipse:
\[ \lambda := \text{eigenvals}(S) \quad \varepsilon := \text{eigenvects}(S) \quad < \text{Eigenvalues and Eigenvectors of } S \]
\[ \Lambda_{\text{sqrt}} := \text{diag}(\sqrt{\lambda}) \]
\[ S_{\text{sqrt}} := \varepsilon \cdot \Lambda_{\text{sqrt}} \cdot \varepsilon^T \quad < \text{Square root matrix of } S \]

Major axis of ellipse M:
\[ p_{x_1} := c \cdot \sqrt{\lambda_1} \cdot \varepsilon^{(1)} \quad p_{x_1} = \left( \begin{array}{c} 4.897984 \\ 0.019102 \end{array} \right) \]
\[ M_{x} := \text{augment}(p_{x_1}, -p_{x_1})^T \]
\[ c_{\text{e}} := \left( S_{\text{sqrt}}(x) \right)^T \quad < \text{points on the confidence ellipse based on } S \]
Plotting in R:

#CONFIDENCE ELLIPSE:
library(latticeExtra)
attach(M)

xyplot(X2 ~ X1,
  scales = "free", aspect='iso',
  par.settings = list(plot.symbol = list(col='blue', cex = 0.5, pch=19)),
  panel = function(x, y, ...) {
    panel.xyplot(x, y, ...)
    panel.ellipse(x,y,lwd=1,level=0.68,col='green', ...)
    panel.ellipse(x,y,lwd=1,level=0.90,col='brown', ...)
    panel.ellipse(x,y,lwd=2,level=0.95,col='red', ...)
  },
  auto.key = list(x = .1, y = .8, corner = c(0, 0)))

^ Although method for calculating Confidence Ellipses is not specified here, they appear to match explicit calculation in MathCad above.

Applying correlation structure to the data:

#APPLYING CORRELATION BETWEEN VARIABLES IN MATRIX M
#USING LINEAR TRANSFORMATION:
A=matrix(c(1,1.5,1.5,1),nrow=2,ncol=2,byrow=TRUE)
A
Mnew1=t(A%*%as.matrix(t(M)))

cor(Mnew1)

> A
[,1] [,2]
[1,] 1.0  1.5
[2,] 1.5  1.0

> cor(Mnew1)
[,1]     [,2]
[1,] 1.0000000 0.9469233
[2,] 0.9469233 1.0000000
AR presents a useful discussion and a graph concerning confidence ellipses calculated in this manner. Confidence ellipses generally do not match univariate confidence intervals (such as for $X_1$ or $X_2$ above) because the multivariate analysis takes into account covariance in $\Sigma$ (or $S$) whereas univariate intervals act as if the variables are independent. As a result, there are two discordant cases:

The multivariate confidence ellipse will determine that a case is outside the confidence limit set by $\alpha$ whereas one or both univariate analysis will consider the same case to be within a confidence intervals. This results because covariance between the variables has not been considered in the univariate analyses.

The multivariate confidence ellipse will determine that a case resides within the confidence ellipse whereas one of both univariate analysis will consider the same case to lie outside the confidence interval. This is an example of "Rao’s paradox" resulting from the fact that the univariate analyses are run on the same dataset thus have a "family-wise" $\alpha$ that should be considered, but is not.