CHAPTER 8

SOME ASPECTS OF MULTIVARIATE ANALYSIS

8.1 INTRODUCTION

In all the problems we have so far considered, observations are made of a single unidimensional response or output $y$. The inference problems that result are called univariate problems. In this and the next chapter, we shall consider problems which arise when the output is multidimensional. Thus, in the study of a chemical process, at each experimental setting one might observe yield $y_1$, density $y_2$, and color $y_3$ of the product. Similarly, in a study of consumer behavior, for each household one might record spending on food $y_1$, spending on durables $y_2$, and spending on travel and entertainment $y_3$. We would then say that a three-dimensional output or response is observed. Inference problems which arise in the analysis of such data are called multivariate.

In this chapter, we shall begin by reviewing some univariate problems in a general setting which can be easily extended to the multivariate case.

8.1.1 A General Univariate Model

It is often desired to make inferences about parameters $\theta_1, \ldots, \theta_k$ contained in the relationship between a single observed output variable or response $y$ subject to error and $p$ input invariables $\xi_1, \ldots, \xi_p$ whose values are assumed exactly known. It should be understood that the inputs could include qualitative as well as quantitative variables. For example, $\xi_i$ might take values of 0 or 1 depending on whether some particular quality was absent or present in which case $\xi_i$ is called an indicator variable or less appropriately a dummy variable.

The Design Matrix

Suppose, in an investigation, $n$ experimental "runs" are made, and the $i$th run consists of making an observation $y_i$ at some fixed set of input conditions $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{ip})$. The $n \times p$ design matrix

$$
\begin{bmatrix}
\xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\
\xi_{21} & \xi_{22} & \cdots & \xi_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\xi_{n1} & \xi_{n2} & \cdots & \xi_{np}
\end{bmatrix}
$$

(8.1.1)
lists the \( p \) input conditions to be used in each of the \( n \) projected runs and the \( u \)th row of \( \xi \) is the vector \( \xi_u \). The phraseology "experimental run", "experimental design" is most natural in a situation in which a scientific experiment is being conducted and in which the levels of the inputs are at our choice. In some applications, however, and particularly in economic studies, it is often impossible to choose the experimental conditions. We have only historical data generated for us in circumstances beyond our control and often in a manner we would not choose. It is convenient here to extend the terminologies "experimental run" and "experimental design" to include experiments designed by nature, but we must, of course, bear in mind the limitations of such historical data.

To obtain a mathematical model for our set-up we need to link the \( n \) observations \( y = (y_1, \ldots, y_n) \) with the inputs \( \xi \). This we do by defining two functions called respectively the expectation function and the error function.

*The Expectation Function*

The expected value \( E(y_u) \) of the output from the \( u \)th run is assumed to be a known function \( \eta \) of the \( p \) fixed inputs \( \xi_u \) employed during that run, involving \( k \) unknown parameters \( \theta = (\theta_1, \ldots, \theta_k) \),

\[
E(y_u) = \eta_u = \eta(\xi_u, \theta). \tag{8.1.2}
\]

The vector valued function \( \eta = \eta(\xi, \theta), \eta^t = (\eta_1, \ldots, \eta_n) \), is called the *expectation function*.

*The Error Function*

The expectation function links \( E(y_u) \) to \( \xi_u \) and \( \theta \). We now have to link \( y_u \) to \( E(y_u) = \eta_u \). This is done by means of an *error distribution function* in \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \). The \( n \) experimental errors \( \varepsilon = y - \eta \) which occur in making the runs are assumed to be random variables having zero means but in general not necessarily independently or Normally distributed. We denote the density function of the errors by \( p(\varepsilon | \pi) \) where \( \pi \) is a set of error distribution parameters whose values are in general unknown.

Finally, then, the output in the form of the \( n \) observations \( y \) and the input in the form of the \( n \) sets of conditions \( \xi \) are linked together by a *mathematical model* containing the error function and the expectation function as follows

\[
\xi \xrightarrow{\eta = \eta(\xi, \theta)} \eta \xrightarrow{p(\varepsilon | \pi)} y. \tag{8.1.3}
\]

This model involves a function

\[
\phi(y, \theta, \pi, \xi) \tag{8.1.4}
\]

of the observations \( y \), the parameters \( \theta \) of the expectation function, the parameters \( \pi \) of the error distribution and the design \( \xi \).
Data Generation Model

If we knew $\theta$ and $\pi$ and the design $\xi$, we could use the function (8.1.4) to calculate the probability density associated with any particular set of data $y$. This data generation model (which might, for example, be directly useful for simulation and Monte-Carlo studies) is the function $f(y, \theta, \pi, \xi)$ with $\theta$, $\pi$, and $\xi$ held fixed and we denote it by

$$p(y \mid \theta, \pi, \xi) = f(y, \theta, \pi, \xi),$$

which emphasizes that the density is a function of $y$ alone for fixed $\theta$, $\pi$, and $\xi$.

The Likelihood Function and the Posterior Distribution

In ordinary statistical practice, we are not directly interested in probabilities associated with various sets of data, given fixed values of the parameters $\theta$ and $\pi$. On the contrary, we are concerned with the probabilities associated with various sets of parameter values, given a fixed set of data which is known to have occurred.

After an experiment has been performed, $y$ is known and fixed (as is $\xi$) but $\theta$ and $\pi$ are unknown. The likelihood function has the same form as (8.1.4), but in it $y$ and $\xi$ are fixed and $\theta$ and $\pi$ are not to be regarded as variables. Thus, the likelihood may be written

$$l(\theta, \pi \mid y, \xi) = f(y, \theta, \pi, \xi).$$

In what follows we usually omit specific note of dependence on $\xi$ and write $l(\theta, \pi \mid y)$ as $l(\theta, \pi \mid y)$.

In the Bayesian framework, inferences about $\theta$ and $\pi$ can be made by suitable study of the posterior distribution $p(\theta, \pi \mid y)$ of $\theta$ and $\pi$ obtained by combining the likelihood with the appropriate prior distribution $p(\theta, \pi)$.

$$p(\theta, \pi \mid y) \propto l(\theta, \pi \mid y) \cdot p(\theta, \pi).$$

An example in which the expectation function is nonlinear and the error distribution is non-Normal was given in Section 3.5. In this chapter, we shall from now on assume Normality but will extend our general model to cover multivariate problems.

8.2 A GENERAL MULTIVARIATE NORMAL MODEL

Suppose now that a number of output responses are measured in each experimental run. Thus, in a chemical experiment, at each setting of the process conditions $\xi_1 =$ temperature and $\xi_2 =$ concentration, observations might be made on the output responses $y_1 =$ yield of product $A$, $y_2 =$ yield of product $B$, and $y_3 =$ yield of product $C$. In general, then, from each experimental run the
m-variate observation

\[ y_{(u)} = (y_{u1}, \ldots, y_{ui}, \ldots, y_{um}) \]

would be available. There would now be \( m \) expectation functions

\[ E(y_{(u)}) = \eta_{(u)} = (\eta_{u1}, \ldots, \eta_{um})' \]

where

\[ E(y_{ui}) = \eta_{ui} = \eta_i(\xi_{ui}, \theta_i) \]

\[ E(y_{ui}) = \eta_{ui} = \eta_i(\xi_{ui}, \theta_i) \]

\[ E(y_{ui}) = \eta_{ui} = \eta_i(\xi_{ui}, \theta_i) \]

\[ E(y_{u}) = \eta_{u} = \eta_u(\xi_u, \theta_u) \]

where \( \xi_{ui} \) would contain \( p_i \) elements \( (\xi_{ui1}, \ldots, \xi_{uip_i}) \) and \( \theta_i \) would contain \( k_i \) elements \( (\theta_{i1}, \ldots, \theta_{i1}, \ldots, \theta_{ik_i}) \). The expectation functions \( \eta_{ui} \) might be linear or nonlinear both in the parameters \( \theta_i \) and the inputs \( \xi_{ui} \). Also, depending on the problem, some or all of the \( p_i \) elements of \( \xi_{ui} \) might be the same as those of \( \xi_u \) and some or all of the elements of \( \theta_i \) might be the same as those of \( \theta_u \). That is to say, a given output would involve certain inputs and certain parameters which might or might not be shared by other outputs.

8.2.1 The Likelihood Function

Let us now consider the problem of making inferences about the \( \theta_i \) for a set of \( n \) \( m \)-variate observations. We assume that the error vector

\[ \varepsilon_{(u)} = y_{(u)} - \eta_{(u)} = (\varepsilon_{u1}, \ldots, \varepsilon_{um})', \quad u = 1, \ldots, n \]

is, for given \( \theta \) and \( \Sigma \), distributed as the \( m \)-variate Normal \( N_n(0, \Sigma) \), and that the runs are made in such a way that it can be assumed that from run to run the observations are independent. Thus, in terms of the general framework of (8.1.4), \( \Sigma = \pi \) are the parameters of an error distribution which is multivariate Normal. We first derive some very general results which apply to any model of this type, and then consider in detail the various important special cases that emerge if the expectation functions are supposed linear in the parameters \( \theta_i \).

The joint distribution of the \( n \) vectors of errors \( \varepsilon = (\varepsilon_{(1)}, \ldots, \varepsilon_{(u)}, \ldots, \varepsilon_{(n)})' \) is

\[ p(\varepsilon | \Sigma, \theta) = \prod_{u=1}^{n} p(\varepsilon_{(u)} | \Sigma, \theta) \]

\[ = (2\pi)^{-mn/2} |\Sigma|^{-n/2} \exp \left( -\frac{1}{2} \sum_{u=1}^{n} \varepsilon_{(u)}' \Sigma^{-1} \varepsilon_{(u)} \right) \]

\[ < \infty \quad \varepsilon_{ui} < \infty, \quad i = 1, \ldots, m, \quad u = 1, \ldots, n \]
where \( \Sigma = (\sigma_{ij}) \) is the \( m \times m \) covariance matrix, \( \Sigma_{-1} = (\sigma_{ij}^{-1}) \) its inverse and \( \Theta \) refers to the complete set of all the \( (k_1 + \cdots + k_m) \) parameters \( \theta_1, \ldots, \theta_m \).

Denoting \( S(\Theta) \) to be the \( m \times m \) symmetric matrix

\[
S(\Theta) = \{S_{ij}(\theta_i, \theta_j)\}
\]

with

\[
S_{ij}(\theta_i, \theta_j) = \sum_{u=1}^{n} e_{ui} e_{uj} = \sum_{u=1}^{n} [y_{ui} - \eta_i(\xi_{ui}, \Theta)] [y_{uj} - \eta_j(\xi_{uj}, \Theta)], \quad i, j = 1, \ldots, m,
\]

then the exponent in (8.2.3) can be expressed as

\[
\sum_{u=1}^{n} e_{ui} \Sigma^{-1} e_{ui} = \text{tr} S(\Theta) \Sigma^{-1} = \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij} S_{ij}(\theta_i, \theta_j)
\]

where \( \text{tr} A \) means the trace of the matrix \( A \). Given the observations, the likelihood function can thus be written

\[
\ell(\Theta, \Sigma | y) \propto p(\varepsilon | \Sigma, \Theta) \\
\propto |\Sigma|^{-n/2} \exp \left[ -\frac{1}{2} \text{tr} \Sigma^{-1} S(\Theta) \right].
\]

To clarify the notation, we emphasize that \( y \) refers to the \( n \times m \) matrix of observations

\[
y = \begin{bmatrix}
y_{11} & \cdots & y_{1i} & \cdots & y_{1m} \\
\vdots & & \vdots & & \vdots \\
y_{ui} & \cdots & y_{ui} & \cdots & y_{um} \\
\vdots & & \vdots & & \vdots \\
y_{ni} & \cdots & y_{ni} & \cdots & y_{nm}
\end{bmatrix} = [y_{1}, \ldots, y_{i}, \ldots, y_{m}] = \begin{bmatrix}
y_{(1)} \\
\vdots \\
y_{(i)} \\
\vdots \\
y_{(m)}
\end{bmatrix},
\]

where \( y_i = (y_{i1}, \ldots, y_{im})' \) is the vector of \( n \) observations corresponding to the \( i \)th response and \( y_{(t)} = (y_{t1}, \ldots, y_{tm})' \) is the vector of \( m \) observations of the \( n \)th experimental run. Similarly, \( \varepsilon \) refers to the \( n \times m \) matrix of errors

\[
\varepsilon = \begin{bmatrix}
e_{11} & \cdots & e_{1i} & \cdots & e_{1m} \\
\vdots & & \vdots & & \vdots \\
e_{ui} & \cdots & e_{ui} & \cdots & e_{um} \\
\vdots & & \vdots & & \vdots \\
e_{ni} & \cdots & e_{ni} & \cdots & e_{nm}
\end{bmatrix} = [e_{1}, \ldots, e_{i}, \ldots, e_{m}] = \begin{bmatrix}
e_{(1)} \\
\vdots \\
e_{(i)} \\
\vdots \\
e_{(m)}
\end{bmatrix}.
\]

8.2.2 Prior Distribution of \( (\Theta, \Sigma) \)

For the prior distribution of the parameters \( (\Theta, \Sigma) \), we shall first assume that \( \Theta \) and \( \Sigma \) are approximately independent so that

\[
p(\Theta, \Sigma) = p(\Theta)p(\Sigma).
\]
We shall further suppose that the parameterization in terms of \( \Theta \) is so chosen such that it is appropriate to take \( \Theta \) as locally uniform,\(^\dagger\)

\[
p(\Theta) \propto \text{constant.} \quad (8.2.8)
\]

For the prior distribution of the \( \frac{1}{2}mn(m+1) \) distinct elements of \( \Sigma \), application of the argument in Section 1.3 for the multiparameter situation leads to the noninformative reference prior

\[
p(\Sigma) \propto |\mathcal{S}(\Sigma)|^{1/2} \quad (8.2.9)
\]

Now,

\[
|\mathcal{S}(\Sigma)| = |\mathcal{S}(\Sigma^{-1})| \left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right|^{-2}, \quad (8.2.10)
\]

where

\[
\left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right| = \left| \frac{\partial (\sigma_{11}, \sigma_{12}, \ldots, \sigma_{mn})}{\partial (\sigma_{11}, \sigma_{12}, \ldots, \sigma_{mn})} \frac{\partial \Sigma^{-1}}{\partial \Sigma^{-1}} \right| \quad (8.2.11)
\]

is the Jacobian of the transformation from the elements \( \sigma_{ij} \) of \( \Sigma \) to the elements \( \sigma^{ij} \) of \( \Sigma^{-1} \). It is shown in Appendix A8.2 that

\[
|\mathcal{S}(\Sigma^{-1})| \propto \left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right| \quad (8.2.12)
\]

and that

\[
\left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right| = |\Sigma|^{m+1}. \quad (8.2.13)
\]

Thus,

\[
p(\Sigma) \propto |\Sigma|^{-\frac{1}{2}(m+1)}. \quad (8.2.14)
\]

In this special case \( m = 1 \), (8.2.14) reduces to

\[
p(\sigma_{11}) \propto \frac{1}{\sigma_{11}} \quad (8.2.15)
\]

which coincides with the usual assumption concerning a noninformative prior distribution for a single variance. Another special case of interest is when the errors \( (e_{a1}, \ldots, e_{am}) \) are uncorrelated, that is, \( \sigma_{ij} = 0 \) if \( i \neq j \). In this case, the same argument leads to

\[
p(\Sigma \mid \sigma_{ij} = 0, i \neq j) = p(\sigma_{11}, \ldots, \sigma_{nm}) \propto \prod_{i=1}^{m} \sigma_{ii}^{-1}. \quad (8.2.16)
\]

\(^\dagger\) As we have mentioned earlier, when the parameter space is of high dimension, the use of the locally uniform prior may be inappropriate and more careful considerations should be given to the structure of the model in selecting a noninformative prior.
8.2.3 Posterior Distribution of \((\theta, \Sigma)\)

Using (8.2.6), (8.2.8), and (8.2.14), the joint posterior distribution of \((\theta, \Sigma)\) is
\[
p(\theta, \Sigma \mid y) \propto |\Sigma|^{-\frac{1}{2}(n+m+1)} \exp \left[ -\frac{1}{2} \text{tr} \Sigma^{-1} S(\theta) \right], \quad -\infty < \theta < \infty, \quad \Sigma > 0,
\]
where the notation \(-\infty < \theta < \infty\) means that each element of the set of parameters \(\theta\) can vary from \(-\infty\) to \(\infty\), and the notation \(\Sigma > 0\) means that the \(\frac{1}{2}m(m+1)\) elements \(\sigma_{ij}\) are such that the random matrix \(\Sigma\) is positive definite.

It is sometimes convenient to work with the elements of \(\Sigma^{-1} = \{\sigma^{ij}\}\) rather than the elements of \(\Sigma\). Since
\[
p(\theta, \Sigma^{-1} \mid y) = p(\theta, \Sigma \mid y) \left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right|^{-1}
\]
(8.2.18)

it follows from (8.2.13) that the posterior distribution of \((\theta, \Sigma^{-1})\) is
\[
p(\theta, \Sigma^{-1} \mid y) \propto |\Sigma^{-1}|^{\frac{1}{2}(n+m+1)} \exp \left[ -\frac{1}{2} \text{tr} \Sigma^{-1} S(\theta) \right], \quad -\infty < \theta < \infty, \quad \Sigma^{-1} > 0.
\]
(8.2.19)

8.2.4 The Wishart Distribution

We now introduce a distribution which is basic in Normal theory multivariate problems. Let \(Z\) be a \(m \times m\) positive definite symmetric random matrix which consists of \(\frac{1}{2}m(m+1)\) distinct random variables \(z_{ij}\) \((i, j = 1, \ldots, m; \ i \geq j\). Let \(q > 0\), and \(B\) be a \(m \times m\) positive definite symmetric matrix of fixed constants. The distribution of \(z_{ij}\)
\[
p(Z) \propto |Z|^{(q+1)/2} \exp \left( -\frac{1}{2} \text{tr} ZB \right), \quad Z > 0 \quad (8.2.20)
\]
obtained by Wishart (1928), is a multivariate generalization of the \(\chi^2\) distribution.

It can be shown that
\[
\int_{\mathbb{S}^{m-1}} |Z|^{(q+1)/2} \exp \left( -\frac{1}{2} \text{tr} ZB \right) dZ = |B|^{-\frac{1}{2}(q+m-1)} 2^{\frac{1}{2}m(m+1)} \Gamma_{m} \left( \frac{q + m - 1}{2} \right)
\]
(8.2.21)

where \(\Gamma_{m}(b)\) is the generalized gamma function, Siegel (1935)
\[
\Gamma_{m}(b) = [\Gamma(\frac{1}{2})]^{2^{m-1}} \prod_{d=1}^{m} \Gamma \left( b + \frac{d - 1}{2} \right), \quad b > \frac{m - 1}{2}. \quad (8.2.22)
\]

We shall denote the distribution (8.2.20) by \(W_{m}(B^{-1}, q)\) and say that \(Z\) is distributed as Wishart with \(q\) degrees of freedom and parameter matrix \(B^{-1}\). For a discussion of the properties of the Wishart distribution, see for example Anderson (1958). Note carefully that the parameterization used in (8.2.20) is different from the one used in Anderson in one respect. In his notation, the
distribution in (8.2.20) is denoted as \( W(B^{-1}, n) \) where \( n = q + m - 1 \) is said to be the degrees of freedom.

As an application of the Wishart distribution, we see in (8.2.19) that, given \( \theta, \Sigma^{-1} \) is distributed as \( W_n[S^{-1}(\theta), n - m + 1] \) provided \( n \geq m \).

**8.2.5 Posterior Distribution of \( \theta \)**

Using the identity (8.2.21), we immediately obtain from (8.2.19) the marginal posterior distribution of \( \theta \) as

\[
p(\theta | y) \propto |S(\theta)|^{-n/2}, \quad -\infty < \theta < \infty,
\]

provided \( n \geq m \).

This extremely simple result is remarkable because of its generality. It will be noted that to reach it we have not had to assume either:

a) that any of the input variables \( \xi_{nt} \) were or were not common to more than one output, or

b) that the parameters \( \theta_i \) were or were not common to more than one output, or

c) that the expectation functions were linear or were nonlinear in the parameters.

This generality may be contrasted with the specification needed to obtain "nice" sampling theory results. For example, a common formulation assumes that the \( \xi_{nt} \) are common, that the \( \theta_i \) are not, and that the expectation functions are all linear in the parameters.

In the special case in which there is only one output response \( y \), (8.2.23) reduces to

\[
p(\theta | y) \propto |S(\theta)|^{-n/2}, \quad -\infty < \theta < \infty,
\]

with \( S(\theta) = \sum_{n=1}^{n} [y_n - \eta(\xi_{nt}, \theta)]^2 \). As we have seen, this result can be regarded as supplying a Bayesian justification of least squares, since the modal values of \( \theta \) (those associated with maximum posterior density) are those which minimize \( S \).

The general result (8.2.23) supplies then, among other things, an appropriate Bayesian multivariate generalization of least squares. The "most probable" values of \( \theta \) being simply those which minimize the determinant \( |S(\theta)| \).

Finally, in the special case \( \sigma_{ij} = 0, i \neq j \), combining (8.2.16) with (8.2.6) and integrating out \( \sigma_{11}, \ldots, \sigma_{nn} \) yields

\[
p(\theta | y) \propto \prod_{i=1}^{m} [S_i(\theta_i)]^{-n/2}, \quad -\infty < \theta < \infty.
\]

**8.2.6 Estimation of Common Parameters in a Nonlinear Multivariate Model**

We now illustrate the general applicability of the result (8.2.23) by considering an example in which:
a) certain of the $\theta$'s are common to more than one output, and

b) the expectation functions are nonlinear in the parameters.

Fig. 8.2.1 Diagrammatic representation of a system $A \rightarrow B \rightarrow C$.

Suppose we have the consecutive system indicated in Fig. 8.2.1, which shows water running from a tank $A$ via a tap opened an amount $\phi_1$ into a tank $B$ which then runs into a tank $C$ via a tap opened an amount $\phi_2$.

If $\eta_1$, $\eta_2$, and $\eta_3$ are the proportions of $A$, $B$, and $C$ present at time $\xi$, with initial conditions ($\eta_1 = 1$, $\eta_2 = 0$, $\eta_3 = 0$), the system can be described by the differential equations

\[
\frac{d\eta_1}{d\xi} = -\phi_1 \eta_1, \\
\frac{d\eta_2}{d\xi} = \phi_1 \eta_1 - \phi_2 \eta_2, \\
\frac{d\eta_3}{d\xi} = \phi_2 \eta_2.
\] (8.2.26)
Systems of this kind have many applications in engineering and in the physical and biological sciences. In particular, the equation (8.2.26) could represent a consecutive first-order chemical reaction in which a substance $A$ decomposed to form $B$, which in turn decomposed to form $C$. The responses $\eta_1, \eta_2, \eta_3$ would then be the mole fractions of $A, B,$ and $C$ present at time $\xi$ and the quantities $\phi_1$ and $\phi_2$ would then be rate constants associated with the first and second decompositions and would normally have to be estimated from data.

If we denote by $y_1, y_2, \text{and} y_3$ the observed values of $\eta_1, \eta_2, \text{and} \eta_3$, then, on integration of (8.2.26), we have the expectation functions

$$E(y_1) = \eta_1 = e^{-\phi_1 \xi},$$

$$E(y_2) = \eta_1 = (e^{-\phi_1 \xi} - e^{-\phi_2 \xi}) \phi_1 / (\phi_2 - \phi_1),$$

$$E(y_3) = \eta_3 = 1 + (-\phi_2 e^{-\phi_1 \xi} + \phi_1 e^{-\phi_2 \xi}) / (\phi_2 - \phi_1),$$

and it is to be noted that for all $\xi$,

$$\eta_1 + \eta_2 + \eta_3 = 1.$$

Observations on $y_1$ could yield information only on $\phi_1$, but observations on $y_2$ and $y_3$ could each provide information on both $\phi_1$ and $\phi_2$. If measurements of more than one of the quantities $(y_1, y_2, y_3)$ were available, we should certainly expect to be able to estimate the parameters more precisely. The Bayesian approach allows us to pool the information from $(y_1, y_2, y_3)$ and makes it easy

| Table 8.2.1 |
| Observations on the yield of three substances in a chemical reaction |
| Time = $\xi_u$ | Yield of $A$ | Yield of $B$ | Yield of $C$ |
| $1/2$ | 0.959 | 0.025 | 0.028 |
| $1/4$ | 0.914 | 0.061 | 0.000 |
| $1$ | 0.855 | 0.152 | 0.068 |
| $2$ | 0.785 | 0.197 | 0.096 |
| $2$ | 0.628 | 0.130 | 0.090 |
| $4$ | 0.617 | 0.249 | 0.118 |
| $4$ | 0.480 | 0.184 | 0.374 |
| $8^+$ | 0.166 | 0.147 | 0.651 |
| $8^+$ | 0.205 | 0.050 | 0.684 |
| $16^+$ | 0.034 | 0.000 | 0.899 |
| $16^+$ | 0.054 | 0.047 | 0.991 |

$^+$ These four runs are omitted in the second analysis.
to appreciate the contribution from each of the three responses. In this example, \( \xi \) is the only input variable and is the elapsed time since the start of the reaction. We denote by \( Y_{10} = (Y_{11}, Y_{12}, Y_{13}) \) a set of \( m = 3 \) observations made on \( \eta_{10}, \eta_{20}, \eta_{30} \) at time \( \xi_{10} \). A typical set of such observations is shown in Table 8.2.1.

In some observations may not be available on all three of the outputs. Thus only the concentration \( y_2 \) of the product \( B \) might be observable, or \( y_2 \) and \( y_3 \) might be known, but there might be no independently measured observation \( y_1 \) of the concentration of \( A \).†

We suppose that the observations of Table 8.2.1 may be treated as having arisen from 12 *independent* experimental runs, as might be appropriate if the runs were carried out in random order in sealed tubes, each reaction being terminated at the appropriate time by sudden cooling. Furthermore, we suppose that \( (y_1, y_2, y_3) \) are functionally independent so that the \( 3 \times 3 \) matrix \( \Sigma \) may be assumed to be positive definite and contains three variances and three covariances, all unknown. It is perhaps most natural for the experimenter to think in terms of the logarithms \( \theta_1 = \log \phi_1 \) and \( \theta_2 = \log \phi_2 \) of the rate constants and to regard these as locally uniformly distributed *a priori*.‡ We shall, therefore, choose as our reference priors for \( (\theta_1, \theta_2) \) and \( \Sigma \) the distributions in (8.2.8) and (8.2.14), respectively.

† When the chemist has difficulty in determining one of the products he sometimes makes use of relations like (8.2.27d) to "obtain it by calculation." Thus he might "obtain" \( y_1 \) from the relation \( y_1 = 1 - y_2 - y_3 \). For the resulting data set, the \( 3 \times 3 \) covariance matrix \( \Sigma \) will of course not be positive definite, and the analysis in terms of three-dimensional responses will be inappropriate. In particular, the determinant of the sums of squares and products which appears in (8.2.23) will be zero *whatever* the values of the parameters. The difficulty is of course overcome very simply. The quantity \( y_1 \) is not an observation and the *data* has two dimensions, not three. The analysis should be carried through with \( y_2 \) and \( y_3 \) which *have* actually been measured. For a fuller treatment of problems of this kind arising because of data dependence or near dependence, see Box, Hunter, and MacGregor (1977).

‡ Suppose that (a) the expectation functions were linear in \( \theta_1(\phi) \) and \( \theta_2(\phi) \) where \( \phi = (\phi_1, \phi_2) \), (b) little was known *a priori* about either parameter compared with the information supplied by the data, and (c) any prior information about one parameter would supply essentially none about the other.

Then, arguing as in Section 1.3, a noninformative reference prior to \( \theta \) should be locally uniform.

Conditions (b) and (c) are likely to be applicable to this problem at least as approximations, but condition (a) is not, because the expectation functions are non-linear in \( \phi_1 \) and \( \phi_2 \) and no general linearizing transformation exists. However, [see for example Beale (1960), and Guttman and Meeter (1965)] the expectation functions are more "nearly linear" in \( \theta_1 = \log \phi_1 \) and \( \theta_2 = \log \phi_2 \). Thus, the assumption that \( \theta_1 \) and \( \theta_2 \) are locally uniform provides a better approximation to a noninformative prior for the rate constants. For reasons we have discussed earlier, the assumption is not critical and, if for example we assume \( \phi_1 \) and \( \phi_2 \) themselves to be locally uniform, the posterior distribution is not altered appreciably.
Expression (8.2.23) makes it possible to compute the posterior density for the parameters assuming observations are available on some or all of the products $A$, $B$, and $C$. Thus, we may consider the posterior distribution of $\mathbf{\theta} = (\theta_1, \theta_2)'$

a) If only $y_2$ of product $B$ are available

\[ p(\mathbf{\theta} | y_2) \propto [S_{22}(\mathbf{\theta})]^{-1/2}, \quad -\infty < \mathbf{\theta} < \infty, \quad (8.2.28a) \]

b) If only $y_3$ of product $C$ are available,

\[ p(\mathbf{\theta} | y_3) \propto [S_{33}(\mathbf{\theta})]^{-1/2}, \quad -\infty < \mathbf{\theta} < \infty, \quad (8.2.28b) \]

c) If only $y_2$ and $y_3$ of $B$ and $C$ are available

\[ p(\mathbf{\theta} | y_2, y_3) \propto \begin{vmatrix} S_{22}(\mathbf{\theta}) & S_{23}(\mathbf{\theta}) \\ S_{32}(\mathbf{\theta}) & S_{33}(\mathbf{\theta}) \end{vmatrix}^{-1/2}, \quad -\infty < \mathbf{\theta} < \infty, \quad (8.2.28c) \]

and

d) If $y_1$, $y_2$, and $y_3$ of the products $A$, $B$, and $C$ are all available

\[ p(\mathbf{\theta} | y) \propto |S(\mathbf{\theta})|^{-1/2}, \quad -\infty < \mathbf{\theta} < \infty, \quad (8.2.28d) \]

where $S(\mathbf{\theta}) = \{S_{ij}(\mathbf{\theta})\}, i, j = 1, 2, 3$.

\[ \theta_2 = \log \phi_2 \]

**Fig. 8.2.2** 99.75% H.P.D. regions for $\theta_1$ and $\theta_2$ for the chemical reaction data.
Since there are only two parameters $\theta_1$ and $\theta_2$, the posterior distributions can be represented by contour diagrams which may be superimposed to show the contributions made by the various output responses. Single contours are shown in Fig. 8.2.2 of the posterior distributions of $\theta_1$ and $\theta_2$ for (a) $y_2$ alone, (b) $y_3$ alone, (c) $y_2$ and $y_3$ jointly, and (d) $y_1$, $y_2$, and $y_3$ jointly. The contours actually shown are those which should correspond to an H.P.D. region containing approximately 99.75% of the probability mass calculated from

$$\log p(\hat{\theta} | \cdot) - \log p(\theta | \cdot) = \frac{1}{2} x^2(2, x), \quad x = 0.0025$$

where $p(\theta | \cdot)$ refers to the appropriate distributions in (8.2.28a-d) and $\hat{\theta}$ the corresponding modal values of $\theta$. In this example, it is apparent, particularly for $y_3$, that the posterior distributions are non-Normal. Nevertheless, the above very crude approximation will suffice for the purpose of the present discussion.

In studying Figure 8.2.2, we first consider the moon-shaped contour obtained from observations $y_3$ on the end product $C$ alone. In any sequential reaction $A \rightarrow B \rightarrow C \rightarrow \ldots$ etc., we should expect that observation of only the end product (C in this case) could provide little or no information about the individual parameters but only about some aggregate of these rate constants. A diagonally attenuated ridge-like surface is therefore to be expected. However, it should be further noted that since in this specific instance $\eta_3$ is symmetric in $\theta_1$ and $\theta_2$ [see expression (8.2.27c)], the posterior surface is completely symmetric about the line $\theta_1 = \theta_2$. In particular, if $(\theta_1, \theta_2)$ is a point of maximum density the point $(\theta_2, \theta_1)$ will also give the same maximum density. In general the surface will be bimodal and have two peaks of equal height symmetrically situated about the equi-angular line. Marginal distributions will thus display precisely the kind of behaviour shown in Fig. A.5.6.1.

Figure 8.2.2 shows how, for this data, the inevitable ambiguity arising when only observations $y_3$ on product $C$ are utilized, is resolved as soon as the additional information supplied by values $y_2$ on the intermediate product $B$ is considered. As can be expected, the nature of the evidence that the intermediate product $y_2$ contributes, is preferentially concerned with the difference of the parameters. This is evidenced by the tendency of the region to be obliquely oriented approximately at right angles to that for $y_3$. By combining information from the two sources we obtain a much smaller region contained within the intersection of the individual regions. Finally, information from $y_1$ which casts further light on the value of $\theta_1$, causes the region to be further reduced.

Data of this kind sometimes occur in which available observations trace only part of the reaction. To demonstrate the effect of this kind of inadequacy in the experimental design, the analysis is repeated omitting the last four observations in Table 8.2.1. As shown in Fig. 8.2.3, over the ranges studied, the contours for $y_2$ alone and $y_3$ alone do not now close. Nevertheless, quite precise estimation is possible using $y_2$ and $y_3$ together and the addition of $y_1$ improves the estimation further.
Fig. 8.2.3 99.75% H.P.D. regions for $\theta_1$ and $\theta_2$, excluding the last four observations.

*Precautions in the Estimation of Common Parameters*

Even in cases where only a single response is being considered, caution is needed in the fitting of functions. As explained in Section 1.1.4, fitting should be regarded as merely one element in the iterative model building process. The appropriate attitude is that when the model is initially fitted it is tentatively entertained rather than assumed. Careful checks on residuals are applied in a process of model criticism to see whether there is reason to doubt its applicability to the situation under consideration.

The importance of such precaution is even greater when several responses are considered. In multivariate problems, not only should each response model be checked individually but they must also be checked for overall consistency. The investigator should in practice *not* revert immediately to a joint analysis of responses. He should:
1) check the individual fit of each response,
2) compare posterior distributions to appraise the consistency of the information from the various responses (an aspect discussed in more detail in Chapter 9).

Only in those cases where he is satisfied with the individual fit and with the consistency shall he revert to the joint analysis.

8.3 LINEAR MULTIVARIATE MODELS

In discussing the general $m$-variate Normal model above, we have not needed to assume anything specific about the form of the $m$ expectation functions $\eta$. In particular, they need not be linear in the parameters† nor does it matter whether or not some parameters appear in more than one of the expectation functions. Many interesting and informative special cases arise if we suppose the expectation functions to be linear in the $\theta$'s. Moreover, as will be seen, the linear results can sometimes supply adequate local approximations for models non-linear in the parameters. From now on then we assume that

$$E(y_{ui}) = \eta_i(\xi_{uai}, \theta_i) = x_{(ui)}'\theta_i, \quad i = 1, \ldots, m, \quad u = 1, \ldots, n,$$

(8.3.1)

where

$$\theta_i = (\theta_{i1}, \ldots, \theta_{ip}, \ldots, \theta_{ik_i})$$

and

$$x_{(ui)} = (x_{u(i)1}, \ldots, x_{u(i)p}, \ldots, x_{u(i)k_i})$$

with

$$x_{u(i)} = \frac{\partial \eta_i(\xi_{uai}, \theta_i)}{\partial \theta_{pi}}$$

independent of all the $\theta$'s.

The $n \times k_i$ matrix $X_i$ whose $u$th row is $x_{(ui)}$ will be called the derivative matrix for the $i$th response.

Our linear $m$-variate model may now be written as

$$y_1 = X_1\theta_1 + \epsilon_1$$

$$\vdots$$

$$y_i = X_i\theta_i + \epsilon_i$$

$$\vdots$$

$$y_n = X_m\theta_m + \epsilon_m.$$

(8.3.2)

Certain characteristics of this mode of writing the model should be noted. In particular, it is clear that while the elements of $x_{(ui)}$ will be functions of the

† Although, so that a uniform density can represent an approximately noninformative prior and also to assist local linear approximation, parameter transformations in terms of which the expectation function is more nearly linear, will often be employed.
elements of the vector input variables $x_{mt}$, they will in general not be proportional to the elements of $x_{mt}$ themselves. Thus if

$$E(y_{mt}) = \theta_1 \log \xi_{mt} + \theta_2 \xi_{mt} \xi_{mt},$$

then

$$x_{mt} = \frac{\log \xi_{mt}}{\xi_{mt}} \quad \text{and} \quad x_{mt} = \frac{\xi_{mt}}{\xi_{mt}}$$

8.3.1 The Use of Linear Theory Approximations when the Expectation is Nonlinear in the Parameters

The specific form of posterior distributions which we shall obtain for the linear case will often provide reasonably close approximations even when the expectation functions $\eta_{mt}$ is nonlinear in $\theta$. This is because we need only that the expectation functions are approximately linear in the region of the parameter space covered by most of the posterior distribution, say within the 95% H.P.D. region. For moderate $n$, this can happen with functions that are highly nonlinear in the parameters when considered over their whole range. Then, in the region where the posterior probability mass is concentrated (say the 95% H.P.D. region), we may expand the expectation function around the mode $\hat{\theta}$,

$$E(y_{mt}) = \eta_{mt} = \eta_{mt}(\xi_{mt}, \hat{\theta}) + \sum_{g=1}^{k_1} x_{mgt}(\theta_{gt} - \hat{\theta}_{gt}), \quad (8.3.3)$$

where

$$x_{mgt} = \frac{\partial \eta_{mt}(\xi_{mt}, \hat{\theta})}{\partial \theta_{gt}} \bigg|_{\theta_{g} = \hat{\theta}_{g}},$$

which is, approximately, in the form of a linear model. Thus, the posterior distributions found from linear theory can, in many cases, provide close approximations to the true distributions. For example, in the univariate case ($m = 1$) with a single parameter $\theta$, the posterior distribution in (8.2.24) would be approximately

$$p(\theta | y) \propto [v s^2 + (\Sigma x^2) (\theta - \hat{\theta})^2]^{-n/2} \quad (8.3.4)$$

where

$$v = n - 1, \quad s^2 = \frac{1}{v} \Sigma [y_{mt} - \eta(\xi_{mt}, \hat{\theta})]^2 \quad \text{and} \quad x = \frac{\partial \eta(\xi_{mt}, \theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}}$$

so that the quantity

$$\frac{\sqrt{\sum x^2} (\theta - \hat{\theta})}{s}$$

would be approximately distributed as $t(0, 1, v).$

† A possibility that can be checked posteriori for any specific case.
Linear Multivariate Models

When, as in the case in which multivariate or univariate least squares is appropriate, a convenient method of calculating the \( \hat{\theta}'s \) is available for the linear case but not for the corresponding nonlinear situation, the linearization may be used iteratively to find the \( \hat{\theta}'s \) for the nonlinear situation.

For example, in the univariate model containing a single parameter \( \theta \), with a first guess \( \theta_0 \), we can write approximately

\[
E(z_{00}) = (\hat{\theta} - \theta_0) x_u
\]

where

\[
z_{00} = y - \eta(\xi, \theta_0) \quad \text{and} \quad x_u = \frac{\partial \eta(\xi, \theta)}{\partial \theta} \bigg|_{\theta = \theta_0}
\]

Applying ordinary least squares to the model, we obtain an estimate of the correction \( \hat{\theta} - \theta_0 \) and hence hopefully an improved “guess” \( \theta_1 \) from

\[
\theta_1 = \theta_0 + \frac{\Sigma z_{00} x_u}{\Sigma x_u^2}
\]

This is the well-known Newton-Gauss method of iteration for nonlinear least squares, Box (1957, 1960), Hartley (1961), Marquardt (1963), and under favorable conditions the successive iterants will converge to \( \hat{\theta} \).

8.3.2 Special Cases of the General Linear Multivariate Model

In general, the joint distribution of \( \theta \) and \( \Sigma \) is given by (8.2.17) and the marginal distribution of \( \theta \) is that in (8.2.23) quite independently of whether \( \eta(\xi, \theta_0) \) is linear in \( \theta_0 \) or not. For practical purposes, however, it is of interest to consider a number of special cases.

For orientation we reconsider for a moment the linear univariate situation discussed earlier in Section 2.7,

\[
y = X \theta + \varepsilon,
\]

where \( y \) is a \( n \times 1 \) vector of observations, \( X \) a \( n \times k \) matrix of fixed elements, \( \theta \) a \( k \times 1 \) vector of parameters and \( \varepsilon \) a \( n \times 1 \) vector of errors. In this case,

\[
\rho(\theta, \sigma^2 \mid y) \propto (\sigma^2)^{-\frac{k+n+1}{2}} \exp \left[ -\frac{S(\theta)}{2\sigma^2} \right], \quad \sigma^2 > 0, \quad -\infty < \theta < \infty,
\]

and

\[
\rho(\theta \mid y) \propto [S(\theta)]^{-n/2}, \quad -\infty < \theta < \infty.
\]

The determinant \( |S(\theta)| \) in (8.2.23) becomes the single sum of squares

\[
S(\theta) = (y - X \theta)' (y - X \theta).
\]

In this linear case, we may write

\[
S(\theta) = (n - k) s^2 + (\hat{\theta} - \theta)' X' X (\hat{\theta} - \theta),
\]
where
\[(n - k)s^2 = (y - \bar{y})' (y - \bar{y}) = (y - X\hat{\theta})' (y - X\hat{\theta})\]
and, assuming \(X\) is of rank \(k\),
\[\hat{\theta} = (X'X)^{-1} X'y\]
so that, writing \(v = n - k\),
\[p(\hat{\theta} | y) \propto \left[ 1 + \frac{(\hat{\theta} - \theta)'X'X(\hat{\theta} - \theta)}{\nu s^2} \right]^{-\frac{\nu + k}{2}}, \quad -\infty < \theta < \infty. \quad (8.3.13)\]

The posterior distribution of \(\theta\) is thus the \(k\) dimensional \(t_{\nu}[\hat{\theta}, \nu s^2(X'X)^{-1}, v]\) distribution. Further, integrating out \(\theta\) from (8.3.9) yields the distribution of \(\sigma^2\),
\[p(\sigma^2 | y) \propto (\sigma^2)^{-\frac{\nu + 1}{2}} \exp \left( -\frac{\nu s^2}{2\sigma^2} \right), \quad \sigma^2 > 0, \quad (8.3.14)\]
so that \(\sigma^2/(\nu s^2)\) has the \(\chi^2_{v-2}\) distribution. All the above results have of course, been already obtained earlier in Section 2.7.

It is clear that the general linear model (8.3.2) which can be regarded as the multivariate generalization of (8.3.8) need not be particularized in any way. The matrices \(X_1, ..., X_m\) may or may not have elements in common; furthermore, the vectors of parameters \(\theta_1, ..., \theta_m\) may or may not have elements in common. Using sampling theory, Zellner (1962, 1963) attempted to study the situation in which the \(X_i\) were not assumed to be identical. The main difficulty with his approach was that the minimum variance estimator for \(\theta\) involves the unknown \(\Sigma\), and the estimators proposed are “optimal” only in the asymptotic sense.

Cases of special interest which are associated with practical problems of importance and which relate to known results include:

a) when the derivative matrices \(X_1 = ... = X_m = X\) are common but the parameters \(\theta_1, ..., \theta_m\) are not,
b) when \(\theta_1 = ... = \theta_m\) but the matrices \(X_1, ..., X_m\) are not, and
c) when \(\theta_1 = ... = \theta_m\) and \(X_1 = ... = X_m\).

In the remaining part of this chapter, we shall discuss case (a). The problem of estimating common parameters which includes (b) and (c) will be treated in the next chapter.

### 8.4 Inferences about \(\theta\) for the Case of a Common Derivative Matrix \(X\)

The model for which \(X_1 = ... = X_m = X\) (so that \(k_1 = ... = k_m = k\)) and \(\theta_1 \neq ... \neq \theta_m\) has received most attention in the sampling theory framework — see, for example, Anderson (1958). From the Bayesian point of view, the problem has been studied by Savage (1961a), Geisser and Cornfield (1965), Geisser (1965a), Ando and Kaufman (1965), and others. In general, the
The multivariate model in (8.3.2) can now be written

\[ y = X\theta + \varepsilon \]

\[ \begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} X \end{bmatrix} \begin{bmatrix} \theta \end{bmatrix} + \begin{bmatrix} \varepsilon \end{bmatrix} \]

\[ n \times m \quad n \times k \quad k \times m \quad n \times m \]

where the notation beneath the matrices indicates that \( y \) is an \( n \times m \) matrix of \( m \)-variate observations, \( \theta \) is a \( k \times m \) matrix of parameters and \( \varepsilon \) an \( n \times m \) matrix of errors.

The model would be appropriate for example if a \( 2^p \) factorial experiment had been conducted on a chemical process and the output \( y_1 = \) product yield, \( y_2 = \) product purity, \( y_3 = \) product density had been measured. The elements of each column of the common matrix \( X \) would then be an appropriate sequence of \( +1 \)'s and \( -1 \)'s corresponding to the experimental conditions and the "effect" parameters \( \theta \), would be different for each output. In econometrics, the model (8.4.1) is frequently encountered in the analysis of the reduced form of simultaneous equation systems.

We note that the \( k \times m \) matrix of parameters

\[ \theta = \begin{bmatrix} \theta_1 & \ldots & \theta_{k1} & \ldots & \theta_{k1m} \\ \vdots & & \vdots & & \vdots \\ \theta_p & \ldots & \theta_{p1} & \ldots & \theta_{pm} \\ \vdots & & \vdots & & \vdots \\ \theta_{kl} & \ldots & \theta_{kl1} & \ldots & \theta_{klm} \end{bmatrix} \]  

(8.4.2)

can be written in the two alternative forms

\[ \theta = [\theta_1, \ldots, \theta_l, \ldots, \theta_m] = \begin{bmatrix} \theta'_1 \\ \vdots \\ \theta'_l \\ \vdots \\ \theta'_m \end{bmatrix} \]

(8.4.3)

where \( \theta_i \) is the \( i \)th column vector and \( \theta'_i \) is the \( i \)th row vector of \( \theta \). For simplicity, we shall assume throughout the chapter that the rank of \( X \) is \( k \).

### 8.4.1 Distribution of \( \theta \)

Consider the elements of the \( m \times m \) matrix \( S(\theta) = \{S_{ij}(\theta, \theta')\} \) of (8.2.4). When \( X_1 = \ldots = X_m = X \), we can write

\[ S_{ij}(\theta, \theta') = (y_j - X\theta')' (y_j - X\theta) \]

\[ = (y_j - X\theta)' (y_j - X\theta) + (\theta_i - \hat{\theta}_i)' X' X (\theta_i - \hat{\theta}_i) \]  

(8.4.4)

where \( \hat{\theta}_i = (X'X)^{-1}X'y_i \) is the least squares estimates of \( \theta_i, i = 1, \ldots, m. \)
Consequently,

$$S(\theta) = A + (\theta - \hat{\theta})' X'X(\theta - \hat{\theta}), \quad (8.4.5)$$

where $\hat{\theta}$ is the $k \times m$ matrix of least squares estimates

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_m \end{bmatrix}$$

and $A$ is the $m \times m$ matrix

$$A = \{a_{ij}\}$$

with

$$a_{ij} = (y_i - X \hat{\theta})' (y_j - X \hat{\theta}), \quad i, j = 1, \ldots, m, \quad (8.4.7)$$

that is, $A$ is proportional to the sample covariance matrix. For simplicity, we shall assume that $A$ is positive definite. From the general result in (8.2.23), the posterior distribution of $\theta$ is then

$$p(\theta \mid y) \propto |A + (\theta - \hat{\theta})' X'X(\theta - \hat{\theta})|^{-n/2}, \quad -\infty < \theta < \infty. \quad (8.4.8)$$

As mentioned earlier, when there is a single output ($m = 1$), (8.4.8) is in the form of a $k$-dimensional multivariate $t$ distribution. The distribution in (8.4.8) is a matrix-variate generalization of the $t$ distribution. It was first obtained by Kairnsagar (1960). A comprehensive discussion of its properties has been given by Dickey (1967b).

### 8.4.2 Posterior Distribution of the Means from a $m$-dimensional Normal

**Distribution**

In the case $k = 1$ where each $\theta_i$ consists of a single element and $X$ is a $n \times 1$ vector of ones, expression (8.4.8) is the joint posterior distribution of the $m$ means when sampling from an $m$-dimensional multivariate Normal distribution $N_m(\theta, \Sigma)$. In this case

$$\theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_m), \quad \hat{\theta} = (\bar{y}_1, \ldots, \bar{y}_i, \ldots, \bar{y}_m),$$

$$X'X = n \quad \text{and} \quad a_{ij} = \sum_{a=1}^{n} (y_{ai} - \bar{y}_i)(y_{aj} - \bar{y}_j), \quad (8.4.9)$$

where

$$\bar{y}_i = \frac{1}{n} \sum_{a=1}^{n} y_{ai}$$
The posterior distribution of $\theta$ can be written

$$p(\theta | y) \propto |A + n(\theta - \hat{\theta})' (\theta - \hat{\theta})|^{-n/2}$$

$$\propto |I + n A^{-1} (\theta - \hat{\theta})' (\theta - \hat{\theta})|^{-n/2}, \quad -\infty < \theta < \infty. \quad (8.4.10)$$

We now make use of the fundamental identity

$$|I_k - P Q| = |I_l - Q P|,$$  

$$\text{where } I_k \text{ and } I_l \text{ are, respectively, } k \times k \text{ and a } l \times l \text{ identity matrices, } P \text{ is a } k \times l \text{ matrix and } Q \text{ is a } l \times k \text{ matrix. Noting that } (\theta - \hat{\theta}) \text{ is a } 1 \times m \text{ vector, we immediately obtain}$$

$$p(\theta | y) \propto |1 + n(\theta - \hat{\theta}) A^{-1} (\theta - \hat{\theta})' A^{-1}|^{-n/2}, \quad -\infty < \theta_i < \infty, \quad i = 1, \ldots, m,$$

$$\quad (8.4.12)$$

which is a $m$-dimensional $t_m [\theta', r^{-1} (n - m)^{-1} A, n - m]$ distribution, a result first published by Geisser and Cornfeld (1963). Thus, by comparing (8.4.12) with (8.3.13), we see that both when $m = 1$ and when $k = 1$, the distribution in (8.4.8) can be put in the multivariate $t$ form.

### 8.4.3 Some Properties of the Posterior Matric-variate $t$ Distribution of $\theta$

When neither $m$ nor $k$ is equal to one, it is not possible to express the distribution of $\theta$ as a multivariate $t$ distribution. As we have mentioned, the distribution in (8.4.8) can be thought of as a matric-variate extension of the $t$ distribution. We now discuss some properties of this distribution.

#### Two equivalent Representations of the Distribution of $\theta$

It is shown in Appendix A8.3 that, for $\nu > 0$,

$$\int_{-\infty < \theta < \infty} |X'X|^{-m/2} \prod_{i=1}^k (\nu + \nu_i)^{-1/2} \Gamma(\nu_i) d\theta$$

$$= c(m, k, \nu) \left|X'X\right|^{-m/2} |A|^{-k/2}, \quad (8.4.13)$$

where

$$\nu = n - (k + m) + 1,$$

$$c(m, k, \nu) = \frac{\prod_{i=1}^k \Gamma(\nu_i)}{\Gamma(\nu + k + m - 1)}, \quad (8.4.14)$$

and $\Gamma_p(\theta)$ is the generalized Gamma function defined in (8.2.22). Thus,

$$p(\theta | y) = [c(m, k, \nu)]^{-1} |X'X|^{m/2} |A|^{-k/2} |X'X - \hat{\theta}'A^{-1}(\theta - \hat{\theta})X'X(\theta - \hat{\theta})|^{-1/2}, \quad -\infty < \theta < \infty. \quad (8.4.15)$$
and we shall say that the \( k \times m \) matrix of parameters \( \theta \) is distributed as 
\[
\mathcal{I}_{\text{ml}}[\tilde{\theta}, (X'X)^{-1}, A, \nu].
\]
Note that by applying the identity (8.4.11), we can write
\[
[I_n + A^{-1}(\theta - \bar{\theta})'X'X(\theta - \bar{\theta})] = [I_k + (X'X)(\theta - \bar{\theta})A^{-1}(\theta - \bar{\theta})]
\]
so that in terms of the \( m \times k \) matrix \( \theta' \) the roles of \( m \) and \( k \) on the one hand and of the matrices \( (X'X)^{-1} \) and \( A \) on the other are simultaneously interchanged. Thus, we may conclude that if
\[
\theta \sim \mathcal{I}_{\text{ml}}[\theta, (X'X)^{-1}, A, \nu],
\]
then
\[
\theta' \sim \mathcal{I}_{\text{ml}}[\theta', A, (X'X)^{-1}, \nu].
\]

It follows from these two equivalent representations that
\[
c(k, m, \nu) = c(m, k, \nu),
\]
that is,
\[
\frac{\Gamma_k \left( \frac{1}{2}(\nu + k - 1) \right)}{\Gamma_k \left( \frac{1}{2}(\nu + k + m - 1) \right)} = \frac{\Gamma_m \left( \frac{1}{2}(\nu + m - 1) \right)}{\Gamma_m \left( \frac{1}{2}(\nu + k + m - 1) \right)}.
\]

**Marginal and Conditional Distributions of Subsets of Columns of \( \theta \)**

We now show that the marginal and conditional distributions of subsets of the \( m \) columns of \( \theta \) are also matrix-variate \( t \) distributions. Let \( m = m_1 + m_2 \) and partition the matrices \( \theta, \tilde{\theta}, \) and \( A \) into
\[
\theta = [\theta_{1*}^{m_1} | \theta_{2*}^{m_2}]_k, \quad \tilde{\theta} = [\tilde{\theta}_{1*}^{m_1} | \tilde{\theta}_{2*}^{m_2}]_k,
\]
\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}_{m_1 \times m_2}.
\]

Then:

a) conditional on \( \theta_{1*} \), the subset \( \theta_{2*} \) is distributed as
\[
\theta_{2*} \sim \mathcal{I}_{\text{ml}}[(\theta_{2*}, H^{-1}A_{22*}, \nu + m_1)],
\]
where
\[
H^{-1} = (X'X)^{-1} + (\theta_{1*} - \bar{\theta}_{1*})A_{11}^{-1}(\theta_{1*} - \bar{\theta}_{1*})',
\]
\[
\tilde{\theta}_{2*} = \bar{\theta}_{2*} + (\theta_{1*} - \bar{\theta}_{1*})A_{11}^{-1}A_{12}.
\]
\[
A_{22*} = A_{22} - A_{21}A_{11}^{-1}A_{12}.
\]

b) \( \theta_{1*} \) is distributed as
\[
\theta_{1*} \sim \mathcal{I}_{\text{ml}}[\tilde{\theta}_{1*}, (X'X)^{-1}, A_{11}, \nu].
\]

To prove these results, we can write
\[
(\theta - \tilde{\theta})A^{-1}(\theta - \tilde{\theta})' = (\theta_{1*} - \bar{\theta}_{1*})A_{11}^{-1}(\theta_{1*} - \bar{\theta}_{1*})' + (\theta_{2*} - \bar{\theta}_{2*})A_{22*}^{-1}(\theta_{2*} - \bar{\theta}_{2*})'.
\]
The determinant on the right-hand side of (8.4.16) can now be written

\[ h_x + (X'X)(\theta - \theta_0)\Lambda^{-1} (\theta - \theta_0)' = |I_k + (X'X)(\theta_{1s} - \bar{\theta}_{1s})\Lambda_{11}^{-1} (\theta_{1s} - \bar{\theta}_{1s})'| \]
\[ \times |I_k + \bar{H}(\theta_{2s} - \bar{\theta}_{2s})\Lambda_{22}^{-1} (\theta_{2s} - \bar{\theta}_{2s})'|. \]  

(8.4.23)

Substituting (8.4.23) into (8.4.15), we see that, given \( \theta_{1s} \), the conditional distribution of \( \theta_{2s} \) is

\[ p(\theta_{2s} | \theta_{1s}, y) \propto |I_k + \bar{H}(\theta_{2s} - \bar{\theta}_{2s})\Lambda_{22}^{-1} (\theta_{2s} - \bar{\theta}_{2s})'|^{-\frac{1}{2}(v+k+n-1)} \]
\[ \times |I_m + A_{11}^{-1} (\theta_{1s} - \bar{\theta}_{1s})'X'X(\theta_{1s} - \bar{\theta}_{1s})|^{-\frac{1}{2}(v+m-1)}. \]  

(8.4.24)

From (8.4.13), the normalizing constant is

\[ c(m_1, k, v + m_1)\]  

(8.4.25)

Thus, given \( \theta_{1s} \), the \( k \times m_1 \) matrix of parameters \( \theta_{1s} \) is distributed as \( t_{m_1}(\bar{\theta}_{2s}, \bar{H}, \Lambda_{22}, k, v + m_1) \). For the marginal distribution of \( \theta_{1s} \), since

\[ p(\theta_{1s} | y) = \frac{p(\theta | y)}{p(\theta_{2s} | \theta_{1s}, y)}, \]

use of (8.4.23) through (8.4.25) yields

\[ p(\theta_{1s} | y) \propto |I_k|^{-m/2} |I_k + (X'X)(\theta_{1s} - \bar{\theta}_{1s})\Lambda_{11}^{-1} (\theta_{1s} - \bar{\theta}_{1s})'|^{-\frac{1}{2}(v+k+n-1)} \]
\[ \times |I_m + A_{11}^{-1} (\theta_{1s} - \bar{\theta}_{1s})'X'X(\theta_{1s} - \bar{\theta}_{1s})|^{-\frac{1}{2}(v+m-1)}. \]  

(8.4.26)

That is, the \( k \times m_1 \) matrix of parameters \( \theta_{1s} \) is distributed as

\[ t_{m_1}(\bar{\theta}_{1s}, (X'X)^{-1}, A_{11}, v). \]

Marginal Distribution of a Particular Column of \( \theta \)

In particular, by setting \( m_1 = 1 \) in (8.4.21), the marginal distribution of \( \theta_{1s} = \theta_s \) is the \( k \)-dimensional multivariate \( t \) distribution

\[ p(\theta_s | y) \propto [1 + a_{11}^{-1} (\theta_s - \bar{\theta}_s)'X'X(\theta_s - \bar{\theta}_s)]^{-\frac{v+k}{2}}, \quad -\infty < \theta_s < \infty, \]  

(8.4.27)

where \( a_{11} = A_{11} \) is now a scalar, that is, \( \theta_s \sim t_k[\bar{\theta}_s, v^{-1}a_{11} (X'X)^{-1}, v] \).

By mere relabeling we may conclude that the marginal distribution of the \( i \)th column of \( \theta \) in (8.4.2) is

\[ p(\theta_i | y) \propto [1 + a_{ii}^{-1} (\theta_i - \bar{\theta}_i)'X'X(\theta_i - \bar{\theta}_i)]^{-\frac{v+k}{2}}, \quad -\infty < \theta_i < \infty. \]  

(8.4.28)

It will be noted that this distribution is identical to that obtained in (8.3.13) when only a single output was considered, except that in (8.3.13)
\[ v = n - k \], but in (8.4.28) \( v = n - k - (m - 1) \). In a certain sense, the reduction of the degrees of freedom by \( m - 1 \) is not surprising. In adopting the multivariate framework, \( m(m - 1)/2 \) additional parameters \( \sigma_{ij} (i \neq j) \) are introduced. A part of the information from the sample is therefore utilized to estimate these parameters and \( (m - 1) \) of them \( (\sigma_{11}, \ldots, \sigma_{(i-1)(i-1)}, \sigma_{(i+1)(i+1)}, \ldots, \sigma_{mm}) \) are connected with \( y_i \). We may say that ‘one degree of freedom is lost’ for each of the \( (m - 1) \) additional parameters.

On the other hand, it is somewhat puzzling that if we ignored the multivariate structure of the problem and treated \( y_i \) as the output of a univariate response, then on the basis of the noninformative prior \( p(\theta, \sigma_{ij}) \propto \sigma_{ij}^{-4} \), we would obtain a posterior multivariate \( t \) distribution for \( \theta \) with \( (m - 1) \) additional degrees of freedom. This would seem to imply that, by ignoring the information from the other \( (m - 1) \) responses \( y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m \), more precise inference about \( \theta \) could be made than when all the \( m \) responses were considered jointly. This phenomenon is related to the “paradox” pointed out by Dempster (1963) and the criticisms of the prior in (8.2.14) by Stone (1964).

The above implication is admittedly perplexing, and further research is needed to clarify the situation. We feel, however, that the multivariate results presented in this chapter are of considerable interest and certainly provide a sensible basis for inference in the common practical situation when \( (n - k) \) is large relative to \( m \).

**Distribution of \( \theta \) Expressed as a Product of Multivariate \( t \) Distributions**

We note that for the partition in (8.4.19), if we set \( m_1 = m - 1 \) and \( m_2 = 1 \), then from (8.4.20) the conditional distribution of \( \theta_{2*} = \theta_m \), given \( \theta_{1*} = [\theta_1, \ldots, \theta_{m-1}] \), is

\[
\theta_m \sim t_k \left( \bar{\theta}_m, (v + m - 1)^{-1} a_{m-1 \ldots (m-1)} H^{-1}, v + m - 1 \right),
\]

where

\[
a_{m-1 \ldots (m-1)} = A_{22.1}.
\]

From the marginal distribution of \( \theta_{1*} = [\theta_1, \ldots, \theta_{m-1}] \) in (8.4.21) if we partition \( \theta_{1*} \) into \( [\theta_{1*} : \theta_{m-1}] \) where \( \theta_{1*} = [\theta_1, \ldots, \theta_{m-1}] \), it is clear that the conditional distribution of \( \theta_{m-1} \), given \( \theta_{1*} \), is again a \( k \)-dimensional multivariate \( t \) distribution. It follows by repeating the process \( m - 1 \) times that, if we express \( p(\theta | y) \) as the product

\[
p(\theta | y) = p(\theta_1 | y) p(\theta_2 | \theta_1, y) \cdots p(\theta_m | \theta_1, \ldots, \theta_{m-1}, y),
\]

then each factor on the right-hand side is a \( k \)-dimensional multivariate \( t \) distribution.

**Marginal and Conditional Distributions of Rows of \( \theta \)**

Results very similar to those given above for column decomposition of \( \theta \) can
now be obtained for the rows of $\theta$. Consider the partitions
\[
\theta' = \begin{bmatrix} \theta_{(1)\star} & \theta_{(2)\star} \end{bmatrix}_m, \quad \hat{\theta}' = \begin{bmatrix} \hat{\theta}_{(1)\star} & \hat{\theta}_{(2)\star} \end{bmatrix}_m
\]
(8.4.31)
\[
(X'X)^{-1} = C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}^{k_1, k_2}, \quad k_1 + k_2 = k,
\]
where it is to be remembered that $\theta_{(1)\star}$ are the first $k_1$ rows and $\theta_{(2)\star}$ the remaining $k_2$ rows of $\theta$. Since the $m \times k$ matrix $\theta'$ is distributed as $t_{m[k', A, (XX')^{-1}, v]}$, it can be readily shown that

a) given $\theta_{(1)\star}$,
\[
\theta_{(2)\star} \sim t_{m[k'(G, C_{22}, v + k_1)]}, \quad (8.4.32)
\]

where
\[
C_{22} = C_{22} - C_{21} C_{11}^{-1} C_{12},
\]
\[
\hat{\theta}_{(2)\star} = \theta_{(2)\star} + (\theta_{(1)\star} - \hat{\theta}_{(1)\star}) C_{11}^{-1} C_{12},
\]
and
\[
G = A + (\theta_{(1)\star} - \hat{\theta}_{(1)\star}) C_{11}^{-1} (\theta_{(1)\star} - \hat{\theta}_{(1)\star})'.
\]

b) Marginally,
\[
\theta_{(1)\star} \sim t_{m[k, A, C_{11}, v]} \quad (8.4.33)
\]
or equivalently,
\[
\theta_{(1)\star} \sim t_{m[k, A, C_{11}, v]} \quad (8.4.34)
\]
where $c_{gg}$ is the $(gg)$th element of $C$.

c) The $g$th row of $\theta$, $\theta_{(g)\star}$, is distributed as
\[
\theta_{(g)\star} \sim t_{m[k'(G, C_{11}, A, v)]}. \quad (8.4.35)
\]

d) The distribution of $\theta$ can alternatively be expressed as the product
\[
p(\theta \mid y) = p(\theta_{(1)} \mid y) p(\theta_{(2)} \mid \theta_{(1)}, y) \cdots p(\theta_{(k)} \mid \theta_{(1)}, \ldots, \theta_{(k-1)}, y), \quad (8.4.35)
\]
where, parallel to (8.4.30), each factor on the right-hand side is an $m$-dimensional multivariate $t$ distribution.

Comparing expression (8.4.35) with the result in (8.4.12), we see that, as was the case with a column vector of $\theta$, the two distributions are of the same form except for the difference in the "degrees of freedom." They now differ by $(k - 1)$, simply because an additional $(k - 1)$ "input variables" are included in the model.

**Marginal Distribution of a Block of Elements of $\theta$**

Finally consider now the partitions
\[
\theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}^{k_1, k_2}, \quad \hat{\theta} = \begin{bmatrix} \hat{\theta}_{11} & \hat{\theta}_{12} \\ \hat{\theta}_{21} & \hat{\theta}_{22} \end{bmatrix}^{k_1, k_2}, \quad (8.4.36)
\]
It follows from (8.4.21) and (8.4.33) that the \( k \times m_1 \) matrix of parameters is distributed as

\[
\theta_{11} \sim t_{k, m_1} (\hat{\theta}_{11}, C_{11}, A_{11}, \nu).
\]

The marginal distributions of \( \theta_{12}, \theta_{21}, \theta_{22} \), and indeed that of any block elements of \( \theta \) can be similarly obtained, and are left to the reader.

In the above we have shown that the marginal and the conditional distributions of certain subsets of \( \theta \) are of the matrix-variate \( t \) form. This, however, not true in general. For example, one can show that neither the marginal distribution of \( (\theta_{11}, \theta_{21}) \) nor the conditional distribution of \( (\theta_{12}, \theta_{21}) \) given \( (\theta_{11}, \theta_{22}) \) is a matrix-variate \( t \) distribution. The problem of obtaining explicit expressions for the marginal and the conditional distributions in general is quite complex, and certain special cases have recently been considered by Dr and Morales (1970) and Tiao, Tan, and Chang (1970).

**Means and Covariance Matrix of \( \theta \)**

From (8.4.28), the matrix of means of the posterior distribution of \( \theta \) is

\[
E(\theta) = \bar{\theta}
\]

and the covariance matrix of \( \theta_i \) is

\[
\text{Cov} (\theta_i) = \frac{a_{i2}}{v - 2} (X'X)^{-1}, \quad i = 1, \ldots, m.
\]

For the covariance matrix of \( \theta_i \) and \( \theta_j \), with no loss in generality we consider the case \( i = 1 \) and \( j = 2 \). Now

\[
E(\theta_1 - \hat{\theta}_1) (\theta_2 - \hat{\theta}_2)' = \underbrace{E(\theta_1 - \hat{\theta}_1) E(\theta_2 - \hat{\theta}_2)}_{\theta_2:|\theta_1}.
\]

If we set \( m_1 = 2 \) in (8.4.21) and perform a column decomposition of the \( k \) matrix \( \theta_{1*} = [\theta_1, \theta_2] \), it is then clear from (8.4.20) that

\[
E(\theta_2 - \hat{\theta}_2) = a_{12}' a_{12} (\theta_1 - \hat{\theta}_1)
\]

so that, as might be expected,

\[
E(\theta_1 - \hat{\theta}_1) (\theta_2 - \hat{\theta}_2)' = \frac{a_{12}}{v - 2} (X'X)^{-1}.
\]

Thus,

\[
E \begin{bmatrix} \theta_1 - \hat{\theta}_1 \theta_2 - \hat{\theta}_2 \end{bmatrix} [ (\theta_1 - \hat{\theta}_1)', (\theta_2 - \hat{\theta}_2) ] = \frac{1}{v - 2} \begin{bmatrix} a_{11} (X'X)^{-1}, a_{12} (X'X)^{-1} \\ a_{21} (X'X)^{-1}, a_{22} (X'X)^{-1} \end{bmatrix} \otimes (X'X)^{-1}
\]

\[
= \frac{1}{v - 2} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes (X'X)^{-1}
\]
where \( \otimes \) denotes the Kronecker product—see Appendix A8.4. In general, if we write the elements of \( \Theta \) and \( \bar{\Theta} \) as
\[
\Theta' = (\theta_1', \ldots, \theta_m'), \quad \bar{\Theta}' = (\bar{\theta}_1', \ldots, \bar{\theta}_m'),
\]
where \( \Theta \) and \( \bar{\Theta} \) are \( km \times 1 \) vectors, then
\[
\text{Cov}(\Theta') = E(\Theta' - \bar{\Theta})(\Theta' - \bar{\Theta})' = \frac{1}{v - 2} A \otimes (X'X)^{-1}.
\]
By a similar argument, if we write
\[
\Theta_0' = (\theta_1'(0), \ldots, \theta_m'(0)), \quad \bar{\Theta}_x' = (\bar{\theta}_1'(0), \ldots, \bar{\theta}_k'(0)),
\]
then
\[
\text{Cov}(\Theta_0') = \frac{1}{v - 2} (X'X)^{-1} \otimes A.
\]

**Linear Transformation of \( \Theta \)**

Let \( P \) be a \( k_1 \times k \) matrix of rank \( k_1 \) and \( Q \) be a \( m_1 \times m_1 \) \( (m_1 \leq m) \) matrix of rank \( m_1 \). Suppose \( \phi \) is the \( k_1 \times m_1 \) matrix of random variables obtained from the linear transformation
\[
\phi = P \Theta Q.
\]
Then \( \phi \) is distributed as \( t_{v, m_1} [P \Theta Q, P(X'X)^{-1}P', Q'AQ, v] \). The proof is left as an exercise for the reader.

**Asymptotic Distribution of \( \Theta \)**

When \( v \) tends to infinity, the distribution of \( \Theta \) approaches a \( km \) dimensional multivariate Normal distribution,
\[
\lim_{v \to \infty} P(\Theta | y) = (\sqrt{2\pi})^{-km} |\Sigma|^{-k/2} |X|^{m/2} \times \exp \left[ -\frac{1}{2} \text{tr} \Sigma^{-1} (\Theta - \bar{\Theta})' XX(\Theta - \bar{\Theta}) \right], \quad -\infty < \Theta < \infty,
\]
where
\[
\Sigma = v^{-1} A,
\]
and we shall say that, asymptotically, \( \Theta \sim N_{km} [\bar{\Theta}, \Sigma \otimes (X'X)^{-1}] \).

To see this, in (8.4.15) let:
\[
Q = v A^{-1} (\Theta - \bar{\Theta})' XX(\Theta - \bar{\Theta}),
\]
\[
= \Sigma^{-1} (\Theta - \bar{\Theta})' XX(\Theta - \bar{\Theta}).
\]
Then, we may write
\[
|I_m + v^{-1} Q| = \prod_{i=1}^m (1 + v^{-1} l_i)
\]
(8.4.46)
where \((\lambda_1, \ldots, \lambda_m)\) are the latent roots of \(Q\). Thus, as \(v \to \infty\)
\[
\lim_{v \to \infty} |I_m + v^{-1}Q|^{-\frac{1}{2}(v + k + m - 1)} = \exp \left( -\frac{1}{2} \sum_{i=1}^{m} \lambda_i \right)
= \exp (-\frac{1}{2} \text{tr} \; Q).
\]
Since
\[
\text{tr} \; Q = (\Theta - \hat{\Theta})' \hat{\Sigma}^{-1} \otimes (X'X)(\Theta - \hat{\Theta}) \quad (8.4.47^*),
\]
where \(\Theta\) and \(\hat{\Theta}\) are the \(km \times 1\) vectors defined in (8.4.42a), and noting that
\[
|\hat{\Sigma}^{-1} \otimes (X'X)| = |\hat{\Sigma}^{-1}|^{k/2} |X'X|^{m/2},
\]
the desired result follows at once.

It follows that
\[
E(\Theta) = \hat{\Theta}, \quad \text{Cov}(\Theta) = \hat{\Sigma} \otimes (X'X)^{-1} \quad (8.4.48a)
\]
or, alternatively,
\[
E(\Theta_a) = \hat{\Theta}_a, \quad \text{Cov}(\Theta_a) = (X'X)^{-1} \otimes \hat{\Sigma} \quad (8.4.48b)
\]
where \((\Theta, \hat{\Theta})\) and \((\Theta_a, \hat{\Theta}_a)\) are defined in (8.4.42a) and (8.4.44a), respectively.

### 8.4.4 H.P.D. Regions of \(\Theta\)

Expressions (8.4.28) and (8.4.34) allow us to make inferences about a specific column or row of \(\Theta\). Using properties of the multivariate \(t\) distribution, H.P.D. regions of the elements of a row or a column can be easily determined.

We now discuss a procedure for the complete set of parameters \(\Theta\), which makes it possible to decide whether a general point \(\Theta = \theta_0\) is or is not included in an H.P.D. region of approximate content \((1 - \alpha)\).

It is seen in expression (8.4.15) that the posterior distribution of \(\Theta\) is a monotonic increasing function of the quantity \(U(\Theta)\), where
\[
U(\Theta) = \frac{|A|}{|A + (\Theta - \hat{\Theta})'X'X(\Theta - \hat{\Theta})^-1|} = |I_m A + A^{-1} (\Theta - \hat{\Theta})'X'X(\Theta - \hat{\Theta})|^{-1} \quad (8.4.49)
\]
Consequently, the parameter point \(\Theta = \theta_0\) lies inside the \((1 - \alpha)\) H.P.D. region if and only if
\[
\Pr \{U(\Theta) > U(\Theta_0) | y\} \leq (1 - \alpha), \quad (8.4.50)
\]

### 8.4.5 Distribution of \(U(\Theta)\)

To obtain the posterior distribution of \(U(\Theta)\) so that (8.4.50) can be calculated we first derive the moments of \(U(\Theta)\). Applying the integral identity (8.4.13) th
4th moment of \( U(\theta) \) is found to be

\[
E[U(\theta)^4 | y] = \frac{c(m, k, v + 2h)}{c(m, k, v)}
\]

\[= \prod_{s=1}^{m} \frac{\Gamma[\frac{1}{2}(v - 1 + s) + h] \Gamma[\frac{1}{2}(v - 1 + k + s)]}{\Gamma[\frac{1}{2}(v - 1 + s)] \Gamma[\frac{1}{2}(v - 1 + k + s) + h]} \]  

(8.4.51)

From (8.4.46) and (8.4.49) it follows that \( U = U(\theta) \) is a random variable defined on the interval (0, 1) so that the distribution of \( U \) is uniquely determined by its moments. Further, expression (8.4.51) shows that distribution of \( U \) is a function of \((m, k, v)\). Adopting the symbol \( U_{m,k,v} \) to mean a random variable whose probability distribution is that implied by the moments in (8.4.51), we have the following general result.

**Theorem 8.4.1** Let \( \theta \) be a \( k \times m \) matrix of constants, \( X'X \) and \( A \) be, respectively, a \( k \times k \) and a \( m \times m \) positive definite symmetric matrix of constants, and \( v > 0 \). If the \( k \times m \) matrix of random variables

\[
\theta \sim t_m [(\bar{\theta}, (X'X)^{-1}, A, v],
\]

then

\[U(\theta) \sim U_{m,k,v}
\]

where

\[U(\theta) = |I_m + A^{-1} (\theta - \bar{\theta}) X'X (\theta - \bar{\theta})|^{-1}.
\]

As noted by Geisser (1965a), expression (8.4.51) correspond exactly to that for the 4th sampling moment of \( U(\theta) \) in the sampling theory framework when \( \theta \) are regarded as fixed and \( y \) random variables. Thus, the Bayesian probability \( P_{R}(U(\theta) > U(\theta_0) | y) \) is numerically equivalent to the significance level associated with the null hypothesis \( \theta = \theta_0 \) against the alternative \( \theta \neq \theta_0 \).

**Some Distributional Properties of \( U_{m,k,v} \).**

Following the development, for example, in Anderson (1958), we now discuss some properties of the distribution of \( U_{m,k,v} \). It will be noted that the notation \( U_{m,k,v} \) here is slightly different from the one used in Anderson. Specifically, in his notation, \( v \) is replaced by \( v + m - 1 \).

a) Since from (8.4.18) \( c(k, m, v) = c(k, m, v) \), the 4th moment in (8.4.51) can be alternatively expressed as

\[
E[U(\theta)^4 | y] = \prod_{s=1}^{m} \frac{\Gamma[\frac{1}{2}(v - 1 + s) + h] \Gamma[\frac{1}{2}(v - 1 + k + s)]}{\Gamma[\frac{1}{2}(v - 1 + s)] \Gamma[\frac{1}{2}(v - 1 + k + s) + h]}.
\]

(8.4.52)

By comparing (8.4.51) with (8.4.52), we see that the roles played by \( m \) and \( k \) can be interchanged. That is, the distribution of \( U_{m,k,v} \) is identical to that of \( U_{k,m,v} \). In other words, the distribution of \( U = U(\theta) \) arising from a multivariate model with \( m \) output variables and \( k \) regression coefficients for each output is identical to that from a
multivariate model with \( k \) output variables and \( m \) regression coefficients for each output. With no loss in generality, we shall proceed with the \( m \)-output model, i.e., the \( U(m,k) \) distribution.

b) Now (8.4.51) can be written

\[
E(U^k | y) = \prod_{s=1}^{m} B\left( \frac{v - 1 + s}{2} + h, \frac{k}{2} \right) \left/ B\left( \frac{v - 1 + s}{2}, \frac{k}{2} \right) \right.,
\]

where \( B(p,q) \) is the complete beta function. The right-hand side is the \( k \)th moment of the product of \( m \) independent variables \( x_1, \ldots, x_m \) having beta distributions with parameters

\[
\left( \frac{v - 1 + s}{2}, \frac{k}{2} \right), \quad s = 1, \ldots, m.
\]

It follows that \( U \) is distributed as the product \( x_1 \cdots x_m \).

c) Suppose \( m \) is even. Then we can write (8.4.53) as

\[
E(U^k | y) = \prod_{t=1}^{m/2} B\left[ \frac{v + 2(t - 1)}{2} + h, \frac{k}{2} \right] \left/ B\left[ \frac{v + 2(t - 1)}{2}, \frac{k}{2} \right] \right. \cdot B\left[ \frac{v - 1 + 2t}{2} + h, \frac{k}{2} \right] \left/ B\left[ \frac{v - 1 + 2t}{2}, \frac{k}{2} \right] \right.,
\]

Using the duplication formula

\[
\Gamma(p + 1/2) \Gamma(p) = \frac{\sqrt{\pi} \Gamma(2p)}{2^{2p-1}},
\]

so that

\[
B(p + 1/2, q) B(p, q) = 2^{2p} B(2p, 2q) B(q, q),
\]

we obtain

\[
E(U^k | y) = \prod_{t=1}^{m/2} B\left[ \frac{v + 2(t - 1) + h}{2}, k \right] \left/ B\left[ \frac{v + 2(t - 1)}{2}, k \right] \right. \cdot \left/ B\left[ \frac{v - 1 + 2t}{2} + h, k \right] \right. \left/ B\left[ \frac{v - 1 + 2t}{2}, k \right] \right.,
\]

where \( z_1, \ldots, z_{m/2} \) are \( m/2 \) independent random variables having beta distributions with parameters \([v + 2(t - 1) + h, k], \ t = 1, \ldots, m/2\), respectively. Thus, \( U \) is distributed as the product \( z_1^2 \cdots z_{m/2}^2 \).

The Case \( m = 1 \).

When \( m = 1 \) so that \( v = n - k \), it follows from (8.4.53) that, \( U \) has the beta distribution with parameters \( [(n - k)/2, k/2] \) so that the quantity \((n - k)(1 - U)/(kU)\) is distributed as \( F \) with \((k, n - k)\) degrees of freedom. This result is of course to be
expected, since for \( m = 1 \), we have \(|A| = (n - k)s^2\), where
\[
s^2 = (n - k)^{-1} (y - \bar{y})' (y - \bar{y}),
\]
so that
\[
\left( \frac{1 - U}{U} \right) \left( \frac{n - k}{k} \right) = \frac{(\theta - \bar{\theta})'X'X(\theta - \bar{\theta})}{ks^2}
\]
which, from (2.7.21), has the \( F \) distribution with \((k, n - k)\) degrees of freedom.

The Case \( m = 2 \).

When \( m = 2, v = n - k - 1 \) and from (8.4.57) \( U^{1/2} \) is distributed as a beta variable with parameters \((n - k - 1, 1)\). Thus, the quantity
\[
\left( \frac{1 - U^{1/2}}{U^{1/2}} \right) \left( \frac{n - k - 1}{k} \right) = \left( I_2 + \Lambda^{-1} (\theta - \bar{\theta})'X'X(\theta - \bar{\theta})^{-1/2} \right) \left( \frac{n - k - 1}{k} \right)
\]
has the \( F \) distribution with \([2k, 2(n - k - 1)]\) degrees of freedom.

8.4.6 An Approximation to the Distribution of \( U \) for General \( m \)

For \( m \geq 3 \), the exact distribution of \( U \) is complicated, see e.g. Schatzoff (1966) and Pillai and Gupta (1969). We now give an approximation method following Bartlett (1938) and Box (1949). In expression (8.4.51), we make the substitutions
\[
M = -\phi v \log U, \quad t = -h(\phi v),
\]
\[
x = \frac{1}{2}v, \quad a_s = \frac{1}{2}(s + k - 1) \quad \text{and} \quad b_s = \frac{1}{2}(s - 1),
\]
where \( \phi \) is some arbitrary positive number, so that
\[
E(U^h | y) = E(\phi^M | y)
\]
\[
= \prod_{s=1}^{m} \frac{\Gamma(x + a_s)}{\Gamma(x + b_s)} \left[ \frac{\phi x(1 - 2t) + x(1 - \phi) + b_s}{\Gamma(x(1 - 2t) + x(1 - \phi) + a_s)} \right].
\]

In terms of the random variable \( M \), (8.4.61) is then its moment-generating function. Taking logarithms and employing Stirling's series (see Appendix A2.2), we obtain the cumulant generating function of \( M \) as
\[
\kappa_M(t) = -\frac{mk}{2} \log (1 - 2t) - \sum_{r=1}^{m} \omega_r [(1 - 2t)^{-r} - 1],
\]
where
\[
\omega_r = \frac{(-1)^r}{r(r + 1)(\phi x)^r} \sum_{s=1}^{m} \left[ B_{r+1}[x(1 - \phi) + b_s] - B_{r+1}[x(1 - \phi) + a_s] \right]
\]
and $B_r(z)$ is the Bernoulli polynomial of degree $r$ and order one. The asymptotic expansion in (8.4.62) is valid provided $\phi$ is so chosen that $x(1 - \phi)$ is bounded. In this case, $\omega_1$ is of order $O[(\phi x)^{-1}]$ in magnitude.

The series in (8.4.62) is of the same type as the one obtains in (2.12.1) for the comparison of the spread of $k$ Normal populations. In particular, the distribution of $M = -\phi v \log U$ can be expressed as a weighted series of $\chi^2$ densities: the leading term having $mk$ degrees of freedom.

The Choice of $\phi$

It follows that, if we take the leading term alone, then to order $O[(\phi x)^{-1}]$, $O[(\phi^{1/2}v)^{-1}]$ the quantity

$$M = -\phi v \log U$$  \hspace{1cm} (8.4.6)

is distributed as $\chi^2_{mk}$ provided $\frac{1}{2}v(1 - \phi)$ is bounded. In particular, if $v$ is taken $\phi = 1$, we then have that $M = -v \log U$ is distributed approximate as $\chi^2_{mk}$.

For moderate values of $v$, the accuracy of the $\chi^2$ approximation can be improved by suitably choosing $\phi$ so that $\omega_1 = 0$. This is because when $\omega_1 = 0$, the quantity $M$ will be distributed as $\chi^2_{mk}$ to order $O[(\phi^{1/2}v)^{-1}]$. Using the fact that

$$B_r(z) = z^r - z + \frac{1}{r}$$  \hspace{1cm} (8.4.6)

it is straightforward to verify that for $\omega_1 = 0$, we require

$$\phi = 1 + \frac{1}{2v} (m + k - 3).$$  \hspace{1cm} (8.4.6)

This choice of $\phi$ gives very close approximations in practice. An example with $v = 9$, $m = k = 2$ will be given later in Section 8.4.8 to compare it approximation with the exact distribution.

It follows from the above discussion that to order $O[(\phi^{1/2}v)^{-2}]$,

$$\Pr \{ U(\theta) > U(\theta_0) | y \} \approx \Pr \{ \chi^2_{mk} < -\phi v \log U(\theta_0) \}$$  \hspace{1cm} (8.4.6)

with

$$\phi = 1 + \frac{1}{2v} (m + k - 3),$$

$$\log U(\theta_0) = -\log |I_n + \theta_0^{-1} (\theta_0 - \theta)' X' X (\theta_0 - \theta)|$$

which can be employed to decide whether the parameter point $\theta = \theta_0$ lies approximately inside or outside the $(1 - \alpha)$ H.P.D. region.

8.4.7 Inferences about a General Parameter Point of a Block Submatrix of $\theta$

In the above, we have discussed inference procedures for (a) a specific column of $\theta$, (b) a specific row of $\theta$, and (c) a parameter point $\theta_0$ for the complete set of
n some problems, we may be interested in making inferences about the parameters belonging to a certain block submatrix of \( \Theta \). Without loss of generality we consider only the problem for the \( k_1 \times m_1 \) matrix \( \Theta_{11} \) defined in (8.4.36). From (8.4.37) and Theorem 8.4.1 (on page 449), it follows that the quantity

\[
U(\Theta_{11}) = \frac{1}{m_1} + A_{11}^{-1} (\Theta_{11} - \bar{\Theta}_{11})' C_{11}^{-1} (\Theta_{11} - \bar{\Theta}_{11})^{-1}
\]

is distributed as \( U(m_1, k_1) \). This distribution would then allow us to decide whether a particular value of \( \Theta_{11} \) lay inside or outside a desired H.P.D. region. In particular, for \( m_1 > 2 \) and \( k_1 > 2 \), we may then make use of the approximation

\[
-\phi_n \log U(\Theta_{11}) \sim \chi^2_{m_1, k_1},
\]

where

\[
\phi_n = 1 + \frac{1}{2} (m_1 + k_1 - 3),
\]

so that the parameter point \( \Theta_{11,0} \) lies inside the \( (1 - \alpha) \) H.P.D. region if and only if

\[
\text{Pr}(U(\Theta_{11}) > U(\Theta_{11,0}) \mid y) = \text{Pr}(\chi^2_{m_1, k_1} < -\phi_n \log U(\Theta_{11,0})) < (1 - \alpha).
\]

8.4.8 An Illustrative Example

An experiment was conducted to study the effect of temperature on the yield of the product \( y_1 \) and the by-product \( y_2 \) of a chemical process. Twelve runs were made at different temperature settings ranging from 161.3°F to 195.7°F. The data are given in Table 8.4.1.

The average temperature employed is \( T = 177.86 \). We suppose a model to be entertained whereby, over the range of temperature explored, the relationships between product yield and temperature and by-product yield and temperature were nearly linear so that to an adequate approximation

\[
E(y_{1u}) = \theta_{11} + \theta_{21} x_u,
\]

\[
E(y_{2u}) = \theta_{12} + \theta_{22} x_u, \quad u = 1, \ldots, 12,
\]

where \( x_u = (T_u - \bar{T})/100 \), the divisor 100 being introduced for convenience in calculation. The parameters \( \theta_{11} \) and \( \theta_{12} \) will thus determine the locations of the yield-temperature lines at the average temperature \( \bar{T} \) while \( \theta_{21} \) and \( \theta_{22} \) will represent the slopes of these lines. The experimental runs were set up independently and we should therefore expect experimental errors to be independent from run to run. However, in any particular run, we should expect the error in \( y_1 \) to be correlated with that in \( y_2 \) since slight aberrations in reaction
Table 8.4.1
Yield of product and by-product of a chemical process

<table>
<thead>
<tr>
<th>Temp, °F</th>
<th>Product $y_1$</th>
<th>By-product $y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>161.3</td>
<td>63.7</td>
<td>20.3</td>
</tr>
<tr>
<td>164.0</td>
<td>59.5</td>
<td>24.2</td>
</tr>
<tr>
<td>165.7</td>
<td>67.9</td>
<td>18.0</td>
</tr>
<tr>
<td>170.1</td>
<td>68.8</td>
<td>20.5</td>
</tr>
<tr>
<td>173.9</td>
<td>66.1</td>
<td>20.1</td>
</tr>
<tr>
<td>176.2</td>
<td>70.4</td>
<td>17.5</td>
</tr>
<tr>
<td>177.6</td>
<td>70.0</td>
<td>18.2</td>
</tr>
<tr>
<td>181.7</td>
<td>73.7</td>
<td>15.4</td>
</tr>
<tr>
<td>185.6</td>
<td>74.1</td>
<td>17.8</td>
</tr>
<tr>
<td>189.0</td>
<td>79.6</td>
<td>13.3</td>
</tr>
<tr>
<td>193.5</td>
<td>77.1</td>
<td>16.7</td>
</tr>
<tr>
<td>195.7</td>
<td>82.8</td>
<td>14.8</td>
</tr>
</tbody>
</table>

conditions or in analytical procedures could simultaneously affect observat-
of both product and by-product yields. Finally, then, the tentative model wi-

\[ y_{u1} = \theta_{11} x_{u1} + \theta_{21} x_{u2} + \varepsilon_{u1} \]

\[ y_{u2} = \theta_{12} x_{u1} + \theta_{22} x_{u2} + \varepsilon_{u2}, \]  

(8.4)

where $x_{u2} = x_u$ and $x_{u1} = 1$ is a dummy variable introduced to “carry”
parameters $\theta_{11}$ and $\theta_{12}$. It was supposed that $(\varepsilon_{u1}, \varepsilon_{u2})$ followed the bivar-
Normal distribution $N_2(\mathbf{0}, \Sigma)$.

Given this setup, we apply the results arrived at earlier in this section to m
inferences about the regression coefficients

\[
\theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \theta_{11} \\ \theta_{12} \end{bmatrix}
\]  

(8.4)

against the background of a noninformative reference prior distribution
$\theta$ and $\Sigma$.

The relevant sample quantities are summarized below:

\[ n = 12 \quad m = k = 2 \]

\[
X'X = \begin{bmatrix} 12 & 0 \\ 0 & 0.14546 \end{bmatrix} \quad C = (X'X)^{-1} = \begin{bmatrix} 0.0833 & 0 \\ 0 & 6.8750 \end{bmatrix}
\]  

(8.4)

\[
A = \begin{bmatrix} 61.4084 & -38.4823 \\ -38.4823 & 36.7369 \end{bmatrix} \quad A^{-1} = \begin{bmatrix} 0.0474 & 0.0496 \\ 0.0496 & 0.0792 \end{bmatrix}
\]  

(8.4)
and
\[
\begin{bmatrix}
71.1417 & 18.0666 \\
54.4355 & -20.0933
\end{bmatrix}
= \begin{bmatrix}
\hat{\theta}_1 \\
\hat{\theta}_2
\end{bmatrix} = \begin{bmatrix}
\theta_{(1)}^t \\
\theta_{(2)}^t
\end{bmatrix}.
\]

The fitted lines
\[
\hat{y}_1 = \hat{\theta}_{11} + \hat{\theta}_{12}(T - \bar{T}) \times 10^{-2}
\]
\[
\hat{y}_2 = \hat{\theta}_{21} + \hat{\theta}_{22}(T - \bar{T}) \times 10^{-2}
\]

together with the data are shown in Fig. 8.4.1. As explained earlier, in a real data analysis, we should pause at this point to critically examine the conditional inference by study of residuals. We shall here proceed with further analysis supposing that such checks have proved satisfactory.

![Fig. 8.4.1 Scatter diagram of the product $$y_1$$ and the by-product $$y_2$$ together with the best fitting lines.](image)

**Inferences about $$\theta_1 = (\theta_{11}, \theta_{21})^t$$**

When interest centers primarily on the parameters $$\theta_1$$ for the product $$y_1$$, we have from (8.4.28) that
\[
p(\theta_1 | y) \propto [61.4084 + (\theta_1 - \bar{\theta}_1)^t X'X(\theta_1 - \bar{\theta}_1)]^{-11/2} \tag{8.4.74}
\]
that is, a bivariate $$\mathcal{N}_2[\bar{\theta}_1, 6.825(X'X)^{-1}, 9]$$ distribution. Since the matrix $$X'X$$ is diagonal, $$\hat{\theta}_{11}$$ and $$\hat{\theta}_{21}$$ are uncorrelated (but of course not independent).
Figure 8.4.2a shows contours of the 50, 75 and 95 per cent H.P.D. region together with the mode \( \hat{\theta}_1 \), from which overall conclusions about \( \theta_1 \) may be drawn.

![Diagram](image)

**Fig. 8.4.2a** Contours of the posterior distribution of \( \theta_1 \), the parameters of the product straight line.

![Diagram](image)

**Fig. 8.4.2b** Contours of the posterior distribution of \( \theta_2 \), the parameters of the by-product straight line.

*Inferences about \( \theta_2 = (\theta_{12}, \theta_{22})' \)*

Similarly, from (8.4.28), the posterior distribution of \( \theta_2 \) for the by-product \( y_2 \) is

\[
p(\theta_2 | y) \propto [36.7369 + (\theta_2 - \bar{\theta}_2)' X'(X \theta_2 - \bar{\theta}_2)]^{-11/2},
\]

(8.4.7)
which is a $t_2[4, 0.08'(X'X)^{-1}, 9]$ distribution. Again, the parameters $\theta_{12}$ and $\theta_{11}$ are uncorrelated. The 50, 75 and 95 per cent H.P.D. contours together with the mode $\hat{\theta}_2$ for this distribution are shown in Fig. 8.4.2b. The contours have exactly the same shape and orientation as those in Fig. 8.4.2a because the same $X'X$ matrix is employed; the spread for $\theta_2$ is however smaller than that for $\theta_1$ since the sample variance from $y_2$ is less than that from $y_1$.

Inferences about $\theta_{(2)} = (\theta_{21}, \theta_{22})'$

In problems of the kind considered, interest often centers on $\theta_{(2)} = (\theta_{21}, \theta_{22})$ which measure respectively the slopes of the yield/temperature lines for the product and by-product. From (8.4.34), the posterior distribution of $\theta_{(2)}$ is

$$p(\theta_{(2)} | y) \propto [6.875 + (\theta_{(2)} - \hat{\theta}_{(2)})'A^{-1}(\theta_{(2)} - \hat{\theta}_{(2)})]^{-11/2},$$

(8.4.76)

that is, a $t_2[4, 0.764, 9]$ distribution. Figure 8.4.3 shows the 50, 75 and 95 per cent contours together with the mode $\hat{\theta}_{(2)}$. Also shown in the same figure are the marginal distributions $t(\hat{\theta}_{21}, 9.90, 9)$ and $t(\hat{\theta}_{22}, 28.05, 9)$ and the

Fig. 8.4.3 Contours of the posterior distribution of $\theta_{(2)}$ and the associated marginal distributions for the product and by-product data.
corresponding 95 per cent H.P.D. intervals for $\theta_{21}$ and $\theta_{22}$. Figure 8.4.2 summarizes, then, the information about the slopes $(\theta_{21}, \theta_{22})$ coming from the data on the basis of a noninformative reference prior. The parameters are negatively correlated; the correlation between $(\theta_{21}, \theta_{22})$ is in fact the sample correlation between the errors $(v_{21}, v_{22})$

$$
r_{12} = \frac{d_{12}}{(d_{11} d_{22})^{1/2}} = -0.81.
$$

(8.4.77)

It is clear from the figure that care must be taken to distinguish between individual and joint inferences about $(\theta_{21}, \theta_{22})$. It could be exceedingly misleading to make inferences from the individual H.P.D. intervals about the parameters jointly (see the discussion in Section 2.7).

**Joint Inferences about $\theta$**

To make overall inferences about the parameters $[\theta_1, \theta_2]$, we need to calculate the distribution of the quantity $U(\theta)$ defined in (8.4.49). For instance, suppose we wish to decide whether or not the parameter point

$$
\theta_0 = [\theta_{10}, \theta_{20}] = \begin{bmatrix} 70 & 17 \\ 65 & -30 \end{bmatrix}
$$

(8.4.78)

lies inside the 95 per cent H.P.D. region for $\theta$. We have

$$(\theta_0 - \theta) X'X(\theta_0 - \theta) = \begin{bmatrix} 31.8761 & -0.6108 \\ -0.6108 & 27.9272 \end{bmatrix}
$$

(8.4.79)

so that

$$
U(\theta_0) = \frac{|A|}{|A + (\theta_0 - \theta) X'X(\theta_0 - \theta)|} = \frac{775.0668}{4503.8878} = 0.1720,
$$

(8.4.80)

Since $m = 2$, we may use the result in (8.4.59) to calculate the exact probability that $U(\theta)$ exceeds $U(\theta_0)$. We obtain

$$
Pr \{U(\theta) > U(\theta_0) \mid y\} = 1 - I_{\chi^2_{\text{2}}} (9, 2) = 1 - 0.0022 = 0.9978.
$$

(8.4.81)

From (8.4.50), we conclude that the point $\theta_0$ lies outside the 95 per cent H.P.D. region of $\theta$. Note that while the point $\theta_0 = (\theta_{10}, \theta_{20})$ is excluded from the 95 per cent region, Figs. 8.4.2a,b show that both the points $\theta_{10}$ and $\theta_{20}$ are included in the corresponding marginal 95 per cent H.P.D. regions. This serves to illustrate once more the distinction between joint inferences and marginal inferences.

**Approximating Distribution of $U = U(\theta)$**

It is informative to compare the exact distribution of $U(\theta)$ with the
Some Aspects of the Distribution of \( \Sigma \)

approximation in (8.4.63) using the present example \((v = 9, m = k = 2)\). From (8.4.59), the exact distribution of \( U \) is found to be

\[
p(U) = \frac{1}{2B(2, 9)} U^{3/2} (1 - U^{1/2}), \quad 0 < U < 1. \tag{8.4.82}
\]

Using the approximation given in (8.4.63) to (8.4.65), we find

\[
\phi = \frac{1}{2}, \quad \phi v = 9.5, \quad M = -9.5 \log U \tag{8.4.83}
\]

and

\[
p(M) = \frac{1}{2} M \exp(-\frac{1}{2} M), \quad 0 < M < \infty.
\]

This implies that the distribution of \( U \) is approximately

\[
p(U) \approx (22.5625) (-\log U) U^{3.75}, \quad 0 < U < 1. \tag{8.4.84}
\]

Table 8.4.2 gives a specimen of the exact and the approximate densities of \( U \) calculated from (8.4.82) and (8.4.84). Although the sample size is only 10, the agreement is very close.

<table>
<thead>
<tr>
<th>( U )</th>
<th>Exact</th>
<th>Approximate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.00098</td>
<td>0.00089</td>
</tr>
<tr>
<td>0.10</td>
<td>0.00973</td>
<td>0.00924</td>
</tr>
<tr>
<td>0.20</td>
<td>0.08900</td>
<td>0.08688</td>
</tr>
<tr>
<td>0.30</td>
<td>0.30098</td>
<td>0.29731</td>
</tr>
<tr>
<td>0.40</td>
<td>0.66947</td>
<td>0.66548</td>
</tr>
<tr>
<td>0.50</td>
<td>1.16498</td>
<td>1.16239</td>
</tr>
<tr>
<td>0.60</td>
<td>1.69708</td>
<td>1.69718</td>
</tr>
<tr>
<td>0.70</td>
<td>2.10935</td>
<td>2.11244</td>
</tr>
<tr>
<td>0.80</td>
<td>2.17561</td>
<td>2.18045</td>
</tr>
<tr>
<td>0.90</td>
<td>1.59706</td>
<td>1.60130</td>
</tr>
<tr>
<td>0.95</td>
<td>0.95220</td>
<td>0.95478</td>
</tr>
</tbody>
</table>

8.5 SOME ASPECTS OF THE DISTRIBUTION OF \( \Sigma \) FOR THE CASE OF A COMMON DERIVATIVE MATRIX \( X \)

We discuss in this section certain results pertaining to the posterior distribution of the elements of the covariance matrix \( \Sigma = \{\sigma_{ij}\} \)\)

\* For the important problem of making inferences about the latent roots and vectors of \( \Sigma \) which is not discussed in this book, see Geisser (1965a) and Tiao and Fienberg (1969).
8.5.1 Joint Distribution of (\(\theta, \Sigma\))

When the \(X_i\)'s are common, the joint posterior distribution of \((\theta, \Sigma)\) in (8.2.1)
can be written

\[
p(\theta, \Sigma | y) \propto |\Sigma|^{-\frac{1}{2}(r+k+2m)} \exp \left\{ -\frac{1}{2} \text{tr} \Sigma^{-1} \left[ A + \left( \theta - \bar{\theta} \right) X'X(\theta - \bar{\theta}) \right] \right\},
\]

\[- \infty < \theta < \infty, \quad \Sigma > 0, \tag{8.5}\]

where \(r = n - (k + m) - 1\) and use is made of (8.4.4) and (8.4.5). The individual and joint inferences about \((\theta_{21}, \theta_{22})\). It could be that exceeding
distribution can be written as the product \(p(\theta | \Sigma, y) = p(\theta | \Sigma) p(\Sigma | y)\).

Conditional Distribution of \(\theta\) given \(\Sigma\)

Given \(\Sigma\), we have that

\[
p(\theta | \Sigma, y) \propto \exp \left\{ -\frac{1}{2} \text{tr} \Sigma^{-1} \left( \theta - \bar{\theta} \right) X'X(\theta - \bar{\theta}) \right\}, \quad - \infty < \theta < \infty, \tag{8.5}\]

which by comparison with (8.4.45), is the \(km\)-dimensional Normal distribution

\[
N_{mk}[\theta, \Sigma \otimes (X'X)^{-1}]. \tag{8.5}\]

Marginal Distribution of \(\Sigma\)

Thus, the marginal posterior distribution of \(\Sigma\) is

\[
p(\Sigma | y) \propto |\Sigma|^{-(k+m)} \exp \left\{ -\frac{1}{2} \text{tr} \Sigma^{-1} \right\}, \quad \Sigma > 0, \tag{8.5}\]

From the Jacobian in (8.2.13), the distribution of \(\Sigma^{-1}\) is

\[
p(\Sigma^{-1} | y) \propto |\Sigma^{-1}|^{\frac{1}{2}(k+m-1)} \exp \left\{ -\frac{1}{2} \text{tr} \Sigma^{-1} \right\}, \quad \Sigma^{-1} > 0, \tag{8.5}\]

which, by comparing with (8.2.20), is recognized as the Wishart distribution \(W_m(\Sigma^{-1}, \nu)\) provided \(\nu > 0\). The distribution of \(\Sigma\) in (8.5.4) may thus be called an \(m\)-dimensional "inverted" Wishart distribution with \(\nu\) degrees of freedom
and be denoted by \(W_m^{-1}(A, \nu)\). From (8.2.21) the normalizing constant for \(\Gamma\) distributions of \(\Sigma\) and \(\Sigma^{-1}\) is

\[
|A|^\frac{1}{2}(k+m-1) 2^\frac{1}{2}(k+m-1) \left[ \Gamma_m \left( \frac{\nu + m - 1}{2} \right) \right]^{-1}. \tag{8.5}\]

Note that when \(m = 1\) and \(\nu = n - k\) the distribution in (8.5.4) reduces to an inverted \(\chi^2\) distribution,

\[
p(\sigma_{11} | y) \propto \sigma_{11}^{-\frac{1}{2}(n-k+2)} \exp \left\{ -\frac{a_{11}}{2\sigma_{11}} \right\}, \quad \sigma_{11} > 0, \tag{8.5}\]

which is the posterior distribution of \(\sigma^2 = \sigma_{11}\) with data from a univariate multiple regression problem (see Section 2.7).

From the results in (8.5.1), (8.5.3), (8.5.4) and (8.4.15), we have the following useful theorem.
Theorem 8.5.1 Let \( \hat{\theta} \) be a \( k \times m \) matrix of constants, \( X'X \) and \( A \) be, respectively, \( k \times k \) and a \( m \times m \) positive definite symmetric matrix of constants, and \( v > 0 \). Then the joint distribution of the elements of the \( k \times m \) matrix \( \theta \) and the \( m \times m \) positive definite symmetric matrix \( \Sigma \) is

\[
\rho(\theta, \Sigma | y) \propto |\Sigma|^{-\frac{v+k+2m}{2}} \exp\left\{-\frac{1}{2} \text{tr} \Sigma^{-1}[A + (\theta - \hat{\theta})'X'X(\theta - \hat{\theta})]\right\}
\]

then, (a) given \( \Sigma \), \( \theta \sim N_m(\hat{\theta}, \Sigma \otimes (X'X)^{-1}) \), (b) \( \Sigma \sim W_m^{-1}(A, v) \) and (c) \( \theta \sim t_m[\hat{\theta}, (X'X)^{-1}, A, v] \).

8.5.2 Some Properties of the Distribution of \( \Sigma \)

Consider the partition

\[
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad m_1 + m_2 = m.
\]

(8.5.8)

We now proceed to obtain the posterior distribution of \( (\Sigma_{11}, \Omega, T) \), where

\[
\Omega = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}
\]

and

\[
T = \Sigma_{11}^{-1} \Sigma_{12}.
\]

(8.5.9)

It is to be remembered that \( \Sigma_{11} \) is the covariance matrix of the \( m_1 \) errors \( (\epsilon_{a1}, \ldots, \epsilon_{am_1}) \), \( \Omega \) and \( T \) are, respectively, the covariance matrix and the \( m_1 \times m_2 \) matrix of "regression coefficients" for the conditional distribution of the remaining \( m_2 \) errors \( (\epsilon_{a(m_1+1)}, \ldots, \epsilon_{am}) \) given \( (\epsilon_{a1}, \ldots, \epsilon_{am_1}) \). Denoting

\[
A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad A_{22 \cdot 1} = A_{22} - A_{21} A_{11}^{-1} A_{12}
\]

(8.5.10)

and

\[
\hat{T} = A_{11}^{-1} A_{12},
\]

it can be readily shown that

a) \( \Sigma_{11} \) is distributed independently of \( (T, \Omega) \),

b) \( \Sigma_{11} \sim W_{m_1}^{-1}(A_{11}, v) \),

c) \( \Omega \sim W_{m_2}^{-1}(A_{22 \cdot 1}, v + m_1) \),

d) \( T \sim t_{m_1 m_2}[(\hat{T}, A_{11}^{-1}, A_{22 \cdot 1}, v + m_1)], \)

\[
U(T) = [I_{m_2} + A_{22 \cdot 1} (T - \hat{T} A_{11} (T - \hat{T}))^{-1}]
\]

\[
\sim U_{(m_2, m_1, v + m_1)}.
\]

(8.5.14)
The above results may be verified as follows. Since $\Sigma$ is positive definite, we express the determinant and the inverse of $\Sigma$ as

$$|\Sigma| = |\Sigma_{11}| |\Omega|$$

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + M,$$

where

$$M = \begin{bmatrix} \Omega^{-1} & -T \Omega^{-1} \\ -\Omega^{-1} & \Omega^{-1} \end{bmatrix}.$$

Expression (8.5.15) may be verified by showing $\Sigma^{-1} \Sigma = I$. Thus, the distribution (8.5.4) can be written

$$p(\Sigma \mid y) \propto \left[ |\Sigma_{11}| |\Omega| \right]^{-\left(\frac{1}{2} + m_1\right)} \exp \left( -\frac{1}{2} \text{tr} \Sigma_{11}^{-1} A_{11} - \frac{1}{2} \text{tr} MA \right), \quad \Sigma > 0. \quad (8)$$

For fixed $\Sigma_{11}$, it is readily seen by making use of (A8.1.1) in Appendix A8.1 that the Jacobian of the transformation from $(\Sigma_{12}, \Sigma_{22})$ to $(T, \Omega)$ is

$$J = \left| \frac{\partial(\Sigma_{12}, \Sigma_{22})}{\partial(T, \Omega)} \right| = |\Sigma_{11}|^{m_1}.$$ (8)

Noting that $M$ does not depend on $\Sigma_{11}$, it follows from (8.5.16) and (8.5.17) that $\rho$ is independent of $(T, \Omega)$ so that

$$p(\Sigma_{11}, T, \Omega \mid y) = p(T, \Omega \mid y)p(\Sigma_{11} \mid y) \quad (8)$$

where

$$p(T, \Omega \mid y) \propto |\Omega|^{-\left(\frac{1}{2} + m_1\right)} \exp \left( -\frac{1}{2} \text{tr} MA \right), \quad \Omega > 0, \quad -\infty < T < \infty \quad (8)$$

and

$$p(\Sigma_{11} \mid y) \propto |\Sigma_{11}|^{-\left(\frac{1}{2} + m_1\right)} \exp \left( -\frac{1}{2} \text{tr} \Sigma_{11}^{-1} A_{11} \right), \quad \Sigma_{11} > 0. \quad (8)$$

Thus, $\Sigma_{11}$ is distributed as the inverted Wishart given in (8.5.11).

From (8.5.10) and (8.5.15), we may express the exponent of the distribution (8.5.19) as

$$\text{tr} MA = \text{tr} \left( T \Omega^{-1} T' A_{11} - T \Omega^{-1} A_{21} - \Omega^{-1} T' A_{12} + \Omega^{-1} A_{22} \right)$$

$$= \text{tr} \Omega^{-1} \left[ A_{22} + (T - \bar{T})' A_{11} (T - \bar{T}) \right]$$

where $\bar{T}$ is in exactly the same form as the joint distribution $(\Sigma, \Theta)$ in (8.5.1). The results in (8.5.12) and (8.5.13) follow by application of Theorem 8.5.1 (on p. 461). Further, by using Theorem 8.4.1 (on p. 449), we obtain (8.5.1-
Distribution of $\sigma_{11}$

By setting $m_1 = 1$ in (8.5.11), the distribution of $\sigma_{11}$ is

$$ p(\sigma_{11} \mid y) \propto \sigma_{11}^{-(\nu+1)} \exp\left(-\frac{d_{11}}{2\sigma_{11}}\right), \quad \sigma_{11} > 0, \quad (8.5.22) $$

a $\chi^2$ distribution with $\nu = n - k - (m - 1)$ degrees of freedom. Comparing with the univariate case in (8.5.7), we see that the two distributions are identical except, as expected, that the degrees of freedom differ by $(m - 1)$. The difference is of minor importance when $m$ is small relative to $n - k$ but can be appreciable otherwise—see the discussion about the posterior distribution of $\theta_i^j$ in (8.4.28).

It is clear that the distribution of $\sigma_{ii}$, the $i$th diagonal element of $\Sigma$, is given by simply replacing the subscripts $(1,1)$ in (8.5.22) by $(i,i)$.

The Two Regression Matrices $\Sigma_{11}^{-1} \Sigma_{12}$ and $\Sigma_{21}^{-1} \Sigma_{22}$

The $m_1 \times m_2$ matrix of "regression coefficients" $T = \Sigma_{11}^{-1} \Sigma_{12}$ measures the dependence of the conditional expectation of the errors $(e_{(m_1+1)}, \ldots, e_{m_2})$ on $(e_1, \ldots, e_{m_1})$. From the posterior distribution of $T$ in (8.5.13) the plausibility of different values of the measure $T$ may be compared in terms of e.g. the H.P.D. regions. In deciding whether a parameter point $T = T_0$ lies inside or outside a desired H.P.D. region, from (8.5.14) and (8.4.66) one may then calculate

$$ \Pr \{ U(T) > U(T_0) \mid y \} = \Pr \{ \chi^2_{m_1 m_2} < -\phi' (\nu + m_1) \log U(T_0) \}, \quad (8.5.23) $$

where

$$ \phi' = 1 + \frac{m_1 + m_2 - 3}{2(\nu + m_1)}. $$

In particular, if $T_0 = 0$ which corresponds to $\Sigma_{12} = 0$, that is, $(e_{m_1+1}, \ldots, e_{m_2})$ are independent of $(e_1, \ldots, e_{m_1})$, then

$$ U(T_0 = 0) = \frac{|A|}{|A_{11}| |A_{22}|}. \quad (8.5.24) $$

Consider now the $m_2 \times m_1$ matrix of "regression coefficients"

$$ Z = \Sigma_{22}^{-1} \Sigma_{21} \quad (8.5.25) $$

It is clear from the development leading to (8.5.13) and (8.5.14) that by interchanging the roles of $m_1$ and $m_2$, we have

$$ Z \sim t_{m_2 \nu} (Z, A_{22}^{-1}, A_{11}^{-1}, v + m_2) \quad (8.5.26) $$

where

$$ Z = A_{22}^{-1} A_{21}, \quad A_{11}^{-1} = A_{11} - A_{12} A_{22}^{-1} A_{21}, $$

and

$$ U(Z) = |I_{m_1} + A_{22}^{-1}(Z - Z') A_{22} (Z - Z)|^{-1} $$

$$ \sim U_{(m_1, \nu + m_2)}. \quad (8.5.27) $$
Thus, in deciding whether the parameter point \( Z_0 \) is included in a desired H.P.D. region, we calculate

\[
\Pr \{ U(Z) > U(Z_0) \mid y \} = \Pr \{ \chi^2_{m_1, m_2} < -\phi'/(v + m_2) \log U(Z_0) \}
\]  

(8.5.28)

where

\[
\phi' = 1 + \frac{m_1 + m_2 - 3}{2(v + m_2)}.
\]

In particular, if \( Z_0 = 0 \) which again corresponds to \( \Sigma_{12} = 0 \), then

\[
U(Z_0 = 0) = \frac{|A|}{|A_{11}| |A_{22}|}.
\]

(8.5.29)

Although \( U(Z_0 = 0) = U(T_0 = 0) \) the probabilities on the right-hand sides of (8.5.23) and (8.5.28) will be different whenever \( m_1 \neq m_2 \). This is not a surprising result. For, it can be seen from (8.5.18) that the distribution of \( T \) is in fact proportional to the conditional distribution of \( \Sigma_{12} \), given \( \Sigma_{11} \). Inferences about the parameter point \( \Sigma_{12} = 0 \) in terms of the probability \( \Pr \{ \chi^2_{m_1, m_2} < -\phi'/(v + m_2) \log U(T_0 = 0) \} \) can thus be interpreted as made conditionally for fixed \( \Sigma_{11} \). That is to say that we are comparing the plausibility of \( \Sigma_{12} = 0 \) with other values of \( \Sigma_{12} \) in relation to a fixed \( \Sigma_{11} \). On the other hand, in terms of \( \Pr \{ \chi^2_{m_1, m_2} < -\phi'/(v + m_2) \log U(Z_0 = 0) \} \), inferences about \( \Sigma_{12} = 0 \) can be regarded as conditional on fixed \( \Sigma_{22} \). Thus, one would certainly not expect that, in general, the two types of conditional inferences about \( \Sigma_{12} = 0 \) will be identical.

### 8.5.3 An Example

For illustration, consider again the product and by-product data in Table 8.4.1. The relevant sample quantities are given in (8.4.73).

When interest centers on the variance \( \sigma_{11} \) of the error \( \epsilon_{ut} \) corresponding to the product \( \nu_1 \), we have from (8.5.22)

\[
\sigma_{11} \sim 61.4084 \chi_{a}^{-2}
\]

(8.5.30)

Using Table II (at the end of this book), limits of the 95 per cent H.P.D. interval of \( \log \sigma_{11} \) in terms of \( \sigma_{11} \) are (3.02, 20.79).

Similarly, the posterior distribution of \( \sigma_{22} \) is such that

\[
\sigma_{22} \sim 36.7369 \chi_{a}^{-2}
\]

(8.5.31)

and the limits of the corresponding 95 per cent H.P.D. interval are (1.81, 12.44).

From (8.5.13), and since \( m_1 = m_2 = 1 \), the posterior distribution of \( T = \sigma_{11}^{-1/2} \sigma_{12} \) is the univariate \( t(\hat{T}, \sigma^2, \nu_1) \) distribution where

\[
\hat{T} = a_{11}^{1/2} a_{12} = -0.621, \quad \nu_1 = 10,
\]
and

\[ s_1^2 = v_1^{-1}a_{11}^{-1}a_{22,1} = \frac{a_{22}^2 - a_{12}^2/a_{11}}{v_1 \times a_{11}} = 0.0206. \]

Thus,

\[ \frac{T + 0.627}{0.143} \sim t_{10} \tag{8.5.32} \]

so that from Table IV at the end of this book, limits of the 95 per cent H.P.D. interval are \((-0.95, -0.31)\). In particular, the parameter point \(\sigma_{11}^2, \sigma_{12} = 0\) (corresponds to \(\sigma_{12} = 0\)) is excluded from the 95 per cent interval.

Finally, from (8.5.26) \( Z = \sigma_{12}^2, \sigma_{12} \) is distributed as \(t(Z, s_2^2, v_2)\) where

\[ Z = a_{22}^{-1}a_{12} = -1.05, \quad v_2 = 10 \]

and

\[ s_2^2 = v_2^{-1}a_{11}^{-1}a_{11,2} = \frac{a_{11}^2 - a_{12}^2/a_{22}}{v_2 \times a_{22}} = 0.0574. \]

Thus,

\[ \frac{Z + 1.05}{0.240} \sim t_{10}. \tag{8.5.33} \]

Limits of the 95 per cent H.P.D. interval are \((-1.58, -0.52)\) and the point \(\sigma_{12}^2 = 0\) is again excluded. Further, from (8.5.14), (8.5.27), (8.5.32) and (8.5.33)

\[ \Pr \{ U(T) > U(0) \mid y \} = \Pr \{ U(Z) > U(0) \mid y \} = \Pr \{ |t_{10}| > 4.37 \} \tag{8.5.34} \]

so that inferences about \(\sigma_{12} = 0\) in terms of either \(T\) or \(Z\) are identical. This is of course to be expected since, for this example, \(m_1 = m_2 = 1\).

8.5.4 Distribution of the Correlation Coefficient \(\rho_{12}\)

The two regression matrices \(\Sigma_{12}^{-1} \Sigma_{12}\) and \(\Sigma_{22}^{-1} \Sigma_{22}\) are measures of the dependence (or association) between the two sets of responses \((y_{w1}, \ldots, y_{wm})\) and \((y_{w(m+1)}, \ldots, y_{wm})\). When interest centers at the association between two specific responses \(y_{wi}\) and \(y_{wj}\), the most natural measure of association is the correlation coefficient \(\rho_{ij} = \sigma_{ij}/(\sigma_i \sigma_j)^{1/2}\). Without loss of generality, we now consider how inferences may be made about \(\rho_{12}\).

By setting \(m_1 = 2\) in the distribution of \(\Sigma_{11}\) in (8.5.11), we can follow the development in Jeffreys (1961, p. 174) to obtain the posterior distribution of the correlation coefficient \(\rho_{12}\) as

\[ p(\rho \mid y) \propto (1 - \rho^2)^{k(v-2)} \int_0^\infty \omega^{-1} \left( \omega + \frac{1}{\omega} - 2pr \right)^{-(v+1)} d\omega, \quad -1 < \rho < 1, \tag{8.5.35} \]
where \( \rho = \rho_{12} \),

\[
r = r_{12} = \frac{a_{12}}{(a_{11} a_{22})^{1/2}}
\]
is the sample correlation coefficient, and the normalizing constant is

\[
2(1 - r^2)^{(
u + 1)/2} \Gamma(\nu + 1) / \left[ \pi^{1/2} \Gamma\left(\frac{\nu}{2}\right) \Gamma\left(\frac{\nu + 1}{2}\right) \right].
\]

It is noted that this distribution depends only upon the sample correlation coefficient \( r \).

To see this, for \( m = 2 \), the posterior distribution of the elements \( (\sigma_{11}, \sigma_{22}, \sigma_{12}) \) in \( \Sigma_{11} \) in (8.5.11) is

\[
p(\sigma_{11}, \sigma_{22}, \sigma_{12} \mid y) \propto \left[ \sigma_{11} \sigma_{22} (1 - \rho^2) \right]^{-(\nu + 2)}
\]

\[
\times \exp \left( -\frac{1}{2(1 - \rho^2)} \left[ \frac{a_{11}}{\sigma_{11}} + \frac{a_{22}}{\sigma_{22}} - \frac{2\rho a_{12}}{(\sigma_{11} \sigma_{22})^{1/2}} \right] \right),
\]

\[
\sigma_{11} > 0, \quad \sigma_{22} > 0, \quad \sigma_{11}\sigma_{22} > \sigma_{12}^2.
\] (8.5.5)

where from (8.5.6), the normalizing constant is,

\[
(a_{11}a_{22} - \sigma_{12}^2)^{(
u + 1)/2} / \left[ 2^{(\nu + 1)/2} \prod_{i=1}^{2} \Gamma\left(\frac{\nu + 2 - i}{2}\right) \right].
\]

We now make the transformation, due to Fisher (1915),

\[
x = \left( \frac{\sigma_{11} \sigma_{22}}{a_{11} a_{22}} \right)^{1/2}, \quad \omega = \left( \frac{\sigma_{11} \sigma_{22}}{a_{11} a_{22}} \right)^{1/2}, \quad \rho = \frac{a_{12}}{(\sigma_{11} \sigma_{22})^{1/2}}.
\] (8.5.3)

The Jacobian is

\[
\left| \frac{\partial(\sigma_{11} \sigma_{22} \sigma_{12})}{\partial(x, \omega, \rho)} \right| = 2a_{11} \sigma_{22} (\sigma_{11} \sigma_{22})^{1/2}
\] (8.5.3)

so that the distribution of \((x, \omega, \rho)\) is

\[
p(x, \omega, \rho \mid y) \propto (1 - \rho^2)^{\nu/2} \omega^{-1} x^{-(\nu + 2)} \exp \left[ -\frac{1}{2(1 - \rho^2)x} \left( \omega + \frac{1}{\omega} - 2\rho \right) \right],
\]

\[
\omega > 0, \quad x > 0, \quad -1 < \rho < 1.
\] (8.5.39)

Upon integrating out \( x \),

\[
p(\omega, \rho \mid y) \propto (1 - \rho^2)^{\nu/2} \omega^{-1} \left( \omega + \frac{1}{\omega} - 2\rho \right)^{-(\nu + 1)}
\]

\[
\omega > 0, \quad -1 < \rho < 1.
\] (8.5.40)

from which we obtain the distribution of \( \rho \) given in (8.5.35).
The Special Case When $r = 0$

When $r = 0$, the distribution in (8.5.35) reduces to

$$p(\rho \mid r = 0) \propto (1 - \rho^2)^{\frac{v}{2}(\nu - 2)}, \quad -1 < \rho < 1,$$

(8.5.41)

which is symmetric at $\rho = 0$, and is identical in form to the sampling distribution of $r$ on the null hypothesis that $\rho = 0$. In this case, if we make the transformation

$$\rho = \frac{t}{(v + t^2)^{1/2}},$$

(8.5.42)

then the distribution of $t$ is

$$p(t) \propto (1 + t^2)^{-(\nu + 1)}, \quad -\infty < t < \infty,$$

so that the quantity $t$ is distributed as $t(0, 1, \nu)$.

The General Case When $r \neq 0$

In general, the density function (8.5.35) cannot be expressed in terms of simple functions of $r$. With the availability of a computer, it can always be evaluated by numerical integration, however. To illustrate, consider again the bivariate product and by-product data introduced in Table 8.4.1. Figure 8.5.1 shows the posterior distribution of $\rho$ calculated from (8.5.35). For this example, $v = 9$ and $r = -0.81$. The distribution is skewed to the right and concentrated rather sharply about its mode at $\rho = -0.87$; it practically rules out values of $\rho$ exceeding $-0.3$.

![Figure 8.5.1 Posterior distributions of $\rho$ for the product and by-product data.](image)
Series Expansions of \( p(\rho \mid y) \)
The distribution in (8.5.35) can be put in various different forms. In particular, it can be expressed as

\[
p(\rho \mid y) \propto \frac{(1 - \rho^2)^{(v-2)}}{(1 - \rho r)^{y+\frac{1}{2}}} S_y(\rho, r), \quad -1 < \rho < 1, \tag{8.5.43}
\]

where

\[
S_y(\rho, r) = 1 + \sum_{l=1}^{\infty} \frac{1}{l!} \left( \frac{1 + \rho r}{8} \right)^l \frac{1}{l+1} \frac{(2\lambda - 1)^2}{(v + s + \frac{1}{2})}
\]

is a hypergeometric series, and the normalizing constant is

\[
(1 - r^{2})^{1/2} \left[ \Gamma(v + 1) \right]^2 \left[ 2^{(v-\frac{1}{2})} \Gamma\left(\frac{v}{2}\right) \Gamma\left(\frac{v+1}{2}\right) \Gamma(v + \frac{1}{2}) \right].
\]

To see this, the integral in (8.5.35) can be written

\[
\int_{0}^{\infty} \omega^{-1} \left[ \omega + \frac{1}{\omega} - 2 \rho r \right]^{-(\nu+1)} d\omega = 2 \int_{1}^{\infty} \omega^{-1} \left[ \omega + \frac{1}{\omega} - 2 \rho r \right]^{-(\nu+1)} d\omega. \tag{8.5.44}
\]

On the right-hand side of (8.5.44), we may make the substitution, again due to Fisher,

\[
u = 1 - \frac{1 - \rho r}{\frac{1}{2}[\omega + (1/\omega)] - \rho r} = \frac{1}{2}[\omega + (1/\omega)] - 1, \tag{8.5.45}
\]

Noting that

\[
\frac{1}{4} \left( \omega + \frac{1}{\omega} \right)^2 - \frac{1}{4} \left( \omega - \frac{1}{\omega} \right)^2 = 1,
\]

we have

\[
\frac{\partial \nu}{\partial \omega} = \omega^{-1} (1 - \omega) (2\omega)^{1/2} \left[ 1 - \frac{1}{2}(1 + \rho r)\omega \right]^{1/2} (1 - \rho r)^{-1/2} \tag{8.5.46}
\]

so that

\[
p(\rho \mid y) \propto \frac{(1 - \rho^2)^{(v-2)}}{(1 - \rho r)^{y+\frac{1}{2}}} \int_{0}^{1} (1 - \omega)^{y} \left[ 1 - \frac{1}{2}(1 + \rho r)\omega \right]^{-1/2} d\omega, \quad -1 < \rho < 1. \tag{8.5.47}
\]

Expanding the last term in the integrand in powers of \( u \) and integrating term by term, each term being a complete beta function, we obtain the result in (8.5.42).
We remark here that an alternative series representation of the distribution of $\rho$ can be obtained as follows. In the integral of (8.5.35), since

$$
\omega^{-1} \left( \omega + \frac{1}{\omega} - 2\rho r \right)^{-(\nu+1)} = \omega^\nu (\omega^2 - 2\rho \omega + 1)^{-(\nu+1)}, \quad (8.5.48)
$$

by completing the square in the second factor on the right-hand side of (8.5.48), we can write the distribution in (8.5.35) as

$$
p(\rho | \gamma) \propto \frac{(1 - \rho^2)^{\frac{\nu}{2}(\nu-2)}}{(1 - \rho^2)^{\frac{\nu+1}{2}}} \omega^\nu \left[ 1 + \frac{(\omega - \rho r)^2}{1 - \rho^2} \right]^{-(\nu+1)} d\omega \quad (8.5.49)
$$

Upon repeated integration by parts, the above integral can be expressed as a finite series involving powers of $[(1 - \rho r)/(1 + \rho r)]^{\nu/2}$ and Student's $t$ integrals. The density function of $\rho$ can thus be calculated from a table of $t$ distribution. This process becomes very tedious when $\nu$ is moderately large so its practical usefulness is limited.

The series $S_r(\rho, r)$ in (8.5.43) has its leading term equal to one, followed by terms of order $\nu^{-l}, l = 1, 2, \ldots$. When $\nu$ is moderately large, we may simply take the first term so that approximately

$$
p(\rho | \gamma) \approx c \frac{(1 - \rho^2)^{\frac{\nu}{2}(\nu-2)}}{(1 - \rho r)^{\frac{\nu+1}{2}}}, \quad -1 < \rho < 1, \quad (8.5.50)
$$

where $c$ is the normalizing constant:

$$
c^{-1} = \int_{-1}^{1} \frac{(1 - \rho^2)^{\frac{\nu}{2}(\nu-2)}}{(1 - \rho r)^{\frac{\nu+1}{2}}} d\rho.
$$

Although evaluation of $c$ would still require the use of numerical methods, it is much simpler to calculate the distribution of $\rho$ using (8.5.50) than to evaluate the integral in (8.5.35) for every value of $\rho$. Table 8.5.1 compares the exact distribution with the approximation using the data in Table 8.4.1. In spite of the fact that $\nu$ is only 9, the agreement is very close.

It is easily seen that the density function (8.5.50) is greatest when $\rho$ is near $r$.

However, except when $r = 0$, the distribution is asymmetrical.

The asymmetry can be reduced by making the transformation

$$
\zeta = \tanh^{-1} \rho = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}, \quad (8.5.51)
$$

due to Fisher (1921). Following his argument, it is found that $\zeta$ is approximately Normal

$$
N \left[ \tanh^{-1} \rho - \frac{5\gamma}{2(\nu + 1)^{-1}}, (\nu + 1)^{-1} \right].
$$
Setting \( m = 2, \ k = 1 \) so that \( v = n - 2 \), the distribution in (8.5.43) is identical to that given by Jeffreys for the case of sampling from a bivariate Normal population. Finally, we note that while we have obtained above the distribution of the specific correlation coefficient \( \rho = \rho_{12} \), it is clear that the distribution of any correlation \( \rho_{ij}, i \neq j \), is given simply by setting \( r = r_{ij} = a_{ij} / (a_{ii} a_{jj})^{1/2} \) in (8.5.35) and its associated expressions.

### Table 8.5.1
Comparison of the exact and the approximate distributions of \( \rho \) for \( v = 5 \) and \( r = -0.81 \)

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Exact</th>
<th>Approximate</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.98</td>
<td>0.2286</td>
<td>0.2283</td>
</tr>
<tr>
<td>-0.96</td>
<td>1.2164</td>
<td>1.2150</td>
</tr>
<tr>
<td>-0.94</td>
<td>2.4867</td>
<td>2.4844</td>
</tr>
<tr>
<td>-0.92</td>
<td>3.5159</td>
<td>3.5134</td>
</tr>
<tr>
<td>-0.90</td>
<td>4.1200</td>
<td>4.1180</td>
</tr>
<tr>
<td>-0.88</td>
<td>4.3281</td>
<td>4.3271</td>
</tr>
<tr>
<td>-0.86</td>
<td>4.2448</td>
<td>4.2448</td>
</tr>
<tr>
<td>-0.84</td>
<td>3.9782</td>
<td>3.9750</td>
</tr>
<tr>
<td>-0.82</td>
<td>3.6141</td>
<td>3.6157</td>
</tr>
<tr>
<td>-0.80</td>
<td>3.2127</td>
<td>3.2149</td>
</tr>
<tr>
<td>-0.70</td>
<td>1.5182</td>
<td>1.5210</td>
</tr>
<tr>
<td>-0.60</td>
<td>0.6584</td>
<td>0.6603</td>
</tr>
<tr>
<td>-0.50</td>
<td>0.2853</td>
<td>0.2865</td>
</tr>
<tr>
<td>-0.40</td>
<td>0.1260</td>
<td>0.1267</td>
</tr>
<tr>
<td>-0.30</td>
<td>0.0569</td>
<td>0.0572</td>
</tr>
<tr>
<td>-0.20</td>
<td>0.0261</td>
<td>0.0263</td>
</tr>
</tbody>
</table>

8.6 **A SUMMARY OF FORMULAE AND CALCULATIONS FOR MAKING INFERENCES ABOUT \( (\theta, \Sigma) \)**

Using the product, by-product data in Table 8.4.1 for illustration, Table 8.6.1 below provides a short summary of the formulae and calculations required for making inferences about the elements of \( (\theta, \Sigma) \) for the linear model with common derivative matrix defined in (8.4.1). Specifically, the model is

\[
y = X\theta + \varepsilon
\]  

(8.6.1)

where \( y = [y_1, \ldots, y_m] \) is a \( n \times m \) matrix of observations, \( X \) is a \( n \times k \) matrix of fixed elements with rank \( k \), \( \theta = [\theta_1, \ldots, \theta_m] \) is a \( k \times m \) matrix of parameters and \( \varepsilon = [e_{i1}, \ldots, e_{im}]' \) is a \( n \times m \) matrix of errors. It is assumed that \( e_{ui}, u = 1, \ldots, n \) are independently distributed as \( N_m(0, \Sigma) \).
Table 8.6.1
Summarized calculations for the linear model $y = X\theta + \epsilon$

1. From (8.4.1), (8.4.4), (8.4.7) and (8.4.13), obtain

$m = 2, \quad k = 2, \quad n = 12, \quad v = n - (m + k) + 1 = 9$

$X'X = \begin{bmatrix} 12 & 0 \\ 0 & 0.15 \end{bmatrix}, \quad C = (X'X)^{-1} = \begin{bmatrix} 0.08 & 0 \\ 0 & 6.88 \end{bmatrix}$

$\theta = (X'X)^{-1}X'y = \begin{bmatrix} 71.14 & 18.07 \\ 54.44 & -20.09 \end{bmatrix}, \quad A = \{ a_{ij} \} \quad i, j = 1, \ldots, m$

$a_{ij} = \langle y_i - X\bar{\theta} \rangle \langle y_j - X\bar{\theta} \rangle, \quad \text{and} \quad A = \begin{bmatrix} 61.41 & -38.48 \\ -38.48 & 36.74 \end{bmatrix}$

2. Inferences about a specific column or row of $\theta$:

Writing

$\theta = [\theta_1, \ldots, \theta_m] = \begin{bmatrix} \theta_{1(1)} \\ \vdots \\ \theta_{m(1)} \end{bmatrix}, \quad \bar{\theta} = [\bar{\theta}_1, \ldots, \bar{\theta}_n] = \begin{bmatrix} \bar{\theta}_{1(1)} \\ \vdots \\ \bar{\theta}_{n(1)} \end{bmatrix}$

then from (8.4.28) and (8.4.34),

$\theta_i \sim t_m \left( \bar{\theta}_i, \frac{a_{ii}}{v} C, v \right), \quad i = 1, \ldots, m,$

and

$\theta_{(g)} \sim t_m \left( \bar{\theta}_{(g)}, \frac{c_{gg}}{v} A, v \right), \quad g = 1, \ldots, k.$

Thus, for $i = 1$ and $g = 2$

$\theta_1 \sim t_2 \left( \begin{bmatrix} 71.14 \\ 54.44 \end{bmatrix}, 6.83 \times \begin{bmatrix} 0.08 & 0 \\ 0 & 6.88 \end{bmatrix}, 9 \right)$

$\theta_{(2)} \sim t_2 \left( \begin{bmatrix} 54.44 \\ 20.09 \end{bmatrix}, 0.76 \times \begin{bmatrix} 61.41 & -38.48 \\ -38.48 & 36.74 \end{bmatrix}, 9 \right)$

3. H.P.D. regions of $\theta$: To decide if a general parameter point $\theta_0$ lies inside or outside the $(1 - \alpha)$ H.P.D. region, from (8.4.63) and (8.4.66), use the approximation

$-\phi v \log U \sim \chi^2_{obs}$

where

$U = U(\theta) = |A| |A + (\theta - \bar{\theta})'X'X(\theta - \bar{\theta})|^{-1} \quad \text{and} \quad \phi = 1 + \frac{m + k - 3}{2v}$

so that $\theta_0$ lies inside the $1 - \alpha$ region if

$-\phi v \log U(\theta_0) < \chi^2(mk, \alpha).$
For the example, \( \phi = 19/18 \). Thus, if
\[
\theta_0 = \begin{bmatrix} 70 & 17 \\ 65 & -30 \end{bmatrix}, \quad \text{then} \quad U(\theta_0) = 0.17
\]
and the point lies inside the \( 1 - \alpha \) region if
\[
-9.5 \log 0.17 = 16.7 < \chi^2(4, \alpha).
\]

4. H.P.D. regions for a block submatrix of \( \theta \): Let
\[
\begin{align*}
\theta &= \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}, \\
&= \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21_0} & \theta_{22_0} \end{bmatrix}, \\
\theta_0 &= \begin{bmatrix} \theta_{11_0} & \theta_{12_0} \\ \theta_{21_0} & \theta_{22_0} \end{bmatrix},
\end{align*}
\]
\[
A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \quad k_1 = k_2.
\]

To decide, for example, if the parameter point \( \theta_{11,0} \) lies inside or outside the \( (1 - \alpha) \) H.P.D. region for \( \theta_{11} \), use the approximation in (8.4.68),
\[
\phi_1 \nu \log U(\theta_{11}) \sim \chi^2_{m_1 k_1},
\]
where
\[
U(\theta_{11}) = |A_{11}| |A_{11} + (\theta_{11} - \theta_{11_0}) C_{11}^{-1} (\theta_{11} - \hat{\theta}_{11})|^{-1}
\]
and \( \phi_1 = 1 + (1/2v)(m_1 + k_1 - 3) \), so that \( \theta_{11,0} \) lies inside the \( (1 - \alpha) \) region if
\[
-\phi_1 \nu \log U(\theta_{11,0}) < \chi^2(m_1 k_1, \alpha),
\]
Thus if \( m_1 = k_1 = 1 \), \( \theta_{11,0} = 70 \), then \( \phi_1 = 17/18 \) and
\[
U(\theta_{11,0} = 70) = 61.41 \left[ \frac{(70 - 71.14)^2}{0.08} \right] = 0.79,
\]
so that \( \theta_{11,0} \) lies inside the \( (1 - \alpha) \) region if
\[-8.5 \log 0.79 = 2.0 < \chi^2(1, \alpha).\]

Note that since \( m_1 = k_1 = 1 \), exact results could be obtained and the above is for illustration only.

5. Inferences about the diagonal elements of \( \Sigma \):

From (8.5.22)
\[
\sigma_{ii} \sim a_{ii} \chi^2_{v_i}, \quad i = 1, \ldots, m.
\]

Thus, for \( i = 1 \), \( \sigma_{ii} \sim 61.41 \chi^2 v_i \).

6. H.P.D. regions for the "regression matrix" \( T \):

Let
\[
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad T = \Sigma_{11}^{-1} \Sigma_{12}, \quad T = A_{11}^{-1} A_{12},
\]
\[
A_{22,1} = A_{22} - A_{21} A_{11}^{-1} A_{12}
\]
Table 8.6.1 Continued

To decide if a parameter point $T_0$ lies inside or outside the $(1-\alpha)$ H.P.D. region, from (8.5.23), use the approximation

$$-\phi'(\nu + m_1) \log U(T) \sim \chi^2_{m,m_2},$$

where

$$\phi' = 1 + \frac{m_1 + m_2 - 3}{2(\nu + m_1)}$$

and

$$U(T) = |A_{22^{-1}}| A_{22^{-1}} + (T - \bar{T})' A_{11}(T - \bar{T})^{-1}$$

so that $T_0$ lies inside the region if

$$-\phi'(\nu + m_1) \log U(T_0) < \chi^2(m_1, \alpha).$$

Thus, for $m_1 = m_2 = 1, \; \phi'(\nu + m_1) = 9.5, \; A_{22^{-1}} = a_{22^{-1}} = 12.63$ and $\bar{T} = 0.63$.

If $T_0 = 0$, then

$$U(T_0 = 0) = 12.63/[12.63 + (0.63)^2 \times 61.41] = 0.34$$

and the point $T_0 = 0$ will lie inside the $(1-\alpha)$ region if

$$-9.5 \log 0.34 = 10.3 < \chi^2(1, \alpha).$$

Note that (i) $U(0) = |A||A_{11}^{-1}A_{22^{-1}}|^{-1}$, and (ii) since $m_1 = m_2 = 1$, exact results are, of course, available.

7. Inferences about the correlation coefficient $\rho_{ij}$: Use the approximating distribution in (8.5.50),

$$p(\rho | y) \propto \frac{(1-\rho^2)^{(n-2)}}{(1-\rho^2)^{n/2}}, \quad -1 < \rho < 1,$$

where

$$\rho = \rho_{ij} \quad \text{and} \quad r = r_{ij} = a_{ij}^2/(a_{ii}a_{jj})^{1/2}, \quad i, j = 1, \ldots, m.$$

Thus,

$$p(\rho | y) \propto \frac{(1-\rho^2)^{n/2}}{(1+0.81\rho)^{n/2}}, \quad -1 < \rho < 1.$$

The normalizing constant of the distribution can be obtained by numerical integration when desired.

---

APPENDIX A8.1

The Jacobians of Some Matrix Transformations

We here give the Jacobians of some matrix transformations useful in multivariate problems. In what follows the signs of the Jacobians are ignored.

a) Let $X$ be a $k \times m$ matrix of $km$ distinct random variables. Let $A$ and $B$ be, respectively, a $k \times k$ and $m \times m$ matrices of fixed elements. If

$$Z_1 = AX, \quad Z_2 = XB, \quad Z_3 = AXB.$$
then the Jacobians are, respectively,

\[
\left| \frac{\partial Z_1}{\partial X} \right| = |A|^m, \quad \left| \frac{\partial Z_2}{\partial X} \right| = |B|^k, \quad \left| \frac{\partial Z_3}{\partial X} \right| = |A|^m |B|^k. \quad (A8.1.1)
\]

b) Let \( X \) be a \( m \times m \) symmetric matrix consisting of \( \frac{1}{2}m(m+1) \) distinct elements and let \( C \) be a \( m \times m \) matrix of fixed elements. If

\[ Y = C X C', \]

then

\[
\left| \frac{\partial Y}{\partial X} \right| = |C|^{m+1}. \quad (A8.1.2)
\]

For proofs, see Deemer and Olkin (1951), based on results given by P. L. Hsu, also Anderson (1958, p. 162).

c) Let \( X \) be a \( m \times m \) nonsingular matrix of random variables and \( Y = X^{-1} \). Then,

\[
\frac{\partial Y}{\partial z} = - Y \frac{\partial X}{\partial z} Y, \quad (A8.1.3)
\]

where \( z \) is any function of the elements of \( X \).

**Proof:** Since \( XY = I \), it follows that

\[
\frac{\partial}{\partial z} (XY) = \left( \frac{\partial X}{\partial z} \right) Y + X \frac{\partial Y}{\partial z} = 0.
\]

Hence

\[
\frac{\partial Y}{\partial z} = - Y \left( \frac{\partial X}{\partial z} \right) Y.
\]

The Jacobians of two special cases of \( X \) are of particular interest.

a) If \( X \) has \( m^2 \) distinct random variables, then from (A8.1.1)

\[
\left| \frac{\partial Y}{\partial X} \right| = |Y|^{2m}. \quad (A8.1.4)
\]

b) If \( X \) is symmetric, and consists of \( \frac{1}{2}m(m+1) \) distinct random variables, then from (A8.1.2)

\[
\left| \frac{\partial Y}{\partial X} \right| = |Y|^{m+1}. \quad (A8.1.5)
\]
APPENDIX A8.2

The Determinant of the Information Matrix of $\Sigma^{-1}$

We now obtain the determinant of the information matrix of $\Sigma^{-1}$ for the $m$-dimensional Normal distribution $N_m(\mu, \Sigma)$. The density is

$$p(y | \mu, \Sigma) = (2\pi)^{-m/2} |\Sigma^{-1}|^{1/2} \exp \left[ -\frac{1}{2} \text{tr} \Sigma^{-1} (y - \mu) (y - \mu)' \right],$$

$$-\infty < y < \infty, \quad (A8.2.1)$$

where $y = (y_1, \ldots, y_m)'$, $\mu = (\mu_1, \ldots, \mu_m)'$, $\Sigma = \{\sigma_{ij}\}$ and $\Sigma^{-1} = \{\sigma^{ij}\}, i, j = 1, \ldots, m$. We assume that $\Sigma$ (and $\Sigma^{-1}$) consists of $\frac{1}{2}m(m+1)$ distinct elements. Taking logarithms of the density function, we have

$$\log p = -\frac{m}{2} \log (2\pi) + \frac{1}{2} \log |\Sigma^{-1}| - \frac{1}{2} \text{tr} \Sigma^{-1} (y - \mu) (y - \mu)'.$$  \quad (A8.2.2)

Differentiating $\log p$ with respect to $\sigma^{ij}$, $i, j = 1, \ldots, m$, $i \geq j$,

$$\frac{\partial \log p}{\partial \sigma^{ij}} = \frac{1}{2} \left[ \frac{1}{|\Sigma^{-1}|} \frac{\partial |\Sigma^{-1}|}{\partial \sigma^{ij}} - (y_i - \mu_i)(y_j - \mu_j) \right]. \quad (A8.2.3)$$

Since $\partial |\Sigma^{-1}|/\partial \sigma^{ij} = \sigma^{ij}$ where $\sigma^{ij}$ is the cofactor of $\sigma^{ij}$, it follows that the first term on the right-hand side of (A8.2.3) is simply $\frac{1}{2} \sigma_{ij}$. Thus, the second derivatives are

$$\frac{\partial^2 \log p}{\partial \sigma^{ij} \partial \sigma^{kl}} = \frac{1}{2} \delta^{ij} \delta^{kl}, \quad \left( \begin{array}{c} i,j = 1, \ldots, m; \quad i \geq j \\ k,l = 1, \ldots, m; \quad k \geq l \end{array} \right). \quad (A8.2.4)$$

Consequently, the determinant of the information matrix is proportional to

$$|\mathcal{I}(\Sigma^{-1})| = \left| -E \left[ \frac{\delta^2 \log P}{\partial \sigma^{ij} \partial \sigma^{kl}} \right] \right| \propto \left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right|. \quad (A8.2.5)$$

From (A8.1.5), we have that

$$\left| \frac{\partial \Sigma}{\partial \Sigma^{-1}} \right| = |\Sigma|^{m+1}. \quad (A8.2.6)$$

APPENDIX A8.3

The Normalizing Constant of the $t_m[\theta, (XX')^{-1}, A, v]$ Distribution

Let $\theta$ be a $k \times m$ matrix of variables. We now show that

$$\int_{-\infty < \theta < \infty} |M + A^{-1} (\theta - \theta') XX (\theta - \theta)'|^{-\frac{1}{2} + (k + m - 1)^2} d\theta \quad (A8.3.1)$$

$$= c(m, k, v) X'X|^{-m/2} |A|^{-k/2},$$
where \( v > 0 \), \( \theta \) is a \( k \times m \) matrix, \( X'X \) and \( A^{-1} \) are, respectively, a \( k \times k \) and a \( m \times m \) positive definite symmetric matrix,

\[
c(m, k, v) = \left[ \Gamma(\frac{1}{2}) \right]^{mk} \frac{\Gamma_m[\frac{1}{2}(v + m - 1)]}{\Gamma_m[\frac{1}{2}(v + k + m - 1)]}
\]

and \( \Gamma_p(b) \) is the generalized Gamma function defined in (8.2.22).

Since \( X'X \) and \( A \) are assumed definite, there exist a \( k \times k \) and a \( m \times m \) nonsingular matrices \( G \) and \( H \) such that

\[
X'X = G'G \quad \text{and} \quad A^{-1} = H'H.
\]  

(A8.3.2)

Let \( T \) be a \( k \times m \) matrix such that

\[
T = G(\theta - \hat{\theta})H.
\]  

(A8.3.3)

Using the identity (8.4.11) we can write

\[
|I_m + A^{-1}(\theta - \hat{\theta})'X'X(\theta - \hat{\theta})| = |I_m + H'(\theta - \hat{\theta})'G'G(\theta - \hat{\theta})H| = |I_m + TT'|
\]

\[
= |I_k + TT'|.  
\]  

(A8.3.4)

From (A8.1.1) the Jacobian of the transformation (A8.3.3) is

\[
\left| \frac{\partial T}{\partial \theta} \right| = |G|^k |H|^m = |X'X|^{k/2}|A^{-1}|^{m/2}.
\]  

(A8.3.5)

Thus, the integral on the left-hand side of (A8.3.1) is

\[
|X'X|^{-m/2}|A^{-1}|^{-k/2} Q_m,
\]  

(A8.3.6)

where

\[
Q_m = \int_{-\infty < T < \infty} |I_k + TT'|^{-k(v + k + m - 1)} dT.
\]  

(A8.3.7)

Let \( T = [t_1, \ldots, t_m] = [T_i, t_m] \) where \( t_i \) is a \( k \times 1 \) vector, \( i = 1, \ldots, m \). Then,

\[
|I_k + TT'| = |I_k + T_i T_i' + t_m t_m'|
\]

\[
= |I_k + T_i T_i'[1 + t_m'(I_k + T_i T_i')^{-1} t_m]|.
\]  

(A8.3.8)

It follows that

\[
Q_m = \int_{-\infty < T_i < \infty} |I_k + T_i T_i'|^{-2(v + k + m - 1)} q_m dT_i,
\]  

(A8.3.9)

where

\[
q_m = \int_{-\infty < t_m < \infty} \left[ 1 + t_m'(I_k + T_i T_i')^{-1} t_m \right]^{-2(v + k + m - 1)} dT_m.
\]

From (A2.1.12) in Appendix A2.1,

\[
q_m = \left[ \Gamma(\frac{1}{2}) \right]^k \frac{\Gamma\left[\frac{1}{2}(v + m - 1)\right]}{\Gamma\left[\frac{1}{2}(v + k + m - 1)\right]} |I_k + T_i T_i'|^{1/2}.
\]
Thus,
\[ Q_m = \left[ \Gamma\left(\frac{k}{2}\right) \right]^2 \frac{\Gamma\left(\frac{1}{2}(\nu + m - 1)\right)}{\Gamma\left(\frac{1}{2}(\nu + k + m - 1)\right)} Q_{m-1}, \]

where
\[ Q_{m-1} = \int_{-\infty < T_1 = 0} \left| I_k + T_1 T_1^* \right|^{-\frac{1}{2}(\nu+n+2+i)} dT_1. \]

The result in (A8.3.1) follows by repeating the process \( m - 1 \) times.

**APPENDIX A8.4**

The Kronecker Product of Two Matrices

We summarize in this appendix some properties of the Kronecker product of two matrices.

**Definition:** If \( A \) is a \( m \times m \) matrix and \( B \) is a \( n \times n \) matrix, then the Kronecker product of \( A \) and \( B \) in that order is the \((mn) \times (mn)\) matrix

\[
A \otimes B = \begin{bmatrix}
    a_{11}B & \cdots & a_{1n}B \\
    \vdots & \ddots & \vdots \\
    a_{m1}B & \cdots & a_{mn}B
\end{bmatrix}.
\]

**Properties:**

i) \((A \otimes B)^\prime = A^\prime \otimes B^\prime\)

ii) If \( A \) and \( B \) are symmetric, then \( A \otimes B \) is symmetric.

iii) When \( A \) and \( B \) are non-singular, then

\((A \otimes B)^{-1} = A^{-1} \otimes B^{-1}\)

iv) \( \text{tr} A \otimes B = \text{tr} A \text{tr} B \)

v) \( |A \otimes B| = |A|^n |B|^m \).

vi) If \( C \) is a \( m \times m \) matrix and \( D \) is a \( n \times n \) matrix, then

\[(A + C) \otimes B = A \otimes B + C \otimes B\]

and

\[(A \otimes B) (C \otimes D) = (AC) \otimes (BD).\]