DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
Madison, Wisconsin

TECHNICAL REPORT NO. 31
May, 1964

THE BAYESIAN ESTIMATION OF COMMON PARAMETERS FROM SEVERAL RESPONSES

by
George E. P. Box and Norman R. Draper

This research was supported by the United States Navy through the Office of Naval Research, under Contract Nonr-1202, Project NR 04Z-222. Reproduction in whole or in part is permitted for any purpose of the United States Government.
The Bayesian Estimation of Common Parameters from Several Responses

George E.P. Box and Norman R. Draper
University of Wisconsin

1. INTRODUCTION

Many, perhaps the majority, of problems arising in the physical and biological sciences yield multivariate data, that is to say, observations are made on several characteristics simultaneously, rather than on only one. For example, in an agricultural experiment on wheat, we might observe the straw yield and the stalk height in addition to the wheat yield.

When tackling univariate problems the basic linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

(1.1)

(where \(\mathbf{y}\) and \(\epsilon\) are \(n\times 1\) vectors, \(\mathbf{X}\) is an \(n\times p\) matrix and \(\beta\) is a \(p\times 1\) vector) is of very general utility. When \(\mathbf{X}\) is suitably chosen it can represent all types of regression and analysis of variance models.

One multivariate generalisation of this linear model is obtained as follows. Suppose each observation has \(k\) elements, i.e.

$$\mathbf{y}^u = (y_1^u, y_2^u, \ldots, y_{ku})$$

for \(u = 1, 2, \ldots, n\); suppose \(\mathbf{X}\) is as it was in the univariate case, an \(n\times p\) matrix; and suppose there are now \(k\) vectors of parameters \(\beta_1, \beta_2, \ldots, \beta_k\), each \(p\times 1\). Write \(\mathbf{y}' = (y_1, y_2, \ldots, y_n)\),

$$\mathbf{y} = \mathbf{X}\mathbf{B} + \mathbf{e}$$

(1.2)

where \(\mathbf{e}\) has the same dimensions as \(\mathbf{y}\), provides a multivariate generalisation of all the univariate regression and analysis of variance models. Note that in this generalisation \(\mathbf{X}, \mathbf{B}\) are now \(n\times p\) and \(k\times p\) matrices but that \(\mathbf{X}\)
remains the same. This particular model has been extensively explored by many authors, including M.S. Bartlett, S.N. Roy, C.R. Rao and S.S. Wilks.

In the univariate case, the Neyman-Pearson likelihood ratio approach to hypothesis testing leads to the use of a criterion \( S_R / (S_R + S_E) \) where \( S_R \) and \( S_E \) are the "regression" and "error" sums of squares respectively. A similar approach for the multivariate model \((1, 2)\) leads to a criterion \( |S_R| / |S_R + S_E| \) where the numerator and denominator are now the determinants of the appropriate \( k \times k \) matrices formed from the "regression" and the "regression + error" sums of squares and cross-products. Further analysis of the criterion can be made using "canonical analysis" in which the linear combinations of the observations \( y_1, y_2, \ldots, y_k \) which give the largest, second largest, \ldots, etc., multiple correlation coefficient with the rows of \( X \) are determined. The first of these linear combinations, called the discriminant function, is useful in problems where discrimination or classification of future data is required.

Many multivariate problems are not appropriately handled by these methods. For example, in what may be called the "specification problems" one is trying to find a combination of levels of \( X \) which will result in a combination of levels for \( y \) which fall within satisfactory sets of limits. The present paper discusses a third multivariate problem, which may be introduced by an example. In studying the mechanism of a chemical reaction the investigator often desires to estimate the basic constants of the system (e.g., rate constants or activation energies). In each experimental run he will usually measure not one but several responses. For example he may measure the yields \( y_1, y_2, \ldots, y_k \) of \( k \) different products.
of the reaction. Now, normally, the physical mechanism will imply $k$
distinct functional relationships corresponding to $y_1$, $y_2$, $...$, $y_k$, each
of which will involve some or all of the parameters $\beta_1$, $\beta_2$, $...$, $\beta_p$. If
(as will not, of course, usually be the case in our example) the func-
tional relationships were linear in the parameters and $\beta = (\beta_1, \beta_2, ..., \beta_p)$
the model would be of form

$$y_i = X_i \beta + \epsilon_i, \quad i = 1, 2, ..., k$$  \(1.3\)

We shall adopt the convention that if any of the $\beta_i$ are not associated
with a particular $y_i$, then a column of zeros will appear in the appropriate
location in the $X$ matrix. In this generalisation of the univariate model
the $n$ observations are, as before, those of a $k$ variate response, but now
there are $k$ different $n \times p$ $X$-matrices, $X_i$, and only a single $p \times 1$ vector
of parameters. Thus the parameters are common to all responses. This
linear model (1.3) is given for purposes of comparison only and will sub-
sequently be regarded as a special case of the more usual non-linear sit-
uation.

An experimenter would certainly expect that the parameters of a system
could be better estimated by combining information from $y_1$, $y_2$, ..., $y_k$.
The object of this paper is to show how it is possible to meet this expec-
tation using Bayesian methods.
2. FORMULATION

Let \( y_{iu} \) \((i=1, 2, \ldots, k; \ u = 1, 2, \ldots, n)\) represent \( n \) sets of observations on each of \( k \) responses. Let \( x_{iu}^q \) \((i=1, 2, \ldots, k; \ u = 1, 2, \ldots, n; \ q = 1, 2, \ldots, r)\) represent \( n \) sets of observations on each of \( k \) sets each of \( r \) independent variables. Consider the \( k \) models

\[
y_{iu}^q = f_i(x_{iu}^q, \theta_h^u) + \varepsilon_{iu}^q
\]

(2.1)

for each \( i = 1, 2, \ldots, k \), where the \( f_i \) are \( k \) response functions of known form, where \( \theta_h^u, \ h = 1, 2, \ldots, m \) are \( m \) unknown parameters and where the \( \varepsilon_{iu} \) denote random errors such that

\[
E(\varepsilon_{iu}) = 0, \ \text{all} \ i, u
\]

\[
E(\varepsilon_{iu} \varepsilon_{jv}) = 0, \ \text{all} \ i, j, u \neq v
\]

\[
E(\varepsilon_{iu}^2) = \sigma_{ii}
\]

\[
E(\varepsilon_{iu} \varepsilon_{ju}) = \sigma_{ij}, \ i \neq j, \ \text{all} \ u.
\]

This implies that if we write \( y_u' = (y_{1u}, y_{2u}, \ldots, y_{ku}) \) to represent the vector of \( u \)th observations \((u = 1, 2, \ldots, n)\) on each of the \( k \) variables then the variance - covariance matrix of \( y_u \) is \( \Sigma \), where

\[
\Sigma = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\
\cdots & \cdots & \cdots & \cdots \\
\sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk}
\end{bmatrix} = \{\sigma_{ij}\}
\]

and where \( \sigma_{ij} = \sigma_{ji} \). Let \( A = \Sigma^{-1} = \{\sigma_{ij}\} = \{\sigma_{ij}\}^{-1} \)
The elements of \( y_u \) and \( y_v \), \( u \neq v \) are uncorrelated. Let \( \theta^* = (\theta_1, \theta_2, \ldots, \theta_m) \) represent the vector of the \( m \) parameters in equations (2.1). Suppose \( y_1, y_2, \ldots, y_R \) follows a multivariate normal distribution. What can be deduced about the possible values of the elements of \( \theta^* \)?

We define the quantities
\[
v_{ij} = \sum_{u=1}^{n} \{y_{1u} - f_i(x_{1u}^q, \theta)\} \{y_{ju} - f_j(x_{ju}^q, \theta)\}
\]
which are the sums of squares and sums of products of the deviations of the observed \( y_{1u} \)'s from their respective models.

Now if the variances and covariances \( \sigma_{ij} \) were known, the likelihood would be a monotonic function of the quadratic form
\[
z = \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma_{ij} v_{ij}^*
\]
Minimisation of this quadratic form provides the generalisation of the method of least squares due to Aitken (1935). A further simplification would result if it were known that the elements of the vector of observations were uncorrelated (\( \sigma_{ij} = 0, \ i \neq j \)). In this case, minimisation of \( z \) would correspond to standard weighted least squares, using the reciprocals of the variances as the weights. In most practical problems the \( \sigma_{ij} \) are unknown and it is this case we shall discuss.
3. **TACKLING THE PROBLEM**

Since the $n$ sets of observations $y'_u = (y'_{1u}, y'_{2u}, \ldots, y'_{ku})$ are independent, the likelihood is

$$p(y \mid \theta, \sigma_{ij}) = (2\pi)^{-n/2} |A|^{-k/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma_{ij} y'_{ij} \right\}$$

(3.1)

where we denote the data by $y' = (y'_1, y'_2, \ldots, y'_n)$.

We shall apply Bayes theorem using "non-informative" prior distributions for $\theta$ and the $\sigma_{ij}$. We suppose that little is known a priori about the values of the constants $\theta$. Specifically we suppose that, when suitably parametrized, the prior distribution of the $\theta$'s will not change very much over a region in which the likelihood is appreciable, i.e., we take a locally uniform prior

$$p(\theta) \propto d\theta.$$  

The invariance theory of Jeffreys (1961) leads to

$$p(\sigma_{ij}) = |A|^{-(k+1)/2} = |A|^{-(k+1)/2}$$

as the appropriate non-informative prior distribution for the variances and covariances $\sigma_{ij}$. Other arguments leading to this same prior distribution have been advanced by Savage (1961). (See also Geisser and Cornfield, 1963 and Tiao and Zellner 1964). Note that

$$p(\sigma_{ij}) = |A|^{-1}^{-(k+1)/2} = |A|^{-(k+1)/2}$$

since the Jacobian $\frac{\partial (\sigma_{ij})}{\partial (\sigma_{ij})} = |A|^{-(k+1)}$ can be used to convert
from \( \sigma_{ij} \) to \( \sigma^{ij} \) or vice versa. (The Jacobian is discussed in the Appendix)

Combining the prior distribution for \( \theta \) and the \( \sigma^{ij} \) with the likelihood,

\[
p(\theta, \sigma^{ij} | \mathbf{y}) \propto (2\pi)^{-nk/2} |\Delta|^{-1/2} \exp \left(-\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma^{ij} v_{ij} \right) d\theta \prod_i d\sigma^{ij}.
\]

(3.2)

To find the marginal distribution of \( \theta \) we can integrate out the \( \sigma^{ij} \) by comparing the right hand side of the joint posterior distribution with the Wishart distribution (Wilks, 1962, page 551, equation 18.2.27), noting that, while Wilks' variables are the \( v_{ij} \), ours are the \( \sigma^{ij} \). It follows that

\[
p(\theta | \mathbf{y}) = C |v_{ij}|^{-n/2}
\]

where

\[
C = \int |v_{ij}|^{-n/2} \, d\theta
\]

is the normalising constant.

We find, therefore, a remarkably simple form for the posterior density involving, apart from a constant, only a power of the determinant of the dispersion matrix which is, of course, a function of \( \theta \) only. In particular, we notice that an appropriate generalisation of least squares is immediately provided. If we wish to obtain point estimates \( \hat{\theta}_h \) for the \( \theta_h \), we should minimise the value of the determinant \( |v_{ij}| \). These estimates, corresponding to the maximum posterior density are also, of course, maximum likelihood estimates.

It is of interest to note the relation between this criterion and the one that would be appropriate if the variance covariance matrix were known.

If \( v_{ij} \) is the cofactor of \( v_{ij} \) in \( |v_{ij}| \), we can write
\[ z_0 = |v_{ij}| = \sum_{i=1}^{k} v_{ij} \bar{v}_{ij} \sum_{j=1}^{k} v_{ij} \bar{v}_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} \bar{v}_{ij} / k \]

thus \( z_0 \) is of the same form as \( z = \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} \sigma_{ij} \) but with quantities proportional to the maximum likelihood estimates of the \( \sigma_{ij} \) replacing the unknown weights \( \sigma_{ij} \).

We now consider this general result in the light of some familiar special cases.

4. **SPECIAL CASES**

4.1 **Univariate linear regression model**

When \( k = 1 \) and \( f_i (x_{1u}, \theta) = \sum_{q=1}^{m} \theta_q x_{1u}^q \), the posterior distribution is readily shown to be

\[
p(\theta) = \text{constant} \left\{ 1 + \frac{(\theta - \hat{\theta})' X' X (\theta - \hat{\theta})}{v s^2} \right\}^{-n/2}
\]

where \( v = n-m \) and \( (n-1)s^2 = \sum_{u=1}^{n} (y_{1u} - \sum_{q=1}^{m} \hat{\theta}_q x_{1u}^q)^2 \), i.e., the well known multivariate - t distribution (Cornish 1954; Dunnett and Sobel 1954).

4.2 **Multivariate independent observations, with linear regression models**

Suppose that for the \( i \)th element of the observation vector there is a linear model, so that

\[
f_i (x_{1u}^q, \theta) = \sum_{q=1}^{m} \theta_q x_{1u}^q, \quad i = 1, 2, \ldots, k.
\]
Suppose further that the k elements of the observation vector 
\( \{y_{1u}, y_{2u}, \ldots, y_{ku}\} \) are independently distributed so that 
\( \sigma_{ij} = 0, \ i \neq j \). Then the posterior distribution of the common \( \theta \) is 
given by the product

\[
p(\theta) = \text{constant} \prod_{i=1}^{k} \left[ 1 + \frac{X_i'X_i(\theta - \hat{\theta}_i)}{\nu s_i^2} \right],
\]

where \( \nu = n-m \), \( s_i^2 = \sum_{u=1}^{n} (y_{iu} - \sum_{q=1}^{m} \theta_q x_{iu}q)^2 \),

\( \hat{\theta}_i \) is the maximum likelihood estimate of \( \theta \) from the \( i \)th set of observations \( i = 1, 2, \ldots, k \), and \( (x_{1u}, x_{2u}, \ldots, x_{mu}) \) is the \( u \)-th row 
\( (u = 1, 2, \ldots, n) \) of \( X_i \), i.e., by the product of \( k \) independent 
multivariate t-distributions.
4.3 Distribution of the weighted mean

Suppose in the previous case \( k = 2, \ m = 1 \) and the models are given by \( f_1 = f_2 = \theta \). The distribution of \( \theta \) is now the product of two univariate t distributions

\[
p(\theta) = \text{constant} \prod_{i=1}^{2} \left( 1 + \frac{(\theta - \hat{\theta}_i)^2}{\nu s_i^2} \right)^{-n/2}
\]

where \( \hat{\theta}_i \) represents the maximum likelihood estimate of \( \theta \) from the \( i \)th set of observations \( i = 1, 2 \). This corresponds to the fiducial distribution of the weighted mean given by Fisher (1961). The weighted mean problem is the simplest example of the more general problem of combining information from several responses. It has an interesting history which we shall mention briefly.

Consider the two models \( f_1 = \theta_1, f_2 = \theta_2 \). The joint posterior distribution of \( y_1 \) and \( y_2 \) would be the product of the two student densities

\[
\prod_{i=1}^{2} \left( 1 + \frac{(\theta_1 - \hat{\theta}_1)^2}{\nu s_i^2} \right)^{-n_i/2}
\]

Suppose we make a change of variables to

\[
\phi_1 = \frac{\theta_1}{s_1^2} + \frac{\theta_2}{s_2^2}
\]

\[
\phi_2 = \theta_1 - \theta_2
\]

On integrating out \( \phi_1 \) we obtain, as is well known, theBehrens–Fisher distribution for the difference of two means. If instead we integrate out
\( \phi_2 \) we obtain, as pointed out by Yates (1939), another Behrens–Fisher distribution. Yates put forward this latter as the distribution of the weighted mean. Fisher (1961) points out, however, that, since in this problem it is given that \( \theta_1 = \theta_2 \), we should not integrate over the joint distribution of \( \theta_1 \) and \( \theta_2 \) but should instead consider the conditional distribution on the line \( \theta = \theta_1 = \theta_2 \). This solution corresponds precisely with our formulation in the special case we mention. Fisher (1961) found, as might be expected, that the distribution of the common \( \theta \) depended very heavily on the size of the difference \( \hat{\theta}_1 - \hat{\theta}_2 \) i.e. on the apparent compatibility of the two estimates. This question of compatibility will also appear as an important desideratum in our examples.

5. AN EXAMPLE OF THE MULTIVARIATE ESTIMATION OF A SINGLE PARAMETER \( \theta \) IN A NON-LINEAR SYSTEM

Consider a chemical reaction in which a product \( A \) is decomposing to form \( B \) in such a way that the rate of reaction is proportional to the proportion \( \eta_1 \) of \( A \) left unchanged. With \( \eta_2 \) representing a proportion of \( A \) which has reacted to form \( B \), the system is described by the differential equations

\[
-\eta_1 = \phi \eta_1 = \bar{\eta}_2, \quad \phi > 0
\]  

(5.1)

where the dot denotes differentiation with respect to time \( t \), with boundary conditions

\[
\eta_1 = 1, \quad \eta_2 = 0 \text{ at } t = 0
\]

\[
\eta_1 = 0, \quad \eta_2 = 1 \text{ at } t = \infty
\]
The equations (5.1) have solution
\[ \eta_1 = e^{-\phi t}, \quad \eta_2 = 1 - e^{-\phi t} \]

We observe \( y_1 \) and \( y_2 \) where
\[ y_{1u} = \eta_{1u} + \epsilon_{1u}, \quad u = 1, 2, \ldots, n, \quad i = 1, 2, \]
and where \( E(\epsilon_{1u}) = 0, \quad E(\epsilon_{1u}^2) = \sigma_{11}^2, \quad E(\epsilon_{1u} \epsilon_{2u}) = \rho \sigma_{11} \sigma_{22} = \sigma_{12}^2 \), where \( \sigma_{11}, \sigma_{22} \) and \( \rho \) are unknown.

In considering a parameter like the specific rate \( \phi \) which is essentially positive, it is probably most realistic to take \( \theta = \ln \phi, -\infty < \theta < \infty \), as locally uniform a priori. This would mean, for example, that having guessed a value for \( \phi \), an experimenter would be about equally preferred to accept a value twice as big as he would to accept one half as big.

Suppose two observations are taken at each of five distinct values of the time \( t \). If a single response, \( y_1 \) alone, or \( y_2 \) alone was available, the posterior distribution of \( \theta \) will be given by \( p(\theta | y_i) \propto v_{11}^{-5} \), \( (i = 1, 2) \). If however both \( y_1 \) and \( y_2 \) were observed, the posterior distribution which now makes use of information from both sources will be
\[ p(\theta | y) \propto (v_{11} v_{22} - v_{12}^2)^{-5}. \]

A set of manufactured data for this type of example is given in the second and third columns of table 1, labelled Example 1. These data were obtained by adding random normal deviates to calculated values and taking \( \sigma_{11} = 0.0004, \quad \sigma_{22} = 0.0016, \quad \sigma_{12} = 0.0004 \) so that the correlation coefficient between the errors was \( \rho_{12} = 0.5 \). Figure 1
shows the posterior distributions for $y_1$ alone, $y_2$ alone and for $y_1$ and $y_2$ taken together.

<table>
<thead>
<tr>
<th>$t_u$</th>
<th>$y_{1u}$</th>
<th>$y_{2u}$</th>
<th>$y_{1u}$</th>
<th>$y_{2u}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{2}$</td>
<td>.907</td>
<td>.127</td>
<td>.907</td>
<td>.05</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>.915</td>
<td>.064</td>
<td>.915</td>
<td>.15</td>
</tr>
<tr>
<td>1</td>
<td>.801</td>
<td>.134</td>
<td>.801</td>
<td>.15</td>
</tr>
<tr>
<td>1</td>
<td>.825</td>
<td>.200</td>
<td>.825</td>
<td>.25</td>
</tr>
<tr>
<td>2</td>
<td>.649</td>
<td>.274</td>
<td>.649</td>
<td>.55</td>
</tr>
<tr>
<td>2</td>
<td>.675</td>
<td>.375</td>
<td>.675</td>
<td>.40</td>
</tr>
<tr>
<td>4</td>
<td>.446</td>
<td>.570</td>
<td>.446</td>
<td>.65</td>
</tr>
<tr>
<td>4</td>
<td>.468</td>
<td>.535</td>
<td>.468</td>
<td>.70</td>
</tr>
<tr>
<td>8</td>
<td>.233</td>
<td>.792</td>
<td>.233</td>
<td>.70</td>
</tr>
<tr>
<td>8</td>
<td>.187</td>
<td>.803</td>
<td>.187</td>
<td>.85</td>
</tr>
</tbody>
</table>

Table 1: Data for examples, $k = 2$, $m = 1$. 
When we consider the posterior distributions from $y_1$ and $y_2$ separately, we see that they present consistent evidence concerning $\theta$. As is to be expected, the precision of the estimate from $y_1$ (which has the smaller variance) is greater than that from $y_2$. Also, even though $y_1$ and $y_2$ are correlated to some extent, the two responses taken together provide a distribution which is sharper than either of those from the individual responses.

**The effect of lack of fit**

In univariate problems, the importance of checking a model tentatively entertained by examination of residuals is becoming recognised. In multivariate problems, checking the model becomes even more important.

For each $y_i$, $i = 1, 2, \ldots, k$, there is a separate model function, any of which may not fit. It is important to realise that when lack of fit occurs for any response, the overall criterion may be upset. We now consider how this occurs.

Suppose the $t$-th function fits badly. Then lack of fit error is superimposed on variance error and the residual quantities $y_{tu} - f_t(x_{tu}, \theta)$ may become excessively large in magnitude even for the indicated "best" value of $\theta$. Thus the $v_{tj}$ ($=v_{jt}$) $j = 1, 2, \ldots, k$ will be affected and so will the cofactors $V_{tj}$ ($=V_{jt}$), $t = 1, 2, \ldots, k$. Now we have seen that in

$$z_0 = \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} V_{ij}/k,$$

the cofactors act as estimated weights so that lack of fit of one factor can affect the weight given to another. It becomes particularly important
therefore to check the fit of all the fitted functions by studying residuals. (A multivariate lack of fit test can be used in conjunction with this less formal examination. It is hoped to discuss this at a later time.)

A second manufactured example (Example 2, Table 1) will serve to demonstrate the situation which may arise when such lack of fit exists. In generating this data, the $y_1$ column was taken as before but a different value of $\theta$ was used to generate the $y_2$ column. On inspection of Figure 2A, we notice immediately that the evidence about $\theta$ from $y_1$ alone contradicts the evidence from $y_2$ alone. This of itself clearly indicates that either one or both of the models are inappropriate. It is instructive to consider the effect of this on the joint criterion. It appears that $y_1$ and $y_2$ together provide less precise evidence about $\theta$ than does $y_1$ alone. Furthermore the maximum value of $p(\theta | y_1, y_2)$ does not lie between the maxima of $p(\theta | y_1)$ and $p(\theta | y_2)$. The reason for this behavior is clear from consideration of the function

$$p(\theta | y_1, y_2) / p(\theta | y_1),$$

which is shown in Figure 2B. This is U-shaped over the important range of $\theta$. If we think of this ratio as the "additional information contributed by $y_2$" we see that because of lack of fit this "additional information" can, as it were, be negative. The inconsistencies mentioned are clearly shown by examining the residuals. These were calculated using the value of $\theta$ which maximizes $p(\theta | y_1, y_2)$ and are given below.
<table>
<thead>
<tr>
<th>$t$</th>
<th>$\epsilon_{1u} = y_{1u} - \hat{y}_{1u}$</th>
<th>$\epsilon_{2u} = y_{2u} - \hat{y}_{2u}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>-.043</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>.008</td>
<td>.057</td>
</tr>
<tr>
<td>1</td>
<td>-.020</td>
<td>-.029</td>
</tr>
<tr>
<td>1</td>
<td>.004</td>
<td>.071</td>
</tr>
<tr>
<td>2</td>
<td>-.027</td>
<td>.226</td>
</tr>
<tr>
<td>2</td>
<td>-.001</td>
<td>.076</td>
</tr>
<tr>
<td>4</td>
<td>-.011</td>
<td>.107</td>
</tr>
<tr>
<td>4</td>
<td>.011</td>
<td>.157</td>
</tr>
<tr>
<td>8</td>
<td>.025</td>
<td>-.092</td>
</tr>
<tr>
<td>8</td>
<td>-.021</td>
<td>.058</td>
</tr>
</tbody>
</table>

Figure 2C shows a plot of the residuals $\epsilon_{1u}$ against time; the behavior of $\epsilon$ strongly suggests abnormality.

6. AN EXAMPLE OF MULTIVARIATE ESTIMATION OF MORE THAN ONE PARAMETER IN A NON-LINEAR SYSTEM

The informative power of this multivariate approach is most striking when there is more than one parameter. This will be illustrated by a further example.

In the study of kinetic mechanisms, a chemist will often be concerned with estimating more than one parameter; for example the rate constants for the system, and will wish to do so using measurements of several chemical products. Perhaps the simplest example which can illustrate this situation is a reaction of the type $A \rightarrow B \rightarrow C$. If $\eta_1$, $\eta_2$ and $\eta_3$ represent the proportions of reactants $A$, $B$, and $C$ present at a particular time $t$, then the system may be described by the differential equations
\[\begin{align*}
\dot{\eta}_1 &= -\phi_1 \eta_1 \\
\dot{\eta}_2 &= \phi_1 \eta_1 - \phi_2 \eta_2 \\
\dot{\eta}_3 &= \phi_2 \eta_2
\end{align*}\]  
(6.1)

with boundary conditions \(\eta_1 = 1\), \(\eta_2 = \eta_3 = 0\) at \(t = 0\), and where the dot denotes differentiation with respect to time \(t\), and \(\phi_1\), \(\phi_2\) are unknown rate constants. Equations (6.1) have solution

\[\begin{align*}
\eta_1 &= e^{-\phi_1 t} \\
\eta_2 &= (e^{-\phi_1 t} - e^{-\phi_2 t}) \phi_1 / (\phi_2 - \phi_1) \\
\eta_3 &= 1 + (-\phi_2 e^{-\phi_1 t} + \phi_1 e^{-\phi_2 t}) / (\phi_2 - \phi_1)
\end{align*}\]  
(6.2)

As before, it is probably most natural for the experimenter to think in terms of the logarithms of the rate constants, \(\theta_i = \ln \phi_i\), \(i = 1, 2\), and to regard these as locally uniformly distributed a priori.

Suppose observations \(y_1\), \(y_2\) and \(y_3\) of all three products were available, two independent multivariate observations being made at each of six distinct values of time \(t\), as shown in Table 2. We shall suppose these observations have arisen from 12 independent experimental runs, as would be the case if each run were carried out in a sealed tube, reaction being terminated at the appropriate time by sudden cooling. There are now three variances and three covariances, all unknown.
<table>
<thead>
<tr>
<th>$t_u$</th>
<th>$y_{1u}$</th>
<th>$y_{zu}$</th>
<th>$y_{su}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{2}$</td>
<td>.959</td>
<td>.025</td>
<td>.028</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>.914</td>
<td>.061</td>
<td>.000</td>
</tr>
<tr>
<td>1</td>
<td>.855</td>
<td>.152</td>
<td>.068</td>
</tr>
<tr>
<td>1</td>
<td>.785</td>
<td>.197</td>
<td>.096</td>
</tr>
<tr>
<td>2</td>
<td>.628</td>
<td>.130</td>
<td>.090</td>
</tr>
<tr>
<td>2</td>
<td>.617</td>
<td>.249</td>
<td>.118</td>
</tr>
<tr>
<td>4</td>
<td>.480</td>
<td>.184</td>
<td>.374</td>
</tr>
<tr>
<td>4</td>
<td>.423</td>
<td>.298</td>
<td>.358</td>
</tr>
<tr>
<td>8</td>
<td>.166</td>
<td>.147</td>
<td>.651</td>
</tr>
<tr>
<td>8</td>
<td>.205</td>
<td>.050</td>
<td>.684</td>
</tr>
<tr>
<td>16</td>
<td>.034</td>
<td>.000</td>
<td>.899</td>
</tr>
<tr>
<td>16</td>
<td>.054</td>
<td>.047</td>
<td>.991</td>
</tr>
</tbody>
</table>

Table 2: Data for example $k = 3$, $m = 2$
The estimation situation is completely portrayed by the posterior distributions. The joint posterior distributions for \( \theta_1 \) and \( \theta_2 \) which result from consideration of a single product, a pair of products jointly, or all three products together may be examined. In view of the fact that \( y_1 \) supplies information on the parameter \( \theta_1 \) only, we have examined only the following distributions: \( p(\theta_1, \theta_2 | y_2) \), \( p(\theta_1, \theta_2 | y_3) \), \( p(\theta_1, \theta_2 | y_2, y_3) \), and \( p(\theta_1, \theta_2 | y_1, y_2, y_3) \). It is particularly instructive to consider these in the form of superimposed contour diagrams. However, it is confusing if a large number of contours are plotted and for clarity, we show in Figure 3 only a single contour for each distribution. In each case, this is the contour which includes approximately 99.75% of the posterior distribution. This Bayesian region is given by that contour for which

\[
\log p(\hat{\theta}) - \log p(\theta) = \frac{1}{2} \chi^2_{2} (1- \alpha), \text{ where } \alpha = .0025.
\]

It is easy to see that this would give the precise region if the joint distribution were multivariate normal or could be made multivariate normal by any transformation of the \( \theta \)'s. Figure 3 repays careful study and provides a very good illustration of the value of considering the whole posterior distribution rather than a single point estimate. Essentially the same point has been repeatedly made by Barnard in discussing the likelihood principle. In any sequential reaction \( A \rightarrow B \rightarrow C \ldots \text{etc.} \ldots \), we should expect that observation of only the end product would provide accurate information on only the sum, or some such function of the specific rates. In this particular example we see, by inspection of the solution
for $\eta_3$ in equations (6.2), that $\eta_3$ is symmetric in $\phi_1$ and $\phi_2$ and hence in $\theta_1$ and $\theta_2$. It follows that, when $y_3$ alone is considered, if any point $(P, Q)$ is included in the Bayesian region, so must be its mirror image $(Q, P)$ in the $\theta_1 = \theta_2$ axis. (This can lead, as in this example, to a double maximum.) This feature is illustrated in Figure 3, where the long, crescent-shaped contour taken from $p(\theta | y_3)$ is symmetrical with respect to the $\theta_1 = \theta_2$ line.

The information supplied by the intermediate product $y_2$ is of essentially different character. From it we obtain information on both specific rates but principally on the difference or ratio of the rates. It is noticeable that this region is obliquely oriented in a direction approximately at right angles to that of $p(\theta | y_3)$.

When information from $y_2$ & $y_3$ is combined, we find, as would be expected, a much smaller region contained within the intersection of the regions for $y_2$ alone and $y_3$ alone. Finally we can consider the effect of adding $y_1$ which provides information on $\theta_1$ only. Study of the probability contour from $p(\theta_1, \theta_2 | y_1, y_2, y_3)$ shows, again as would be expected, that the region is changed principally in being narrowed in the $\theta_1$ direction. Table 3 shows, in the second and third columns, the maximum likelihood estimates available from the various sources.

This example perhaps serves to illustrate the caution and common sense which ought to be applied in analysing this type of problem and the necessity for considering each problem individually and not hoping for "automatic" answers from a computer program. The authors know of one case where
an elaborate consecutive mechanism involving ten constants was pos-
tulated, but observations were taken on the single end product only
and on no intermediate product.

An iterative non-linear estimation routine run on a computer converged
only very slowly on to what appeared to be nonsensical answers. This
was undoubtedly due to the peculiarities of the ten-dimensional likeli-
hood surface, which probably contained multidimensional ridges and multi-
ple maxima. Observations on additional responses and application of
the theory of this paper, would have eliminated many, if not all, of the
ambiguities.

In the complete set of data just used the reaction is almost complete
when the last observation is taken. In practice, for one reason or another,
data of this kind occur in which the available observations trace only part
of the course of the reaction. We have illustrated the effect of this by
omitting the last four observations on \( y_1 \), \( y_2 \), and \( y_3 \) and repeating
the analysis. As shown in Figure 4, over the ranges studied, the contours
for \( p(0|y_2) \) and \( p(0|y_3) \) do not close. Nevertheless quite precise estima-
tion is possible using \( y_2 \) and \( y_3 \) together and the addition of \( y_1 \) improves
the estimation further. The maximum likelihood estimates are given in
Table 3.
Table 3: Maximum likelihood estimates for \( \theta \) and \( \phi \)

<table>
<thead>
<tr>
<th></th>
<th>Example 3 ( n=12 )</th>
<th>Example 4 ( n=8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_2 )</td>
<td>((-1.561, -0.685))</td>
<td>((-1.715, -0.942))</td>
</tr>
<tr>
<td>( y_3^* )</td>
<td>((-1.619, -0.635))</td>
<td>((-1.273, -1.022))</td>
</tr>
<tr>
<td>( y_2, y_3 )</td>
<td>((-1.585, -0.695))</td>
<td>((-1.561, -0.701))</td>
</tr>
<tr>
<td>( y_1, y_2, y_3 )</td>
<td>((-1.570, -0.707))</td>
<td>((-1.565, -0.707))</td>
</tr>
</tbody>
</table>

*Note: For \( y_3^* \) alone, interchange of the coordinates for \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) (and for \( \hat{\phi}_1 \) and \( \hat{\phi}_2 \)) also provides maximum likelihood estimates; see text.
Appendix (see Section 3)

We use results due to Hsu, given by Deemer and Olkin (1951); the full notation of that paper is not required however.

**Result:** If \( \Sigma = \{ \sigma_{ij} \} \) and \( A = \Sigma^{-1} = \{ \sigma_{ij}^{-1} \} = \{ \sigma^{ij} \} \)

are two symmetric matrices of order \( k \), then the Jacobian

\[
\frac{\partial (\sigma_{ij})}{\partial (\sigma^{ij})} = D(\sigma_{ij}, \sigma^{ij}) = D(A, \Sigma)
\]

has value \( |A|^{-(k+1)} \).

**Proof:** By Property 5B.3, page 366 of Deemer and Olkin, \( D(A, \Sigma) = D(A^*, \Sigma^*) \),

where \( A^* = dA \) and \( \Sigma^* = d\Sigma \) are the matrices of differentials. Taking differentials of \( \Sigma = A^{-1} \) provides

\[
\Sigma^* = -A^{-1} A^* A^{-1}
\]

by Theorem 4.4, page 357, of Deemer and Olkin. By Theorem 3.7 and Corollary 3.7, page 348 of Deemer and Olkin, and noting that since \( A \) is symmetric, \( A^{-1} = (A')^{-1} \) we see that \( D(A^*, \Sigma^*) = |A|^{-(k+1)} \)

whence the result follows.
REFERENCES


REFERENCES


Example 2

Figure 2A: Posterior Distributions
Figure 2 B: Plot of $p(\theta | y_1, y_2) / p(\theta | y_1)$ against $\theta$, example 2
Figure 2 C: Residuals Plotted Against Time, Example 2
Figure 3: 99.75% Confidence Regions
For $\theta_1$ and $\theta_2$
For Example 3 ($N = 12$)
Figure 4: 99.7% Confidence Regions for $O_1$ and $O_2$ for Example 4 ($N = 8$)