MODELING INTERLABORATORY DIFFERENCES:
A BAYESIAN ANALYSIS

by

Irwin Guttman and Ingram Olkin

Technical Report No. 261
June 1989

Prepared under the Auspices
of
National Science Foundation
DMS 87-08083
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
MODELING INTERLABORATORY DIFFERENCES:
A BAYESIAN ANALYSIS

by

Irwin Guttman and Ingram Olkin

University of Toronto and Stanford University

Abstract

It has long been observed that independent laboratories differ in reporting the results of repeated experiments. The problem is to detect those laboratories that might be considered aberrant. The bulk of studies dealing with this model have been based on a linear structure. The present analysis takes a different view for which a Bayesian analysis is natural.

Key words and phrases: Detection of outliers, analysis of variance models.
MODELING INTERLABORATORY DIFFERENCES:
A BAYESIAN ANALYSIS

by

Irwin Guttman and Ingram Olkin

University of Toronto and Stanford University

1. Preliminaries

A prototype experiment is one in which a sample is divided into \( M \) parts, one for each of \( M \) laboratories. The \( j \)-th laboratory then subdivides its part into \( m_j, j = 1, \ldots, M \), units each of which is tested. The results of the experiment consist of measurements \( y_{ij}, (i = 1, \ldots, m_j; j = 1, \ldots, M) \).

Figure 1 provides an example of such an experiment. This figure exhibits some radical differences between laboratories, as well as more modest differences within laboratories. Since some of the laboratories are similar the problem is to distinguish between those laboratories that are similar and those that are aberrant.

A standard procedure is to remove the aberrant laboratories and to use the remaining laboratories that are similar as a basis for estimating a common mean value. This problem is not new and has been studied from a variety of points of view.

The problem of interlaboratory differences was already noted forty years ago by Grubbs (1948). Later Youden (1959) provided data from 34 laboratories for chemical analysis on a granite rock and a dibase rock in which there were clear indications of discrepancies in the results. The method of analysis was to remove outliers, and to use the remaining observations for subsequent analyses.

John Mandel in a series of papers [Mandel (1958), Mandel and Lashof (1959), Mandel and Paule (1970), Mandel (1976), Mandel (1978)] has relied on a linear model analysis of variance framework. In so doing some parameters represent laboratories, some represent observations, and there may be interaction terms. Several variants are also available, see e.g., Jaech (1979). Kramer (1967) uses a mean deviation in an ANOVA context. Maximum
likelihood estimators for some linear models are obtained by Jordan and de Alvaro (1979). Rocke (1983) provides a method of analysis that is robust to the existence of outliers and long-tailed distributions of random effects.

A different approach was proposed by Olkin and Sobel (1987). Here a ranking and selection procedure was used to generate a subset that excludes aberrant laboratories. The present approach is similar in that the goal is the detection of aberrant laboratories. However, the approach adopts a Bayesian analysis to accomplish this.

Assume that an unknown number, $k$, of the laboratories are aberrant. Normally, $k$ is small relative to the total number $M$ of laboratories. If $k$ is permitted to be large then it is not clear which laboratories are aberrant and which are not. For example, when $k = 0$ or $k = M$, then all the laboratories are homogeneous, but we cannot distinguish whether they are aberrant or not. Thus, we normally expect that $1 \leq k < M/2$. Because of the symmetry between aberrant and non-aberrant laboratories, we need not be concerned about the case $k = 0$ or $k = M$.

To accommodate the potential for interlaboratory differences assume that

$$Ey_{ij} = \mu + \alpha_j, \quad j = 1, \ldots, M,$$

where the $\alpha$'s represent laboratory effects.

In the simplest model we assume that the laboratories function independently, and that replications within each laboratory are independent with a common variance $\sigma^2$. This assumption can be relaxed. However, the essential ingredients of the proposed analysis are most readily illustrated in this model.

Suppose for the moment that $k$ is fixed but unknown. There are $\binom{M}{k}$ possible partitions of the $M$ laboratories into $k$ groups. For each partition we compute the posterior probability that laboratories $j_1, \ldots, j_k$ have means $\mu + \alpha_{j_\ell}, \ell = 1, \ldots, k$, and that the remaining laboratories have a common mean $\mu$. For simplicity of notation we write $\{1, 2, \ldots, k\}$ for $\{j_1, j_2, \ldots, j_k\}$ whenever no confusion arises. Let $\mathcal{K}$ denote the set of $k$ indices $(i_1, \cdots, i_k)$, this set has $\binom{M}{k}$ elements. We write $\sum_{\mathcal{K}}$ to denote a sum over the elements in $\mathcal{K}$.

Denote the posterior probabilities as $C(j_1, \ldots, j_k)$, or more simply as $C(1, \ldots, k)$. Our
proposed procedure is to determine, for each fixed $k = 1, 2, \ldots, [M/2]$, the maximum of the $\binom{M}{k}$ posterior probabilities. The set of maxima will then permit a determination of the aberrant laboratories. The value $k_0$ of $k$ that satisfies

$$
\max_k \max_{j_1, \ldots, j_k} C(j_1, \ldots, j_k) \equiv C^*(1, \ldots, k_0)
$$

is used to determine the number of aberrant laboratories.

In normal practice we expect the sequence of maximum posterior probabilities $C^*(1), C^*(1, 2), \ldots, C^*(1, 2, \ldots, [M/2])$ to be either monotone or unimodal. In either case the value of $k_0$ is readily obtained numerically.

2. A Bayesian Approach

As indicated previously, we are concerned with the case of $M$ laboratories that generate observations $y$ on a response variable, where there is the possibility that $k$ of the laboratories produce biased results, modeled as follows. Let $y_{ij}$, $i = 1, \ldots, m_j$, $j = 1, \ldots, M$ be the $i$-th determination of $y$ from laboratory $j$. For a set of laboratories labeled $1, 2, \ldots, M$, assume that

\begin{equation}
(2.1) \quad y_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad j = 1, \ldots, k,
\end{equation}

\begin{equation}
(2.1) \quad y_{ij} = \mu + \varepsilon_{ij}, \quad j = k + 1, \ldots, M,
\end{equation}

for which the $\varepsilon_{ij}$'s are independent, identically distributed normal random variables with a common variance $\sigma^2$, and where $\alpha_1, \ldots, \alpha_k$ are the non-zero shift parameters. We use the notation $\mathcal{M}_{\{k\}}$ to denote the model (2.1) and to indicate that laboratories $1, 2, \ldots, k$ generate spurious observations in the sense that there is a shift of the mean.

The likelihood function for model (2.1) is proportional to

\begin{equation}
L(\mathcal{M}_{\{k\}}; \mu, \alpha; \sigma^2 \mid y) \propto \frac{1}{\sigma^m} \exp - \frac{1}{2\sigma^2} \sum_{j=1}^{M} \sum_{i=1}^{m_j} (y_{ij} - \bar{y}_{.j})^2 \times \exp - \frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{k} m_j (\bar{y}_{.j} - \mu - \alpha_j)^2 + \sum_{k+1}^{M} m_j (\bar{y}_{.j} - \mu)^2 \right\},
\end{equation}
where \( \alpha = (\alpha_1, \ldots, \alpha_k) \), \( \bar{y}_j = \sum_{i=1}^{m_j} y_{ij}/m_j \), and \( m = \sum_{1}^{M} m_j \).

Assume that to a good approximation the priors on \( \mu, \alpha, \sigma^2 \), and \( M_{\{k\}} \) are independent with

\[
p(M_{\{k\}}; \mu, \alpha; \sigma^2) \propto 1/\sigma^2
\]

so that we are employing the usual non-informative priors on \( \mu \) and \( \alpha \). Then the posterior \( M_{\{k\}} \) given the data, and of course \( k \), is such that

\[
p(M_{\{k\}}; \mu, \alpha; \sigma^2 \mid \text{data}) \propto \frac{1}{\sigma^{m+2}} \exp - \frac{1}{2\sigma^2} \left\{ S_W + \sum_{1}^{k} m_j(\bar{y}_j - \mu - \alpha_j)^2 + \sum_{k+1}^{M} m_j(\bar{y}_j - \mu)^2 \right\},
\]

where

\[
S_W = \sum_{j=1}^{M} \sum_{i=1}^{m_j} (y_{ij} - \bar{y}_j)^2
\]
is within the sum of squares. Hence

\[
p(M_{\{k\}}; \mu, \sigma^2 \mid \text{data}) \propto \frac{1}{\sigma^{m-k+2}(\prod_{1}^{k} m_j)^{1/2}} \exp - \frac{1}{2\sigma^2} \left\{ S_W + \sum_{k+1}^{M} m_j(\bar{y}_j - \mu)^2 \right\}.
\]

It is straightforward to verify that

\[
\sum_{k+1}^{M} m_j(\bar{y}_j - \mu)^2 = \bar{m}(\bar{y} - \mu)^2 + \sum_{k+1}^{M} m_j(\bar{y}_j - \bar{y})^2,
\]

where \( \bar{m} = \sum_{k+1}^{M} m_j \), \( \bar{y} = \sum_{k+1}^{M} m_j \bar{y}_j/\bar{m} \). Integration of (2.5) with respect to \( \mu \) yields

\[
p(M_{\{k\}}; \sigma^2 \mid \text{data}) \propto \frac{1}{\sigma^{m-k+1}(\prod_{1}^{k} m_j)^{1/2} \bar{m}^{1/2}} \exp - \frac{1}{2\sigma^2} \tilde{S},
\]

where \( \tilde{S} = S_W + \tilde{S}_B \), \( \tilde{S}_B = \sum_{k+1}^{M} m_j(\bar{y}_j - \bar{y})^2 \) is a between sum of squares, and \( \tilde{S} \) is the total sum of squares that includes the between sums of squares based on the subscripts \( k + 1, \ldots, M \).

Finally, integration of (2.7) with respect to \( \sigma^2 \) yields
\begin{equation}
C(1, \ldots, k) \equiv p(\mathcal{M}_k \mid \text{data}) = c \left( \bar{S}^{m-k-1} \prod_{1}^{k} m_j^{1/2} \bar{m}^{1/2} \right)^{-1};
\end{equation}

where the constant, \( c \), is obtained from the expression

\begin{equation}
c^{-1} = \sum_{\mathcal{K}} \left( \bar{S}^{m-k-1} \prod_{1}^{k} m_j^{1/2} \bar{m}^{1/2} \right)^{-1}.
\end{equation}

Note that if \( m_1 = m_2 = \ldots = m_M \), then \( \prod_{1}^{k} m_j^{1/2} \bar{m}^{1/2} \) is constant so that the sum in (2.9) requires only the determination of \( \bar{S} \) over all partitions.

To estimate \( k \) \((k \geq 1)\) we determine

\begin{equation}
\max_{C(1, \ldots, k) = C^*(1, \ldots, k)},
\end{equation}

where the maximum is over the set of \( \mathcal{K} \) indices. For each value of \( k \), \( C^*(1, \ldots, k) \) tells us which set of \( k \) laboratories generates spurious observations. Then conditional on \( k \geq 1 \), we determine

\begin{equation}
\max_{1 \leq k \leq [M/2]} C^*(1, \ldots, k).
\end{equation}

which is designed to single out that value of \( k \) that tells us the number of aberrant laboratories.

3. The Marginal Posterior Distribution of the Mean \( \mu \).

As noted before, a primary goal in the use of the model (2.1) is to make inferences about the parameter \( \mu \). To do this suppose that conditional on \( k \), data is generated as in (2.1). The likelihood, conditional on \( k \), is proportional to

\begin{equation}
L \propto \sum_{\mathcal{K}} \frac{1}{\sigma^m} \exp - \frac{1}{2\sigma^2} \left\{ \bar{S} + \bar{m}(\bar{y} - \mu)^2 + \sum_{1}^{k} m_j(\bar{y}_j - \mu - \alpha_j)^2 \right\}.
\end{equation}

Combining the likelihood with the prior of (2.3), followed by an integration with respect to \( \alpha \) yields

\begin{equation}
p(\mu, \sigma^2 \mid \text{data}, k) \propto \sum_{\mathcal{K}} \frac{1}{\sigma^{m-k+2}} \prod_{1}^{k} \frac{1}{m_j^{1/2}} \exp - \frac{1}{2\sigma^2} \left\{ \bar{S} + \bar{m}(\bar{y} - \mu)^2 \right\}.
\end{equation}
A second integration with respect to $\sigma^2$ yields

\begin{equation}
(3.3) \quad p(\mu | \text{data}, k) \propto \sum_{\mathcal{K}} \frac{1}{\prod_{j=1}^{k} m_j^{1/2} \tilde{S}^{m-k/2}} \left(1 + \frac{\tilde{m}(\tilde{y} - \mu)^2}{\tilde{S}}\right)^{-\frac{m-k}{2}},
\end{equation}

or equivalently,

\begin{equation}
(3.4) \quad p(\mu | \text{data}, k) \propto \sum_{\mathcal{K}} \frac{1}{\prod_{j=1}^{k} m_j^{1/2} \tilde{m}^{1/2} \tilde{S}^{m-k-1}} \left(\frac{(\tilde{m}/\tilde{S})^{1/2} \Gamma\left(\frac{m-k}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{m-k-1}{2}\right)} \frac{1}{[1 + \left(\tilde{y} - \mu\right)^2]^{m-k/2}} \right).
\end{equation}

Using (2.8) and (2.9), (3.4) can be rewritten as

\begin{equation}
(3.5) \quad p(\mu | \text{data}, k) = \sum_{\mathcal{K}} C(1, \ldots, k) f(\mu | \tilde{y}, \tilde{m}, \tilde{S}; m - k - 1),
\end{equation}

where $f$ denotes the general univariate $t$-density

\begin{equation}
(3.6) \quad f(x | a, b, c; d) = \frac{\sqrt{b} \Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi} c \Gamma\left(\frac{d}{2}\right)} \left\{1 + \frac{b(x - a)^2}{c}\right\}^{-\frac{d+1}{2}}.
\end{equation}

Using properties of the $t$-density, from (3.5) we readily obtain

\begin{equation}
(3.7) \quad \mathcal{E}(\mu | \text{data}, k) = \sum_{\mathcal{K}} C(1, \ldots, k) \tilde{y},
\end{equation}

\begin{equation}
\mathcal{E}(\mu^2 | \text{data}, k) = \sum_{\mathcal{K}} C(1, \ldots, k) \left[\tilde{y}^2 + \frac{\tilde{S}}{\tilde{m}(m - k - 3)}\right],
\end{equation}

from which we can calculate $\text{Var}(\mu | \text{data}, k)$.

In the special case $k = 0$, when there are no aberrant laboratories,

\begin{equation}
\mathcal{E}(\mu | \text{data}, k = 0) = \bar{y},
\end{equation}

\begin{equation}
\text{Var}(\mu | \text{data}, k = 0) = (m - 1)S^2/[m(m - 3)],
\end{equation}

where $S^2 = \sum_{i=1}^{M} \sum_{j=1}^{m_j} (y_{ij} - \bar{y})^2/(m - 1)$. 
It is now routine to calculate $\text{Var}(\mu \mid \text{data}, k)$ for each $k$, and to determine

$$\min_k \text{Var}(\mu \mid \text{data}, k).$$

Both this criterion and that of (2.10) help in determining the final choice for $k$.

4. An Example

Vapor pressure induced when heating a standard amount of dinitro toluene solution to the given temperature of $25^\circ C$ is observed at each of eight laboratories. Seven laboratories make 3 determinations and one laboratory makes 2 determinations, the results of which are in Table 1. The data is plotted in Figure 1.

**Table 1**

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Readings</th>
<th>Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13.91</td>
<td>17.39</td>
</tr>
<tr>
<td>B</td>
<td>126.96</td>
<td>128.27</td>
</tr>
<tr>
<td>C</td>
<td>27.83</td>
<td>31.31</td>
</tr>
<tr>
<td>D</td>
<td>17.02</td>
<td>20.15</td>
</tr>
<tr>
<td>E</td>
<td>16.52</td>
<td>20.58</td>
</tr>
<tr>
<td>F</td>
<td>151.30</td>
<td>153.04</td>
</tr>
<tr>
<td>G</td>
<td>40.02</td>
<td>41.17</td>
</tr>
<tr>
<td>H</td>
<td>17.47</td>
<td>18.58</td>
</tr>
</tbody>
</table>

**Figure 1 Here**

For this data, the complete listing of results of $C(1, \ldots, k)$ of (2.8) - (2.9) for the $\binom{8}{k}$ combinations was made, for which the largest with $k = 1, 2, 3, 4$ are
Table 2

Determination of the Number $k$ of Spurious Results

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\max C(j_1, \ldots, j_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.985 = $C(6)$</td>
</tr>
<tr>
<td>2</td>
<td>.996 = $C(2, 6)$</td>
</tr>
<tr>
<td>3</td>
<td>.238 = $C(2, 6, 7)$</td>
</tr>
<tr>
<td>4</td>
<td>.098 = $C(2, 3, 6, 7)$</td>
</tr>
</tbody>
</table>

Note that the maximum values are unimodal in $k$ and that

$$C^*(1, 2, 3, 4) = \max_k C(j_1, \ldots, j_k) = C(2, 6).$$

Hence, we estimate $k = 2$, and assert that laboratories 2 (laboratory B) and 6 (laboratory F) generate observations that are dissimilar to those of other laboratories.

With $k = 2$ fixed the conditional posterior $p(\mu|\text{data, } k = 2)$ of $\mu$ is obtained from (3.5). Its graph is given in Figure 2.

Figures 2 and 3 Here

From (3.5) we have

$$E(\mu \mid \text{data, } k = 2) = 24.87,$$
$$\text{Var}(\mu \mid \text{data, } k = 2) = 5.06.$$

The conditional posterior $p(\mu|\text{data, } k = 0)$ is graphed in Figure 3. Its mean and variance are (see (3.8))

$$E(\mu \mid \text{data, } k = 0) = 50.52,$$
$$\text{Var}(\mu \mid \text{data, } k = 0) = 123.56.$$

The densities for the cases $k = 0$ and $k = 2$ can be seen to be divergent, as are the means and variances. It is clear from Table 1 or Figure 2 that the choice $k = 0$ is unwarranted, and that the estimate $k = 2$ confirms an intuitive analysis.

An inspection of the means in the data of Table 1 indicates a separation of the laboratories into two groups: laboratories (1,8, 4,5, 3,7) or (A, H, D, E, C, G) and (2,6) or
(B,F) with means \((17.4, 18.6, 20.2, 20.6, 31.3, 41.2)\) and \((128.3, 153.0)\), respectively. This raises the question whether laboratories 3 and 7 are in some way anomalous. The analysis based on posterior probabilities point up this possibility, but ultimately do not label these as dissimilar.

5. The Posterior Distribution of \(k\)

Our starting point is the model in (2.1), for which we presume that the prior information, as conveyed by \(p(k, \mu, \alpha, \sigma^2)\), has the following structure:

\[
p(k, \mu, \alpha, \sigma^2) = p(\mu, \sigma^2) p(k, \alpha | \mu, \sigma^2) = p(\mu, \sigma^2) p_k(k) p(\alpha | k, \sigma^2)
\]

with \(-\infty < \alpha_j < \infty, j = 1, \ldots, k; -\infty < \mu < \infty; \sigma^2 > 0, 0 \leq k \leq K_*\) and where \(K_*\) is known in advance (\(K_*\) usually \(\ll [M/2]\)). Here, we assume that

\[
p(\mu, \sigma^2) = C(n_0, v_0, s_0^2)(\sigma^2)^{-n_0/2+1} \exp\left\{-\frac{1}{2\sigma^2} [v_0 s_0^2 + n_0 (\mu - \mu_0)^2]\right\},
\]

where

\[
C(n_0, v_0, s_0^2) = \frac{\sqrt{n_0} (v_0 s_0^2)^{(n_0-1)/2}}{2^{n_0/2} \sqrt{\pi} \Gamma\left(\frac{n_0}{2}\right)},
\]

\[
p_k(k) = C(L, K_*) \Lambda^k / k!, 0 \leq k \leq K_*,
\]

with \(\Lambda\) given and

\[
p(\alpha | k, \sigma^2) = \begin{cases} 1 & \text{if } \alpha = 0, \\ (2\pi C_\sigma \sigma^2)^{-k/2} \exp\left\{-\frac{1}{2 C_\sigma \sigma^2} \sum_{j=1}^{k} (\alpha_j - \alpha_j 0)^2\right\} & 1 \leq k \leq K_*, \end{cases}
\]

where \(C_\sigma\) is a given constant, \(\alpha_j 0, j = 1, \ldots, k\) are given numbers, and \([C(L, k)]^{-1} = \sum_{k=0}^{K_*} \Lambda^k / k!\). That is, if \(k \geq 1\), then \(\alpha\) has a distribution whose components are such that

\[V(\alpha_j) / \sigma^2 = C_\sigma,\text{ a known number.}[\text{The choice of the prior for } k \text{ could be otherwise, as for example, } p_k(k) = 1 / (K_* + 1), \text{ } k = 0, 1, \ldots, K_*].\]

Once data has been observed as in (2.1), we have

\[
p(k, \mu, \alpha, \sigma^2 | \text{data} ) = C p(\mu, \sigma^2) \sum_{k \sim} p(k, \alpha | \mu, \sigma^2) p(y | k, \mu, \alpha, \sigma^2),
\]
where

\begin{equation}
C^{-1} = \sum_{k=0}^{K_\ast} p_K(k) \int_{\sigma^2} \int_{\mu} \int_{\alpha} \int_{\sigma^2} p(\alpha|k, \sigma^2) p(\mu, \sigma^2) p(y|k, \mu, \alpha, \sigma^2) d\alpha d\mu d\sigma^2.
\end{equation}

We establish the first term of (5.6) and then the \((k + 1)st\) term, \(1 \leq k \leq K_\ast\). The procedures necessary to do this will help us in determining \(p(k|\text{data})\), the posterior of \(k\), for after all,

\begin{equation}
p(k|\text{data}) = \int_{\sigma^2} \int_{\mu} \int_{\alpha} p(k, \mu, \alpha, \sigma^2|\text{data}) d\alpha d\mu d\sigma^2
= C p_K(k) \int_{\sigma^2} \int_{\mu} \int_{\alpha} p(\alpha|k, \sigma^2) p(\mu, \sigma^2) p(y|k, \mu, \alpha, \sigma^2) d\alpha, d\mu d\sigma^2;
\end{equation}

that is, the integration with respect to \(\alpha, \mu, \text{and } \sigma^2\) needed in (5.7) is the required integration needed in (5.6).

For the first term of (5.6), say \(u_0\), we have that

\begin{equation}
u_0 = p_K(0) \int_{\sigma^2} \int_{\mu} \int_{\alpha} p(\alpha|k = 0, \sigma^2) p(\mu, \sigma^2) p(y|k = 0, \mu, \alpha, \sigma^2) d\alpha d\mu d\sigma^2.
\end{equation}

Now using (5.4), we have that

\begin{equation}
u_0 = p_K(0) \int_{\sigma^2} \int_{\mu} 1 \times p(\mu, \sigma^2) p(y|k = 0, \mu, \alpha = 0, \sigma^2) d\mu d\sigma^2.
\end{equation}

It is straightforward to show that

\begin{equation}
p(y|k = 0, \mu, \alpha = 0, \sigma^2) = (2\pi\sigma^2)^{-m/2} \exp \left\{ -\frac{1}{2\sigma^2} \left\{ m(\mu - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m_j} (y_{ij} - \bar{y})^2 \right\} \right\}
\end{equation}

and using (5.10) in (5.9), we find that

\begin{equation}
u_0 = \frac{2^{n_0/2} C(\Lambda, K_\ast) C(n_0, v_0, s_0^2) \Gamma \left( \frac{m+n_0-1}{2} \right)}{\pi^{(m-1)/2}(m+n_0)^{1/2}} \times \left\{ v_0 s_0^2 + \frac{mn_0}{m+n_0} (\mu_0 - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m_j} (y_{ij} - \bar{y})^2 \right\}^{-(m+n_0)/2},
\end{equation}
where $m = \sum_{j=1}^{M} m_j$.

To find the form of $u_k$, the $(k+1)$st term of $C^{-1}$ (defined in (5.6)), for $1 \leq k \leq K_*$, we have that

$$u_k = C(\Lambda, K_*) \frac{\Lambda^k}{k!} \int_{\sigma^2} \int_{\mu} \int_{\alpha} p(\alpha|k \geq 1, \sigma^2) p(\mu, \sigma^2) p(\gamma|k \geq 1, \mu, \alpha, \sigma^2) \, d\alpha \, d\mu \, d\sigma^2.$$ (5.12)

Using (5.2) for $p(\mu, \sigma^2)$ the second line of (5.4) for $p(\alpha|k, \sigma^2)$, and $(1/(2\pi)^{m/2})$ times the expression to the right of the sign of proportionality in (2.2), we find (the Appendix gives details — a completion in the square for the $\alpha_j$'s and then $\mu$ is necessary) that for $1 \leq k \leq K_*$,

$$u_k = \frac{C(\Lambda, K_*)[\Lambda^k/k!] C(n_0, v_0, s_0^2) \Gamma(n_0 + m - 1)}{\pi^m} \frac{2^{n_0/2}}{A_k^{1/2}} \tilde{W}^{-(n_0 + m - 1)/2},$$ (5.13)

where

$$A_k = \left\{ \prod_{j=1}^{k} (1 + m_jC_\alpha)^{1/2} \right\} \left\{ \sum_{j=1}^{k+2} b_j \right\}$$ (5.14)

and

$$\tilde{W} = v_0s_0^2 + \tilde{S} + \sum_{j=1}^{k+2} b_j(\mu_{0j} - \bar{\mu}_0)^2,$$ (5.15)

with

$$\bar{\mu}_0 = \sum_{j=1}^{k+2} b_j\mu_{0j}/\sum_{j=1}^{k+2} b_j,$$ (5.16)

and where we have defined

$$b_j = m_j/(m_jC_\alpha + 1) \text{ for } 1 \leq j \leq k, \quad b_{k+1} = n_0, \quad b_{k+2} = \tilde{m},$$ (5.17)

$$\mu_{0j} = \tilde{y}_j - \alpha_{j0} \text{ for } 1 \leq j \leq k, \quad \mu_{0,k+1} = \mu_0, \quad \mu_{0,k+2} = \tilde{y},$$

and we note that

$$\sum_{j=1}^{k+2} b_j = (n_0 + \tilde{m}) + \sum_{j=1}^{k} m_j/(m_jC_\alpha + 1).$$ (5.18)
(The symbols \( \tilde{S} \) and \( \tilde{m} \) are defined in Section 2.) Hence we find that the posterior of \( k \), given the data from the \( M \) laboratories, is

\[
p(k | \text{data}) = u_k / \sum_{k=0}^{K*} u_k.
\]

Note that for all \( k, 0 \leq k \leq K* \), the common factors

\[C(\Lambda, K*)C(n_0, v_0, s_0^2) \Gamma \left( \frac{n_0 + m - 1}{2} \right) 2^{n_0/2} \pi^{-(m-1)/2}\]

are present in the “definition” of \( u_k \) so that there is cancellation, and we thus re-define \( p(k | \text{data}) \) as

\[
p(k | \text{data}) = \tilde{u}_k / \sum_{k=0}^{K*} \tilde{u}_k,
\]

where

\[
\tilde{u}_0 = (m + n_0)^{-1/2} \left\{ v_0 s_0^2 + \frac{mn_0}{m + n_0} (\mu_0 - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m_j} (y_{ij} - \bar{y})^2 \right\}^{-\frac{m+n_0-1}{2}},
\]

\[
\tilde{u}_k = \frac{\Lambda^k}{k! A_k^{1/2} W} \tilde{W}^{-\frac{m+n_0-1}{2}}, \quad k \geq 1.
\]

**Appendix**

The purpose of this Appendix is to provide details for the establishment of (5.13). We need to treat the case \( 1 \leq k \leq K* \) and \( k = 0 \) separately.

**Case:** \( 1 \leq k \leq K* \)

Referring to (5.12), and using (5.4) along with (5.2) for the priors of \( \alpha | k \geq 1 \), and \( p(\mu, \sigma^2) \), respectively, we find that

\[
(A.1) \quad u_k = [C(\Lambda, K*)A^k / k!] \int_{\sigma^2} d\sigma^2 \int_{\mu} d\mu \int_{\alpha} d\alpha \sim \frac{C(n_0, v_0, s_0^2)}{(\sigma^2)^{(n_0+2)/2}} \times
\]

\[
\exp\left\{ -\frac{1}{2\sigma^2} [v_0 s_0^2 + n_0 (\mu - \mu_0)^2] \times (s \pi \sigma^2 C_\alpha)^{-k/2} \exp -\frac{1}{2C_\alpha \sigma^2} \sum_{j=1}^{k} (\alpha_j - \alpha_{j0})^2 \right\}
\]

\[
\times (2\pi \sigma^2)^{-m/2} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \bar{S} + \tilde{m}(\mu - \bar{y})^2 + \sum_{j=1}^{k} m_j (\bar{y}_j - \mu - \alpha_j)^2 \right] \right\},
\]
where it is easy to see that the last line of the integrand of (A.1), namely, \( p(\bar{y} \mid k, \alpha, \mu, \sigma^2) \), can be written in this indicated form when \( k \geq 1 \) (see (2.2) and the definition of \( \bar{S} \) in (2.7)). The program for evaluating the integral of (A.1) is as follows — first complete the “squares” in \( \alpha_j \), and then integrate out with respect to the \( \alpha_j \). This yields a quadratic in \( \mu \). Complete the “square” in \( \mu \) and integrate out \( \mu \). This leaves an integration to be performed with respect to \( \sigma^2 \), which is an integration of an inverted gamma variable. This yields the result (5.13), from which the result given in (5.20) follows.

(i) Integration with respect to \( \sim \)

First collect terms involving \( \alpha_j \) in the exponential in (A.1). Apart from \([-1/2\sigma^2]\), we have for \( \alpha_j \) terms,

\[
(A.2) \quad (\alpha_j - \alpha_j^0)^2/C_\alpha + m_j[\alpha_j - (\bar{y}_j - \mu)]^2,
\]

which is easily seen to be

\[
(A.3) \quad [(1 + m_j C_j)/C_j](\alpha_j - \bar{\alpha}_j)^2 + \frac{m_j}{(1 + m_j C_\alpha)} (\alpha_j^0 - \bar{y}_j + \mu)^2
\]

\[= [(1 + m_j C_j)/C_j](\alpha_j - \bar{\alpha}_j)^2 + \frac{m_j}{(1 + m_j C_\alpha)} (\mu - (\bar{y}_j - \alpha_j^0))^2
\]

where, for the record,

\[
(A.4) \quad \bar{\alpha}_j = [\alpha_j/C_\alpha + m_j(\bar{y}_j - \mu)]/[(C_\alpha^{-1} + m_j).
\]

Integration with respect to a single \( \alpha_j \) yields

\[
(A.5) \quad \frac{\sqrt{2\pi C_\alpha \sigma^2}}{\sqrt{1 + m_j C_\alpha}} \exp \left( -\frac{1}{2\sigma^2} \left( \frac{m_j}{1 + m_j C_\alpha} \right) (\mu - (\bar{y}_j - \alpha_j^0))^2 \right),
\]

so that integration with respect to \( (\alpha_1, \ldots, \alpha_k) \) is, after using (A.5) and (A.1),

\[
(A.6) \quad u_k = \frac{[C(\Lambda, K*)A^k/k!]}{(2\pi)^{m/2} \prod_1^k [m_j C_\alpha + 1]^{1/2}}
\]

\[ \times \int_\sigma^2 d\sigma^2 \int_\mu d\mu \frac{1}{(\sigma^2)^{n_0+m+1}} \exp \left( -\frac{1}{2\sigma^2} [v_0 s_0^2 + \bar{S}] \right)
\]

\[ \times \exp \left( -\frac{1}{2\sigma^2} \left\{ n_0(\mu - \mu_0)^2 + \bar{m}(\mu - \bar{y})^2 + \sum_{j=1}^k \frac{m_j}{1 + m_j C_\alpha} [\mu - (\bar{y}_j - \alpha_j^0)]^2 \right\} \right).
\]
The integrand of (A.6) is quadratic in $\mu$ of the form

$$
(A.7) \quad \sum_{j=1}^{k+2} b_j (\mu - \mu_{0j})^2,
$$

where the $b_j$'s and $\mu_{0j}$'s are defined in (5.17). Using the definition of the mean of the $\mu_{0j}$'s given in (5.16), (A.7) can be written as

$$
(A.8) \quad B_k (\mu - \bar{\mu}_0)^2 + \sum_{j=1}^{k+2} b_j (\mu_{0j} - \bar{\mu}_0)^2,
$$

where $B_k = \sum_{j=1}^{k+2} b_j$. Thus, integration with respect to $\mu$ yields

$$
(A.9) \quad (2\pi \sigma^2 / B_k)^{1/2} \exp - \frac{1}{2\sigma^2} \sum_{j=1}^{k+2} b_j (\mu_{0j} - \bar{\mu}_0)^2
$$

so that, using (A.6) and (A.9), we obtain

$$
(A.10) \quad u_k = \frac{[C(\Lambda, K_*)\Lambda^k / k!]}{(2\pi)^{(m-1)/2} \prod_{1}^{k+2} \frac{1}{1 + m_j C_\alpha}]^{1/2} B_k^{1/2}} \times \int_{\sigma^2} d\sigma^2 \frac{1}{(\sigma^2)^{(n_0 + m - 1)/2 + 1}} \exp - \frac{1}{2\sigma^2} \left[ v_0 s_0^2 + \ddot{S} + \sum_{j=1}^{k+2} b_j (\mu_{0j} - \bar{\mu}_0)^2 \right].
$$

It is now straightforward to integrate with respect to $\sigma^2$ to yield,

$$
(A.11) \quad u_k = \frac{[C(\Lambda, K_*)\Lambda^k / k!]}{(2\pi)^{(m-1)/2} A_k^{1/2} \bar{W}^{(n_0 + m - 1)/2}} C(n_0, v_0, s_0^2) \Gamma \left( \frac{n_0 + m - 1}{2} \right),
$$

where $\bar{W}$ is defined in (5.15) and $A_k$ is defined in (5.14). This completes the proof of (5.13) for this case.

For convenience define $\bar{u}_k$ by the relation

$$
(A.12) \quad \bar{u}_k = \frac{\Lambda^k / k!}{A_k^{1/2} \bar{W}^{(n_0 + m - 1)/2}}, \quad 1 \leq k \leq K^*,
$$

so that

$$
(A.13) \quad u_k = \frac{2^{n_0/2} C(\Lambda, K_*) C(n_0, v_0, s_0^2)}{\pi^{(m-1)/2}} \Gamma \left( \frac{n_0 + m - 1}{2} \right) \bar{u}_k
$$
Case: $k = 0$.

We wish to establish (5.11). From (5.2), (5.9) and (5.10) we obtain

(A.14)

\[
    u_0 = C(\Lambda, K_*) \int_{\sigma^2} d\sigma^2 \int_{\mu} d\mu \frac{C(n_0, v_0, s_0^2)}{(\sigma^2)(n_0+2)/2} \\
    \times \exp \left\{ -\frac{1}{2\sigma^2} [v_0s_0^2 + n_0(\mu - \mu_0)^2] \right\} \times \frac{1}{(2\pi\sigma^2)^{m/2}} \exp -\frac{1}{2\sigma^2} \left\{ m(\mu - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m} (y_{ij} - \bar{y})^2 \right\}.
\]

Complete the square for the two quadratic terms in $\mu$ in the exponent of (A.15)

(A.15)

\[
    (n_0 + m)(\mu - \bar{\mu})^2 + \frac{n_0 m}{n_0 + m}(\mu_0 - \bar{y})^2,
\]

where

(A.16)

\[
    \bar{\mu} = (n_0\mu_0 + m\bar{y})/(n_0 + m).
\]

Hence, integration with respect to $\mu$ yields

(A.17)

\[
    \left( \frac{2\pi\sigma^2}{n_0 + m} \right)^{1/2} \exp -\frac{1}{2\sigma^2} n_0 m(\mu_0 - \bar{y})^2,
\]

so that, using (A.16) and (A.15) we have

(A.18)

\[
    u_0 = C(\Lambda, K_*) \int_{\sigma^2} d\sigma^2 \frac{C(n_0, v_0, s_0^2)}{(2\pi)^{(m-1)/2}(n_0 + m)^{1/2} (\sigma^2)^{(m+n_0+1)/2}} \int_{\mu} d\mu \frac{1}{(2\pi\sigma^2)^{m/2}} \exp -\frac{1}{2\sigma^2} \left\{ v_0s_0^2 + \frac{n_0 m}{n_0 + m}(\mu_0 - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m} (y_{ij} - \bar{y})^2 \right\}.
\]

Using (A.11), we have after simplification that

(A.19)

\[
    u_0 = \frac{C(\Lambda, K_*)C(n_0, v_0, s_0^2)\Gamma\left(\frac{m+n_0-1}{2}\right) 2^{n_0/2}}{\pi^{(m-1)/2}(n_0 + m)^{1/2} \left[ v_0s_0^2 + \frac{n_0 m}{n_0 + m}(\mu_0 - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m} (y_{ij} - \bar{y})^2 \right]}^{(m+n_0-1)/2},
\]

which establishes (5.11).

Define

(A.20)

\[
    \bar{u}_0 = \frac{1}{(n_0 + m)^{1/2}} \left[ v_0s_0^2 + \frac{n_0 m}{n_0 + m}(\mu_0 - \bar{y})^2 + \sum_{j=1}^{M} \sum_{i=1}^{m} (y_{ij} - \bar{y})^2 \right]^{-\frac{m+n_0-1}{2}},
\]
so that

\begin{equation}
(A.21) \quad u_0 = C(\Lambda, K_*)C(n_0, v_0, s_0^2) \Gamma \left( \frac{m + n_0 - 1}{2} \right) \frac{2^{n_0/2}}{\pi^{(m-1)/2}} \tilde{u}_0,
\end{equation}

which then yields

\[ p(k | \text{data}) = \tilde{u}_k / \sum_{j=0}^{K_*} \tilde{u}_j, \]

and completes the derivation of (5.20).

**Acknowledgements**

The authors wish to acknowledge partial support from National Science Foundation Grant DMS-87-08083 and from NSERC (Canada) Grant A8743. The authors would like to thank Robert Phillips, University of Toronto, for computing assistance.
References


Vapor Pressure (mPa) at 25 deg C

Lab 1
Lab 2
Lab 3
Lab 4
Lab 5
Lab 6
Lab 7
Lab 8

Dinitrotoluene

Figure 1

Vapor pressure readings from eight laboratories
Figure 2
Posterior distribution of $y$ for $k=2$ spurious laboratories