AN INFORMATION ANALYSIS OF
MULTINOMIAL EXPERIMENTS AND
CONTINGENCY TABLES

by

D. V. GOKHALE AND SOLOMON KULLBACK

FSU Statistics Report M343

May, 1975

The Florida State University
Department of Statistics
Tallahassee, Florida
AN INFORMATION ANALYSIS OF
MULTINOMIAL EXPERIMENTS AND
CONTINGENCY TABLES

by

D. V. Gokhale and Solomon Kullback

The University of California, Riverside and the Florida State University

ABSTRACT

The purpose of this paper is to explain and illustrate applications of the principle of minimum discrimination information estimation to the analysis of frequency or count data. The methods reviewed here are applicable when one or several multinomial experiments are to be analyzed, subject to a set of linear constraints on the underlying probabilities. The approach gives rise to log-linear models. The analysis of a general set of linear constraints is presented and illustrated with some numerical examples.
1. Introduction

The purpose of this paper is to explain and illustrate applications of the principle of minimum discrimination information (MDI) estimation to the analysis of frequency or count data. The MDI approach, as will be seen, provides a unified treatment for problems of various kinds. The analysis yields MDI estimates of cell frequencies which can be obtained routinely by convergent iterative computer algorithms. The estimates have nice properties in the sense that, in general, they are best asymptotically normal (BAN) and for a large subclass of problems they are maximum likelihood estimates.

The methods reviewed here are applicable when one or several multinomial experiments are to be analyzed, subject to a set of linear constraints on the underlying probabilities. In addition to the MDI estimates of cell frequencies, MDI statistics are provided for an analysis of information into contributions from each of a series of nested hypotheses. Thus, if \( H_1 \) and \( H_2 \) are two linear hypotheses such that \( H_2 \) implies \( H_1 \), then the MDI statistic for testing departure of the sample from the estimate under \( H_2 \) can be expressed as a sum of two components, one measuring the departure of the estimate under \( H_2 \) from the estimate under \( H_1 \) and the other measuring the departure of the sample from the estimate under the weaker hypothesis \( H_1 \). Important features of the MDI statistics are that they are distributed as chi-square in large samples and that they are additive as are also the associated degrees of freedom.

The situations in which MDI analysis can be applied include those in which the objective in "fitting" or "smoothing" of data, in which case the estimated frequency distribution is required to satisfy a certain (usually small) number of constraints which are satisfied by the observed distribution.
This problem is called an internal constraints problem (ICP). The observed
distribution can then be said to be "explainable" in terms of a smaller
number of linear functions of cell-frequencies. Analysis of contingency
tables by fitting marginals is an important particular case of ICP. The MDI
approach gives rise to loglinear models. Their treatment in contingency
tables is recently reviewed and illustrated by Ku and Kullback [1974].

The present paper illustrates MDI analysis of problems of a different
nature. In these problems, the observed distribution does not satisfy the
linear constraints of interest and the objective is to find an estimate of a
distribution that satisfies the given constraints and is as "close" as possible
to the observed distribution (with discrimination information between the two
distributions minimized). The estimated distribution then conserves all
the properties of the observed distribution except those determined by the
given constraints. Such problems are called external constraints problems
(ECP). Theoretical details of this case are given by Gokhale [1973] and
Kullback [1974]. The discussion by Simon [1973] covers a similar class of
problems.

Before formulating the problem symbolically, we give in Section 2,
several examples of ECP. In Section 3, the MDI analysis of a general set of
linear constraints is presented. In Section 4, the analysis is illustrated
with the help of numerical data from two examples of Section 2.
2. Examples: External Constraints Problem (ECP)

In this section, we give examples in which MDI analysis is applicable. For frequency data with quantitative values of the random variable we consider the following hypothesis for several samples:

(i) equality of population means when the common value is specified,
(ii) equality of population means when the common value is not specified, and
(iii) equality of population means and variances.

For data represented as a set of contingency tables, the random variable need not be necessarily quantitative (e.g. eye-color). But even in such qualitative cases, the MDI approach can be used to analyze the data under many hypotheses, including those of

(i) specified marginals
(ii) homogeneity of certain multinomials,
(iii) equality of marginal probabilities and/or individual probabilities in several contingency tables,
(iv) no interaction on a linear scale.

An important step in the analysis of all such problems is the formulation of the hypothesis in terms of linear constraints on the underlying probabilities. This is illustrated below. For the sake of brevity, the three hypotheses relating to count data on quantitative random variables are discussed with the help of one example.

In multidimensional contingency tables, with the same number of categories for each dimension, MDI analysis in ECP is applicable under hypotheses of symmetry and marginal homogeneity. For an illustration, the reader is referred to Kullback [1974]. For a different formulation of such problems and a different computer algorithm see Ireland et. al [1969] and Kullback [1971a, 1971b].
Ex. 1: In animal-trapping experiments, a certain standard type of trap can hold at most three animals. A new trap, which can hold up to five animals, is under trial. Two independent random samples of respective sizes 100 and 120 traps are laid out under homogeneous conditions. The results (artificial) are shown below:

Table 1

Animal trapping data on two trap-types

<table>
<thead>
<tr>
<th>No. of animals trapped</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of traps: Standard</td>
<td>38</td>
<td>34</td>
<td>20</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>No. of traps: New</td>
<td>37</td>
<td>32</td>
<td>31</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>120</td>
</tr>
</tbody>
</table>

Consider the following hypotheses:

\( H_1 \): The population means for both the traps are equal to 1.
\( H_2 \): The two population means are equal.
\( H_3 \): The two populations have the same means and variances.

We have two independent multinomial experiments, the first with four cells and the second with six. Let us denote the values of the random variable in the first experiment by \( x_{1i} \), \( i = 1, 2, 3, 4 \). Thus \( x_{11} = 0, x_{12} = 1, x_{13} = 2, x_{14} = 3 \). The corresponding cell-probabilities are denoted by \( p_{1i} \), \( i = 1, 2, 3, 4 \). In the second experiment with the new trap, the values of the random variable are denoted by \( x_{2i} \), \( i = 1, \ldots, 6 \) with \( p_{2i} \) the respective probabilities. Note that the random variables are quantitative.

There are two natural constraints on the probabilities: \( \Sigma p_{1i} = 1 \) and \( \Sigma p_{2i} = 1 \). The hypothesis \( H_1 \) that the population means are both equal to 1 corresponds to further linear constraints given by...
It is convenient to express the constraints in matrix notation. For this purpose, let

\[ p^* = (p_{11}, p_{12}, p_{13}, p_{14}, p_{21}, p_{22}, p_{23}, p_{24}, p_{25}, p_{26}) \]  

Then the constraints can be expressed in the form \( B_1 p = \theta_1 \), say where

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
x_{11} & x_{12} & x_{13} & x_{14} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26}
\end{bmatrix}
\]

and

\[
\theta_1^* = (1, 1, 1, 1).
\]

The first two rows of \( B_1 \) express the natural constraints and the third and fourth row correspond to (1) and (2).

In hypothesis \( H_2 \), the common value of the population mean is not specified. Hence \( H_2 \) imposes only one constraint in addition to the natural ones. It is
\[ \Sigma_1 x_{1i} p_{1i} = \Sigma_1 x_{2i} p_{2i} \text{ i.e. } \Sigma_1 x_{1i} p_{1i} - \Sigma x_{2i} p_{2i} = 0. \] (4)

The matrices \( B_2 \) and \( \theta_2 \) corresponding to \( H_2 \) can be written as

\[ B_2 = \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 & 1 & 1 \\
x_{11} & x_{12} & x_{13} & x_{14} & -x_{21} & -x_{22} & -x_{23} & -x_{24} & -x_{25} & -x_{26} \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & -1 & -2 & -3 & -4 & -5 & 0
\end{bmatrix} \]

and

\[ \theta_2 = (1, 1, 0). \]

Hypothesis \( H_3 \) postulates equality of means and variances. Given that the means are equal, equality of variances is the same as equality of second moments about the origin. Using this fact, we can write the linear constraints corresponding to \( H_3 \) as

\[ \Sigma_1 x_{1i} p_{1i} - \Sigma_1 x_{2i} p_{2i} = 0 \] (5)

\[ \Sigma_1 x_{1i}^2 p_{1i} - \Sigma_1 x_{2i}^2 p_{2i} = 0. \] (6)

Thus

\[ B_3 = \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
x_{11} & x_{12} & x_{13} & x_{14} & -x_{21} & -x_{22} & -x_{23} & -x_{24} & -x_{25} & -x_{26} \\
x_{11}^2 & x_{12}^2 & x_{13}^2 & x_{14}^2 & -x_{21}^2 & -x_{22}^2 & -x_{23}^2 & -x_{24}^2 & -x_{25}^2 & -x_{26}^2 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & -1 & -2 & -3 & -4 & -5 & 0 \\
0 & 1 & 2 & 3 & 0 & -1 & -2 & -3 & -4 & -5
\end{bmatrix} \]
and
\[ \theta_3^- = (1, 1, 0, 0). \]

Note that hypothesis \( H_1 \) could have also been stated in terms of the constraints (1) and (4). This shows that there may be several equivalent formulations of the hypothesis under study. Of course, the estimates of cell frequencies and the test statistics are unchanged under different but equivalent formulations of the hypotheses. The loglinear representation of the estimates does depend on the formulation. Note also that \( H_1 \) can be stated in terms of constraints (1), (2) and (4). But the three are not independent. Any two imply the remaining one. By convention, we specify \( H_1 \) in terms of independent constraints only.

It may be pointed out that \( H_1 \) implies \( H_2 \). Hypothesis \( H_3 \) also implies \( H_2 \). Thus \( H_2 \) is a weaker hypothesis than \( H_1 \) or \( H_3 \). This fact is interestingly reflected in their corresponding \( B \) matrices. If \( H_1 \) is formulated in terms of constraints (1) and (4), the rows of its \( B \)-matrix contain all the rows of \( B_2 \).

Similarly the rows of \( B_3 \) contain all the rows of \( B_2 \). If an hypothesis implies another (weaker) hypothesis, we can always find two \( B \)-matrices corresponding to the two hypotheses such that all the rows of the \( B \)-matrix of the weaker hypothesis are contained in the \( B \)-matrix of the other.

Remarks in the above two paragraphs apply to contingency tables also.

**Ex. 2:** In the data given by Fisher [1954 p. 244], the following 2 × 2 table represents seedling counts on self-fertilized heterozygotes for two factors in maize, starchy vs. sugary and green vs. white base leaf.

<table>
<thead>
<tr>
<th></th>
<th>Green</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starchy</td>
<td>1997</td>
<td>906</td>
</tr>
<tr>
<td>Sugary</td>
<td>904</td>
<td>32</td>
</tr>
</tbody>
</table>
In accordance with genetic theory, the marginal probabilities for starchy category and for Green category should be 0.75 each.

We now have a single multinomial experiment with four cells. The "values" of the random variable are qualitative; (Starchy, Green), (Starchy, White), (Sugary, Green) and (Sugary, White). But these values do not appear in the constraints so the approach is still applicable. The probabilities \( p_{ij} \), \( i = 1, 2, j = 1, 2 \), obey the following constraints specified by the genetic hypothesis.

\[
\begin{align*}
\ p_{11} + p_{12} &= 0.75 \\
\ p_{11} + p_{21} &= 0.75
\end{align*}
\]  

(7)  

(8)

With \( \theta = (p_{11}, p_{12}, p_{21}, p_{22}) \), the constraints (7) and (8) together with the natural constraint \( \sum \sum p_{ij} = 1 \) are put in the form \( B \theta = 0 \) by letting

\[
B = \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0
\end{bmatrix}
\]

and

\[
\theta = (1, 0.75, 0.75)
\]

The constraint \( p_{21} + p_{22} = 0.25 \), for example, is not included since it is implied by the natural constraint and (7).

The cells of the contingency table are indexed in lexicographic order as \((11, 12, 21, 22)\) above. (See also Ex. 5). This data was also considered in Ireland et.al (1968) using a different algorithm viz adjustments of marginals.
Ex. 3 (Homogeneity of binomials): The following table is based on the data of the hourly distribution of live and still births, from B. Hill [1955].

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Number of live births</th>
<th>Number of still births</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midnight</td>
<td>8691</td>
<td>283</td>
<td>8974</td>
</tr>
<tr>
<td>6 A.M.</td>
<td>8839</td>
<td>337</td>
<td>9176</td>
</tr>
<tr>
<td>12 Noon</td>
<td>6892</td>
<td>328</td>
<td>7220</td>
</tr>
<tr>
<td>6 P.M.</td>
<td>7802</td>
<td>324</td>
<td>8126</td>
</tr>
</tbody>
</table>

For the sake of illustration the data are regarded as independent samples from 4 binomial populations corresponding to the 4 time intervals of the day. The hypothesis of interest is that probabilities of live births are the same for the four intervals. Writing the probabilities in lexicographic order,

\[ p^* = (p_{11}, p_{12}, p_{21}, p_{22}, p_{31}, p_{32}, p_{41}, p_{42}) \]

the hypothesis can be expressed as

\[ p_{11} = p_{21} = p_{31} = p_{41}. \]  

(9)

The above equalities represent three independent constraints which can be taken to be

\[ p_{11} = p_{21} \]  

(10)

\[ p_{11} = p_{31} \]  

(11)

\[ p_{11} = p_{41} \]  

(12)

For this hypothesis, the formulation \[ B_p = 0 \] is achieved by setting \[ ~ \sim \sim \sim \]
\[
B = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
\end{bmatrix}
\]

and

\[
\theta^* = (1, 1, 1, 1, 0, 0, 0)
\]

**Ex. 4 (No interaction on a linear scale):**

In the data given by Bartlett (1935) are found the result of an experiment designed to investigate the propagation of plum root stocks from root cuttings. There were 240 cuttings for each of the four treatments.

<table>
<thead>
<tr>
<th>At Once ( i = 1 )</th>
<th>In Spring ( i = 2 )</th>
</tr>
</thead>
</table>
| \begin{tabular}{c}
Long \\
\( j = 1 \)
\end{tabular} | \( j = 2 \) | \begin{tabular}{c}
Long \\
\( j = 1 \)
\end{tabular} | \( j = 2 \) |
| \begin{tabular}{c}
Dead \( k = 1 \)
\end{tabular} | \begin{tabular}{c}
Short \\
\( k = 2 \)
\end{tabular} | \begin{tabular}{c}
Short \\
\( k = 2 \)
\end{tabular} | \begin{tabular}{c}
Short \\
\( k = 2 \)
\end{tabular} |
| 84 | 133 | 156 | 209 |
| 156 | 107 | 84 | 31 |
| 240 | 240 | 240 | 240 |

Thus we have four independent binomials. The null hypothesis of no interaction on a linear scale is

\[
\text{Ho: } p_{111} - p_{121} - p_{211} + p_{221} = 0,
\]

which can be expressed in the \( Bp = \theta \) formulation by setting

\[
p^* = \{p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, p_{222}\},
\]
\[ B = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0
\end{bmatrix} \]

and

\[ \theta^* = (1, 1, 1, 1, 0). \]

This data was also considered by Snedecor and Cochran (1967), Bhapkar and Koch (1968), Berkson (1972), Kullback (1974).

Ex. 5 (Several contingency tables, equality of marginal and individual probabilities): In the data given by Gail [1974, p. 97] we have two contingency tables given by

\[
\begin{array}{ccc}
20 & 5 & 3 \\
6 & 4 & 2
\end{array}
\quad \begin{array}{ccc}
15 & 15 & 2 \\
10 & 5 & 5
\end{array}
\]

Thus we have two independent multinomial experiments, each with six cells.

The hypothesis \( H_1 \) of equality of marginal probabilities is expressed in terms of probabilities \( p_{hij} \), \( h = 1, 2, i = 1, 2, j = 1, 2, 3, \) as

\[
p_{111} + p_{112} + p_{113} = p_{211} + p_{212} + p_{213},
\]

\[
p_{111} + p_{121} = p_{211} + p_{221},
\]

\[
p_{112} + p_{122} = p_{212} + p_{222},
\]

which can be expressed in the \( \mathbb{P} p = 0 \) formulation by letting

\[ \theta^* = (p_{111}, p_{112}, p_{113}, p_{121}, p_{122}, p_{123}, p_{211}, p_{212}, p_{213}, p_{221}, p_{222}, p_{223}), \]
\[
B = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\
\end{bmatrix}
\]

and

\[
\theta^* = (1, 1, 0, 0, 0).
\]

The hypothesis \( H_2 \) of equality of individual probabilities specified by the constraints

\[
P_{1ij} = P_{2ij}, \quad i = 1, 2, \quad j = 1, 2, 3,
\]

can be expressed as \( B_p = 0 \) with

\[
B = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
\end{bmatrix}
\]

and

\[
\theta^{**} = (1, 1, 0, 0, 0, 0, 0, 0)
\]

using constraints which are independent.

3. General formulation and MDI analysis

To formulate the general problem we need the following notation:

1. \( k \), the number of independent multinomial experiments \((k \geq 1)\).

2. \( d_h \), the number of cells in the \( h \)-th experiment, \( h = 1, 2, \ldots, k \).

Note that \( d_h \) need not be the same for all \( h \). Let \( d = \sum_{h=1}^{k} d_h \).
(3) \( p_{hi} \), the probability corresponding to the \( i \)-th cell, \( i = 1, 2, \ldots, d_h \) in the \( h \)-th experiment, \( h = 1, 2, \ldots, k \). For contingency tables the index \( i \) stands for the cell indexed in lexicographic order. The probabilities are assumed to be all positive. They are written as a matrix

\[
P^* = \begin{pmatrix} p_{11}, p_{12}, \ldots, p_{1d_1}, p_{21}, \ldots, p_{kd_k} \end{pmatrix}
\]

(13)

(4) \( B \) and \( \theta \), are matrices which express the constraints imposed by the hypothesis and the natural constraints \( \sum_i p_{hi} = 1, h = 1, \ldots, k \) in the form \( Bp = \theta \). If the hypothesis specifies \( m \) independent constraints, the order of \( B \) is \( (k + m) \times d \), the first \( k \) rows of \( B \) representing the natural constraints. The order of \( \theta \) is \( (k + m) \times 1 \) with the first \( k \) elements in it being unity.

(5) \( f_{hi} \), the observed frequency in the \( i \)-th cell of the \( h \)-th experiment. It is assumed that all the \( f_{hi} \) are positive. Let \( n_h = \sum_i f_{hi} \) be the number of observations in the \( h \)-th experiment. The total of all observations is denoted by \( N = \sum_h n_h \). The quantities \( w_h = n_h/N \), \( h = 1, \ldots, k \), are called weights. The observed frequency distributions are denoted by \( \pi_{ohi} = f_{hi}/n_h, i = 1, \ldots, d_h, h = 1, \ldots, k \). They can be written as a matrix similar to (13) above

\[
\pi^* = \begin{pmatrix} \pi_{011}, \pi_{012}, \ldots, \pi_{0d_1}, \pi_{021}, \ldots, \pi_{0kd_k} \end{pmatrix}
\]

(14)

Now, let \( \pi \) be an arbitrary matrix consisting of \( k \) probability distributions each on the \( d_h \) cells, \( h = 1, \ldots, k \); such that \( \pi_{hi} > 0 \) for all \( h \) and \( i \). Thus

\[
\pi^* = \begin{pmatrix} \pi_{11}, \ldots, \pi_{1d_1}, \pi_{21}, \ldots, \pi_{kd_k} \end{pmatrix}
\]

(15)

The principle of minimum discrimination information estimation and the subsequent analysis are based on the estimates \( p^*_{hi} \) of the cell probabilities
or the estimates \( f_{hi}^* = n_{hi}^* p_{hi}^* \) of the cell-frequencies obtained so as to minimize the discrimination information

\[
I(p; \pi) = \sum_{h=1}^{k} \sum_{i=1}^{d_h} p_{hi} \ln \left( \frac{p_{hi}}{n_{hi}^*} \right)
\]

subject to the condition that \( p^* \) satisfy the constraints \( Bp^* = 0 \). (See Kullback [1959], for theoretical treatment.)

We emphasize the fact that the \( \pi \) in (15) is arbitrary. In applications a suitable choice of \( \pi \) is made according to the problem at hand. In internal constraints problems the observed distributions \( \pi_0 \) satisfy the constraints, so that \( Bp = B\pi_0 \). But the objective is to find \( p \) subject to the given constraints and those implied by the given ones but no others. A proper choice of \( \pi \) is then any set of distributions which satisfies constraints already implied by some of the rows of \( B \). From the point of computer programming convenience \( \pi \) is taken to be the collection of \( k \) uniform distributions, that is, a set of distributions which satisfy the constraints of the first \( k \) rows of \( B \). In external constraints problems, the observed distributions do not satisfy the constraints \( Bp = 0 \). But, rather than "smoothing" the observed data as in the case of internal constraints, the objective now is to determine from all the distributions \( p \) which satisfy \( Bp = 0 \) the one which is closest to the observed distributions in the MDI sense. Hence in external constraints problems \( \pi \) is taken to be \( \pi_0 \).

For the sake of clarity of understanding, we now present an iterative algorithm for the one-sample case, essentially a Newton-Raphson type procedure (Kullback, 1974). Later, we shall see how the \( k \)-sample case can be transformed so that the one-sample algorithm is applicable. For a different algorithm see Gokhale [1973].

For \( k = 1 \), that is, the one-sample case, we will drop the subscript corresponding to the sample number in \( d_i, x_{1i} \) and \( f_{1i} \):
Step 1: $B_p = 0, B = \begin{bmatrix} B_1 \\ \sim \sim \sim \sim B_2 \end{bmatrix}, B_1$ is $l \times d$, $B_2$ is $m \times d$ and

$\theta = \begin{bmatrix} 1 \\ \sim \sim \sim \sim \theta^* \end{bmatrix}, \theta^*$ is $m \times 1$,

Step 2: $Bf = N\phi, \phi = \begin{bmatrix} 1 \\ \sim \sim \sim \sim \hat{\theta} \end{bmatrix}, \hat{\theta}$ is $m \times 1$, $f$ is the matrix of observed frequencies, of order $d \times 1$, $N = \sum f_i$,

Step 3: $D$ is the $d \times d$ diagonal matrix with $f_i$ in the $(ii)$-th cell,

Step 4: $S = BDB^* = \begin{bmatrix} S_{11} & S_{12} \\ \sim \sim \sim \sim S_{21} & S_{22} \end{bmatrix}, S_{11}$ is $1 \times 1$, $S_{21} = S_{12}$

is $1 \times m$, $S_{22}$ is $m \times m$,

Step 5: $S_{22.1} = S_{22} - S_{21} S_{11} S_{12}$

Step 6: $\Lambda = N\theta - N\hat{\theta} = \begin{bmatrix} 0 \\ \sim \sim \sim \sim \hat{\theta} \end{bmatrix} - \begin{bmatrix} N\theta^* \\ \sim \sim \sim \sim \sim \sim \sim \hat{\theta} \end{bmatrix} = \begin{bmatrix} \delta \end{bmatrix}$

$\delta = N\theta^* - N\hat{\theta}$ is $m \times 1$,

Step 7: $t^{(s)} = S_{22.1}^{-1}(s) \delta^{(s)}, s = 0, 1, 2...$
Let $\ln y$ denote a $d \times 1$ matrix and $\ln f$ the $d \times 1$ matrix of $\ln f_1, \ln f_2, \ldots, \ln f_d$ where $f_1, f_2, \ldots, f_d$ are the observed frequencies.

**Step 8:**\( \begin{align*} \tau(s) &= \tau(s) + t(s), \quad \tau(0) = 0, s = 0, 1, 2, \ldots. \end{align*} \)

**Step 9:**\( \begin{align*} \ln y(s) &= \ln f + \sum_{s^2} \tau(s), \quad s = 1, 2, \ldots. \end{align*} \)

**Step 10:**\( \begin{align*} y_1(s), y_2(s), \ldots, y_d(s), \quad s = 1, 2, \ldots. \end{align*} \)

**Step 11:**\( \begin{align*} L(s) &= \frac{N}{\ln y_1(s) + \ldots + y_d(s)}, \quad s = 1, 2, \ldots. \end{align*} \)

**Step 12:**\( \begin{align*} \ln f_1(s) &= L(s) + \ln y_1(s), \quad s = 1, 2, \ldots, \\
\ln f_d(s) &= L(s) + \ln y_d(s). \end{align*} \)

**Step 13:**\( \begin{align*} f_1(s), f_2(s), \ldots, f_d(s), \quad s = 1, 2, \ldots. \end{align*} \)

In step 7, $s = 0$ corresponds to the values computed in steps 1 to 6 using the observed frequencies, and $s = 1, 2, \ldots$ corresponds to the procedures in steps 1 to 6, however using the values $f_1(s), f_2(s), \ldots, f_d(s)$ from step 13. Note that in step 9 $\ln f$ is always composed of the original observations.

The iteration is continued until the maximum of the absolute values of the differences between successive iterates is less than a specified small positive number.

The final iterate $f(s)$ is the required MDI estimate $f^*$ and $2I(f^*: f)$ can be computed. It is asymptotically (for large $N$) distributed as chi-square with $m$ degrees of freedom.
Analysis of Information:

Suppose that two hypotheses $H_1$ and $H_2$ are being considered in a problem. Let $H_1$ correspond to $B_1 p \sim \theta_1$ and $H_2$ to $B_2 p \sim \theta_2$. Suppose further that $H_2$ is stronger than $H_1$, that is, the constraints $B_2 p = \theta_2$ imply $B_1 p = \theta_1$. If $m_2$ is the number of external constraints in $B_2$ and $m_1$, those in $H_1$, then $m_2 > m_1$. If $2I(f^*_1; f)$ and $2I(f^*_2; f)$ are the MDI statistics corresponding to $H_1$ and $H_2$, they are respectively asymptotically distributed as chi-square with $m_1$ and $m_2$ degrees of freedom. Further

$$2I(f^*_2; f) = 2I(f^*_2; f^*_1) + 2I(f^*_1; f) \quad (17)$$

and $2I(f^*_2; f^*_1)$, which measures the contribution to $2I(f^*_2; f)$ by the additional constraints imposed by $H_2$ but not by $H_1$, is distributed asymptotically as chi-square with $m_2 - m_1$ degrees of freedom. Such a procedure of partitioning the MDI statistic into components as in (17) above is called analysis of information.

The procedure can be extended in an obvious manner to a string of nested hypotheses, one successively implying the next.

Let us not return to $k$ samples and show how that case can be transformed into a one-sample case. Note that, for any matrix

$$p^* = (p_{11}, p_{12}, \ldots, p_{1d_1}, p_{21}, \ldots, p_{kd_k})$$

(18)

the positive probabilities $p_{hi}$ constitute a probability distribution for each $h$, $h = 1, 2, \ldots k$. Thus, for given positive constants $w_h$, $h = 1, 2, \ldots k$, with $\sum_{h=1}^k w_h = 1$, the numbers $w_h p_{hi}$ constitute a probability distribution over the $d$ cells, that is,
\[ E_{k}^{k} \sum_{i=1}^{d_1} w_{hi} p_{hi} = \sum_{h=1}^{k} w_{h} \sum_{i=1}^{d_1} p_{hi} \]

\[ = \sum_{h=1}^{k} w_{h} \]

\[ = 1. \]

The matrix of these probabilities \( w_{h} p_{hi} \), \( i = 1, 2, \ldots d_h \), \( h = 1, \ldots k \) can be written as

\[ W_p = (w_{1} p_{11}, w_{1} p_{12}, \ldots, w_{1} p_{1d_1}, w_{2} p_{21}, \ldots, w_{k} p_{kd_k})^{\top} \]

where \( W \) is a \( d \times d \) diagonal matrix, the elements in the first \( d_1 \) diagonal positions being all equal to \( v_1 \) and so on.

Now the original constraints \( B p = \theta \) can be written as

\[ (B W^{-1})(W p) = \theta \]

which now represent constraints on the probability matrix \( W_p \) over the \( d \) cells. Thus the problem is reduced to a one-sample case over \( d \) cells with probability vector \( W_p \) and constraints given by (19).

The modification of the algorithm for the \( k \)-sample case is not too complex but we do not present it here.

The analysis of information discussed earlier carries over without any change to \( k \)-samples.

Once the matrix of observed frequencies \( f \), the matrix \( B \) and the matrix \( \theta \) are constructed for any given problem, the complete analysis can be done by using convergent iterative computer algorithms suggested in Gokhale [1973] or by using their improved versions (Gokhale [1974], Kullback [1974]). These computer programs are available in the libraries of the computer centers at the George Washington University and University of California, Riverside.
Depending on the nature of \( B \), sometimes it is possible to get non-
iterative (direct) estimates of \(~\)\( \tilde{P}^* \). However, rather than examining this
aspect, we recommend to an applied statistician the routine application of
computer methods.

We also use the notation \( \ln(19) \) to express the loglinear representation of
the MDI estimate. Set
\[
P = Wp, \quad \tilde{P}^* = Wp^*, \quad II = Ww, \quad C = EW^{-1}
\]
and let \( \ln \tilde{P}^* \) be a \( d \times 1 \) matrix with entries in lexicographic order as in
\( \sim \) \( II \) \( \sim \)
(18). The loglinear representation is
\[
\ln \tilde{P}^* = C'\tau
\]
where the \( \tau \)'s are such that (19) is satisfied. We remark that the computer
program carrying out the iterative algorithm described above also provides
the values of the \( \tau \)'s as output.

4. Numerical Examples

In this section we apply the preceding methods of MDI analysis to two
examples (Ex. 2 and Ex. 5) of Section 2. The first example illustrates the
one-sample algorithm described in Section 3. The second example illustrates
analysis of information using a computer output.

For the sake of brevity, we give only the essential calculations,
leaving details to the reader. With reference to Ex. 2, of Section 2,

Step 1:
\[
B = \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0
\end{bmatrix}, \quad \theta = \begin{bmatrix}
1 \\
0.75 \\
0.75
\end{bmatrix}, \quad \theta^* = \begin{bmatrix}
0.75 \\
0.75
\end{bmatrix}, \quad \sim
\]
Step 2:

\[ f = \begin{bmatrix} 1997 \\ 906 \\ 904 \\ 32 \end{bmatrix}, \quad B_f = N_f = \begin{bmatrix} 3839 \\ 2903 \\ 2901 \end{bmatrix}, \quad \theta = \begin{bmatrix} 0.7562 \\ 0.7557 \end{bmatrix}, \quad N = 3839, \]

Step 3:

\[ D_f = \begin{bmatrix} 1997 & 0 & 0 & 0 \\ 0 & 906 & 0 & 0 \\ 0 & 0 & 904 & 0 \\ 0 & 0 & 0 & 32 \end{bmatrix}, \]

Step 4:

\[ s = \begin{bmatrix} 3839 & 2403 & 2901 \\ 2903 & 2903 & 1997 \\ 2901 & 1997 & 2901 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}, \]

Step 5:

\[ s_{22}^{-1} = \begin{bmatrix} 0.00153 & 0.00042 \\ 0.00042 & 0.00153 \end{bmatrix}, \]

Step 6:

\[ \delta = N\theta^* - N\theta = \begin{bmatrix} -23.75 \\ -21.75 \end{bmatrix}, \quad \Delta = \begin{bmatrix} 0 \\ \delta \end{bmatrix} \]
Step 7:
\[ t'(0) = g^{-1}(0) d(0) \approx 22.1 \]
\[ = \begin{bmatrix} 0.00153 & 0.00042 \\ 0.00042 & 0.00153 \end{bmatrix} \begin{bmatrix} -23.75 \\ -21.75 \end{bmatrix} \]
\[ = \begin{bmatrix} -0.0454725 \\ -0.0432525 \end{bmatrix} \]

Step 8:
\[ (\tau u)(1) \approx \begin{bmatrix} -0.0454725 \\ -0.0432525 \end{bmatrix} \]

Step 9:
\[ B \sim_{2} (\tau u)(1) = \begin{bmatrix} 0.033725 \\ 0.0454725 \\ 0.0432525 \\ 0 \end{bmatrix} \]

Step 10:
\[ y_{1}^{(1)} = 1827.448 \]
\[ y_{2}^{(1)} = 865.733 \]
\[ y_{3}^{(1)} = 865.733 \]
\[ y_{4}^{(1)} = 32.000 \]

Step 11:
\[ L^{(1)} = 0.066805 \]

Step 12:
\[ z^{(1)} = \begin{bmatrix} 1953.701 \\ 925.535 \\ 925.543 \\ 34.211 \end{bmatrix} \]
Now
\[ \delta'(1) = N \hat{\theta}^* - N \hat{\theta}'(1) = \begin{bmatrix} 0.014 \\ 0.006 \end{bmatrix}. \]

Compared to either \( N \hat{\theta}^* \) or to \( N \hat{\theta}'(1) \), \( \delta'(1) \) can be practically regarded as zero; there will be no change in further iterates up to two places of decimals. Hence we can take
\[ f^* = f(1). \]

Then \( 2I(f^* : f) \) is easily calculated as 2.022. This is not significant at 5% level, the critical value of chi-square with two degrees of freedom being 5.491.

For Gail's data (Ex. 5 of Section 2) the MDI estimates of cell frequencies under the two hypotheses \( H_1 \) (of marginal homogeneity) and \( H_2 \) (of total homogeneity) are obtained by using a computer program. They are as follows:

<table>
<thead>
<tr>
<th>Cell Index</th>
<th>( H_1 )</th>
<th>( H_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>16.504</td>
<td>15.901</td>
</tr>
<tr>
<td>112</td>
<td>6.466</td>
<td>8.559</td>
</tr>
<tr>
<td>113</td>
<td>4.042</td>
<td>2.808</td>
</tr>
<tr>
<td>121</td>
<td>6.320</td>
<td>7.419</td>
</tr>
<tr>
<td>122</td>
<td>6.604</td>
<td>4.219</td>
</tr>
<tr>
<td>123</td>
<td>2.064</td>
<td>3.095</td>
</tr>
<tr>
<td>211</td>
<td>18.263</td>
<td>19.686</td>
</tr>
<tr>
<td>212</td>
<td>12.705</td>
<td>10.596</td>
</tr>
<tr>
<td>213</td>
<td>2.476</td>
<td>3.476</td>
</tr>
<tr>
<td>221</td>
<td>9.996</td>
<td>9.186</td>
</tr>
<tr>
<td>222</td>
<td>3.477</td>
<td>5.223</td>
</tr>
<tr>
<td>223</td>
<td>5.083</td>
<td>3.832</td>
</tr>
</tbody>
</table>
The analysis of information is obtained as

<table>
<thead>
<tr>
<th>Component due to</th>
<th>Information</th>
<th>D. F.</th>
<th>chi-square(5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_2 )</td>
<td>2I(( f_2^*; f ) = 9.008</td>
<td>5</td>
<td>11.070</td>
</tr>
<tr>
<td>( H_1 )</td>
<td>2I(( f_2^<em>; f_1^</em> ) = 4.675</td>
<td>2</td>
<td>5.991</td>
</tr>
<tr>
<td></td>
<td>2I(( f_1^*; f ) = 4.333</td>
<td>3</td>
<td>7.815</td>
</tr>
</tbody>
</table>

We see that the (stronger) hypothesis of total homogeneity is accepted, hence the hypothesis of marginal homogeneity is also accepted.
References


