Biometrika (1978), 65, 3, *pp.* 603–8
Printed in Great Britain

Singularities in the analysis of binomial data

BY PASCALE ROUSSEAU AND DAVID SANKOFF
Centre de recherches mathématiques, Université de Montréal

SUMMARY

Transformed additive models accounting for the factors influencing a binomial probability sometimes give rise to infinite maximum likelihood estimates. Necessary and sufficient conditions for a given pattern of singular estimates to be produced are proved in terms of the scalability of the array of relative frequency data.

Some key words: Binomial probability; Divergent estimate; Guttman scale; Maximum likelihood; Transformed additive model.

1. INTRODUCTION

Arrays of zeros and ones in two or more dimensions may be analysed in terms of scales (Guttman, 1944), the idea being to permute the rows, columns, etc. of the array so that it becomes as monotonic as possible, according to some criterion, in each dimension. Arrays of relative frequencies are also sometimes analysed in this way, especially if most or many of the entries are zero or one. More commonly, however, frequency data are fitted by transformed additive models which decompose the binomial probabilities underlying the data into separate row, column, etc., effects (Cox, 1970; Haberman, 1974). In this paper we show that by suitably formulating the notion of a scale, and by using Haberman's extension of the maximum likelihood criterion to allow singular estimates, the scaling and parametric approaches to the analysis of frequency data are made to coincide. We work in two dimensions, but our definitions, theorem and proofs carry over in obvious ways to higher dimensions.

Let P be an $m \times n$ array of binomial probabilities where each p_{ij} independently generates r_{ij} successes out of $N_{ij} > 0$ trials. We assume that

$$p_{ij} = f(\mu + a_i + b_j), \tag{1}$$

where f is a strictly increasing function on $[-\infty, +\infty]$ whose range is the entire interval $[0, 1]$, and which assures the strict concavity of the likelihood function, e.g. the cumulative normal or the logistic transform. Unlike untransformed linear models where use of least squares assures finite parameter estimates, transformed models for fitting binomial frequencies admit divergent maximum likelihood estimates for μ, the a_i and the b_j. Various sets of necessary and sufficient conditions for finite estimates have been studied by Haberman (1974, 1977), Wedderburn (1976) and others. Our main result here is a theorem giving necessary and sufficient conditions, in terms of the scalability of the data matrix r/N, for obtaining the various possible patterns of singularities in the parameter estimates. These singularities are interpretable in a well-defined way, even when the model apparently involves the addition of two infinite quantities of opposite sign.

2. EXISTENCE AND UNIQUENESS

In the finite case, e.g. when all $0 < r_{ij} < N_{ij}$, as we will prove later for model (1) to have a unique solution it is necessary to impose constraints such as

$$\sum_{i=1}^{m} \hat{a}_i = \sum_{j=1}^{n} \hat{b}_j = 0. \tag{2}$$

Otherwise, nonuniqueness would be a consequence of the identity

$$\hat{\mu}+(\hat{a}_i+x)+(\hat{b}_j-x)=\hat{\mu}+\hat{a}_i+\hat{b}_j. \tag{3}$$

That maximum likelihood estimates can diverge may be seen already for a single binomial probability $p=f(\mu)$, where $r/N=1$ if and only if $\hat{p}=1$ and $\hat{\mu}=f^{-1}(\hat{p})=\infty$. Also, $r/N=0$ if and only if $\hat{\mu}=-\infty$. Analogous conditions, to be specified below, on the r_{ij}/N_{ij} within submatrices of the $m\times n$ data matrix, can also lead to divergent estimates for μ and/or some of the a_i and b_j. To interpret such singularities, we must extend (1), otherwise if $\hat{a}_i=\infty$ and $\hat{b}_j=-\infty$ then $\hat{p}_{ij}$ is undefined. Constraint (2) must also be reformulated to account for possibilities such as $\hat{a}_i=\infty$, $\hat{a}_j=-\infty$.

For a given data set, let L be the least upper bound of the likelihood function over all sets of finite parameter configurations satisfying (2). By the continuity of f and the likelihood function, we can find a sequence of such sets of finite estimates $\hat{\mu}^{(k)}$, $\hat{a}_i^{(k)}$, $\hat{b}_j^{(k)}$ with likelihoods L_k satisfying $\lim L_k=L$, and such that the values for each parameter also approach a limit, possibly infinite. The strict concavity of the likelihood function under (2) ensures the uniqueness of these limits, which we write as $\hat{\mu}$, $\hat{a}_i$ and $\hat{b}_j$.

Defining

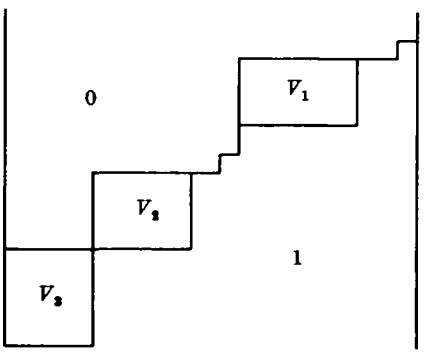
$$\hat{p}_{ij}=\lim_{k\to\infty} f(\hat{\mu}^{(k)}+\hat{a}_i^{(k)}+\hat{b}_j^{(k)}), \tag{4}$$

we note that this estimate will have likelihood L and is a consistent extension of (1). Though (2) no longer has meaning if some of the parameter estimates are infinite, the fact that it holds for each set $\{a_1^{(k)},\ldots,a_m^{(k)}\}$ and each $\{b_1^{(k)},\ldots,b_n^{(k)}\}$ suffices to ensure the uniqueness of any finite estimates among $\hat{\mu}$, the $\hat{a}_i$ and the $\hat{b}_j$. This extension of the maximum likelihood criterion is discussed by Haberman (1974, Appendix B).

3. CANONICAL FORMS

A matrix of probabilities in scaled form consists of a partition of the rows into subsets each containing one or more consecutive rows, and a similar partition of the columns; in the submatrix or block where a subset of rows intersects a subset of columns, the entries are all zeros or are all ones, except that each subset of rows or columns may have all entries strictly between zero and one everywhere within at most one block, called its variable block. In addition, the entries in each row and column are respectively nondecreasing from left to right and from top to bottom. Thus a scaled form has two roughly triangular regions, one in the upper left-hand corner which contains only zeros, and the other in the lower right-hand containing ones. These regions may touch at one or more places, sandwiching between them one or more rectangular blocks each containing only terms not zero or one, as illustrated in Table 1.

Table 1. *Matrix in scaled form. Terms in blocks V_1, V_2 and V_3 strictly between zero and one*



A blockwise-scaled matrix also involves row and column partitions and also permits at most one variable block per subset of rows or subset of columns. Within the variable block, however, zeros and ones are permitted and the nondecreasing condition need not apply there.

THEOREM. *If an array $\hat{P}$, together with possibly divergent estimates $\hat{\mu}, \hat{a}_i, \hat{b}_j$, is a maximum likelihood analysis for binomial data with relative frequency array F, then there is a unique relabelling of rows and columns which puts $\hat{P}$ into scaled form. This also puts F into blockwise-scaled form with corresponding variable blocks. For any other blockwise-scaled arrangement of F, its variable blocks cover those of $\hat{P}$.*

Proof. We show first that the possibly infinite value

$$\lim_{k\to\infty}(\hat{a}_i^{(k)} - \hat{a}_j^{(k)}) \tag{5}$$

defines a total order among the $\hat{a}_i$. A similar order exists among the $\hat{b}_j$. This is trivial if all $\hat{a}_i$ are finite, and even if no more than one $\hat{a}_i = \infty$ and no more than one $\hat{a}_j = -\infty$. However, if $\hat{a}_1 = \ldots = \hat{a}_s = -\infty$, where $s > 1$ for example, there is the question of whether a unique finite or infinite limit exists when $1 \leqslant i < j \leqslant s$ in (5). To answer this, we set up a supplementary estimation problem involving only the first s rows of F.

We estimate new row effects $\hat{c}_1, \ldots, \hat{c}_s$ and then compare these to $\hat{a}_1, \ldots, \hat{a}_s$. More precisely, we find s sequences $\hat{c}_1^{(k)}, \ldots, \hat{c}_s^{(k)}$ containing only finite terms and satisfying the constraint, for each k

$$\sum_{i=1}^{s} \hat{c}_i^{(k)} = 0, \tag{6}$$

where each sequence has a unique finite or infinite limit and such that the set of $\hat{q}_{ij}$ defined by

$$\lim_{k\to\infty} f(\nu^{(k)} + \hat{c}_i^{(k)} + d_j^{(k)}) \tag{7}$$

has maximum likelihood. So that the $\hat{c}_i^{(k)}$ be comparable to the $\hat{a}_i^{(k)}$, a special constraint is imposed: the sequences $\nu^{(k)} = \hat{\mu}^{(k)} + s^{-1}\Sigma\hat{a}_i^{(k)}$ as well as $d_1^{(k)} = \hat{b}_1^{(k)}, \ldots, d_n^{(k)} = \hat{b}_n^{(k)}$ are fixed in advance according to the solution of the original m-row problem. How can we be sure that a unique maximum likelihood solution exists to this estimation problem under the special constraint? For each fixed k, by the continuity of f and the likelihood function, we can find sequences $\hat{c}_1^{(k,l)}, \ldots, \hat{c}_s^{(k,l)}$ for $l = 1, 2, \ldots$, such that as $l \to \infty$, $\lim f(\nu^{(k)} + \hat{c}_i^{(k,l)} + d_j^{(k)})$ represents a most likely solution, of likelihood V_k, and where, by the concavity of the likelihood function, the sequences approach unique limits. We then choose an l large enough so that $V_k^{(l)}$, the likelihood of the probabilities $f(\nu^{(k)} + \hat{c}_i^{(k,l)} + d_j^{(k)})$ differs from V_k by less than ε/k, and set each $\hat{c}_i^{(k)} = \hat{c}_i^{(k,l)}$. Since for each j, the sequence $d_j^{(k)}$ approaches a unique limit, continuity and concavity considerations again ensure that the $\hat{q}_{ij}$ defined by the sequences $\nu^{(k)}; \hat{c}_1^{(k)}, \ldots, \hat{c}_s^{(k)}; d_1^{(k)}, \ldots, d_n^{(k)}$ have maximum likelihood $V = \lim V_k$, and that the sequences approach unique limits.

Now that we have shown that the special s-row problem admits a unique maximum likelihood solution, we compare this to that of the full problem. For each k, suppose that the likelihood of the parameter set $\hat{\mu}^{(k)}; \hat{a}_1^{(k)}, \ldots, \hat{a}_s^{(k)}; \hat{b}_1^{(k)}, \ldots, \hat{b}_n^{(k)}$ over the first s rows is W_k, and $\lim W_k = W$. By setting

$$c_{(i)}^{*(k)} = \hat{a}_i^{(k)} - \frac{1}{s}\sum_{i=1}^{s} \hat{a}_i^{(k)} \tag{8}$$

we obtain a set of estimates for the s row problem satisfying (6). Since

$$\nu^{(k)} + c_i^{*(k)} + d_j^{(k)} = \hat{\mu}^{(k)} + \hat{a}_i^{(k)} + \hat{b}_j^{(k)} \tag{9}$$

the likelihood of this new parameter set is also $W_k \leqslant V_k$. Going to the limit, we have that $W \leqslant V$. Similarly setting

$$a_i^{*(k)} = \hat{a}_i^{(k)} + \nu^{(k)} - \hat{\mu}^{(k)} \tag{10}$$

we have that $V_k \leqslant W + \varepsilon/k$, and so $V \leqslant W$. Thus $W = V$ and the uniqueness of the maximum likelihood solution ensures that the sequences $c_1^{*(k)}, \ldots, c_s^{*(k)}$ approach this solution in the s row problem.

From (6), if some $c_i^* = -\infty$, there must be at least one $c_j^* = \infty$ and vice versa, so that

$$\lim_{k \to \infty} (c_i^{*(k)} - c_j^{*(k)}) \tag{11}$$

is defined and equals $-\infty$. If all c_i^* are finite, then (11) is also well defined for any $1 \leqslant i, j \leqslant s$. From (8), as $k \to \infty$,

$$\begin{aligned} \lim (\hat{a}_i^{(k)} - \hat{a}_j^{(k)}) &= \lim (c_i^{*(k)} + \nu^{(k)} - \hat{\mu}^{(k)} - c_j^{*(k)} - \nu^{(k)} + \hat{\mu}^{(k)}) \\ &= \lim (c_j^{*(k)} - c_i^{*(k)}) \end{aligned} \tag{12}$$

and we have thus established the existence of additional order relations among the $\hat{a}_i$.

We can repeat the above procedure if there remain $t < s$ negatively infinite $\hat{a}_i$ among which (5) remains undefined, and so on. Similarly, we may order the positively infinite $\hat{a}_i$, and then proceed to order the $\hat{b}_j$.

Relabelling the rows and columns according to the order we have established puts $\hat{P}$ into scaled form, with the subsets of the row and column partitions defined by the subsets of the $\hat{a}_i$ and the $\hat{b}_j$, within which all limits of form (5) are finite. For if $\lim f(\hat{\mu}^{(k)} + \hat{a}_i^{(k)} + \hat{b}_j^{(k)})$ is strictly between zero and one, then $\lim (\hat{\mu}^{(k)} + \hat{a}_i^{(k)} + \hat{b}_j^{(k)})$ is finite and so are all those involving $\hat{a}_h^{(k)}$ and $\hat{b}_l^{(k)}$ within the same subsets as $\hat{a}_i^{(k)}$ and $\hat{b}_j^{(k)}$ respectively. In any block to the left of such a variable block, the entries will be zero since $\lim (\hat{b}_j^{(k)} - \hat{b}_g^{(k)}) = \infty$ for column g in a subset which precedes that containing column j. The same for any block above the variable block, while for those below and/or to the right the entries will be identically one. Elsewhere, since $\hat{\mu}^{(k)} + \hat{a}_i^{(k)} + \hat{b}_j^{(k)}$ can only vary by a finite amount within a block, the entries will be identically one or identically zero. The fact that the $\hat{a}_i$ are ordered, as are the $\hat{b}_j$, assures the nondecreasing condition on the entries in any column and those in any row.

Having shown that $\hat{P}$ is in scaled form, we consider a set of $N_{ij} > 0$ binomial trials where $\hat{p}_{ij} = 0$. Since $\hat{p}_{ij}^r (1 - \hat{p}_{ij})^{N-r} = 0$ if $r > 0$, then $\hat{p}_{ij} = 0$ cannot form part of a maximum likelihood solution if $r_{ij} > 0$. Similarly if $\hat{p}_{ij} = 1$, then $r_{ij} = N_{ij}$. Thus F is in blockwise-scaled form with the same variable blocks as $\hat{P}$.

It remains to prove the last statement of the theorem. Given a relative frequency matrix F in blockwise-scaled form, it suffices to prove that outside a variable block, if $r_{ij} = 0$, then $\hat{p}_{ij} = 0$, and if $r_{ij} = N_{ij}$, then $\hat{p}_{ij} = 1$, in the maximum likelihood solution.

If $r_{gh} = 0$ outside a variable block, it is in a block containing only zero relative frequencies. Suppose that the right-most column in this block is the tth column. All columns in this block contain only zeros from row 1 down to some row s. The case $s = m$ will be dealt with separately. Generally, row $s+1$ is the first row in a variable block or possibly a block of ones. All $r_{ij} = 0$ for $i \leqslant s, j \leqslant t$ while all $r_{ij} = N_{ij}$ for $i \geqslant s+1, j \geqslant t+1$ as can be easily deduced from the blockwise scaling condition. We have thus divided our matrix into four submatrices. The upper left $s \times t$ matrix Z contains only zeros, the lower right $m-s \times n-t$ matrix J contains only ones. The lower left is any $(m-s) \times t$ matrix M_1 and the upper right is any

$(n-t)\times s$ matrix M_2. All we need show is that the matrix $\hat{P}$ contains the same submatrices Z and J in the same positions as F.

Consider $\hat{P}^{(k)}$, an estimate matrix approximating $\hat{P}$ based on finite parameter estimates satisfying (2). We adjust the parameter estimates to improve the likelihood. We add x_1 to each of $\hat{a}_{s+1}^{(k)},\ldots,\hat{a}_m^{(k)}$ to form $\tilde{a}_{s+1}^{(k)},\ldots,\tilde{a}_m^{(k)}$ and add x_2 to $\hat{a}_1^{(k)},\ldots,\hat{a}_s^{(k)}$ to form $\tilde{a}_1^{(k)},\ldots,\tilde{a}_s^{(k)}$. Similarly we add y_1 to $\hat{b}_1^{(k)},\ldots,\hat{b}_t^{(k)}$ and y_2 to $\hat{b}_{t+1}^{(k)},\ldots,\hat{b}_n^{(k)}$ to form the $\tilde{b}_j^{(k)}$. We add z to $\hat{\mu}^{(k)}$ to give $\tilde{\mu}^{(k)}$. For any x_1 we can always choose the other increments so that constraint (2) remains satisfied and so that the $\hat{p}_{ij}$ in the submatrices corresponding to M_1 and M_2 remain unchanged. In the process of carrying out this adjustment, we find that each entry in the upper left corner decreases by $x_1 m/s$ and each one in the lower right increases by the same amount. Thus by making x_1 large enough we can assure that each $\tilde{p}_{ij}^{(k)}<\varepsilon/k$ for $1\leqslant i\leqslant s$, $1\leqslant j\leqslant t$ and $\tilde{p}_{ij}^{(k)}>1-\varepsilon/k$ for $s+1\leqslant i\leqslant m$, $t+1\leqslant j\leqslant n$. Then $\tilde{P}^{(k)}$ will have likelihood $\tilde{L}_k\geqslant L_k$, since $\tilde{p}_{ij}^{(k)}=\hat{p}_{ij}^{(k)}$ for M_1 and M_2, but the $\tilde{p}_{ij}^{(k)}$ are more likely outside M_1 and M_2 where $r_{ij}=0$ or $r_{ij}=N_{ij}$.

Since the $\hat{P}^{(k)}\to\hat{P}$, so do the $\tilde{P}^{(k)}$, and hence $\hat{p}_{gh}=0$. If $s=m$, the proof follows an adjustment involving y_1, y_2 and z only. Analogous reasoning proves $r_{gh}=N_{gh}$ outside a variable block of F only if $\hat{p}_{gh}=1$.

COROLLARY 1. *A data set gives estimates all finite if and only if for any row and column permutations, the upper left $s\times t$ submatrix does not contain only zeros at the same time as the lower $(m-s)\times(n-t)$ submatrix contains only ones, $0\leqslant s\leqslant m$, $0\leqslant t\leqslant n$ (Haberman, 1977, p. 821).*

COROLLARY 2. *A data set where $0<r_{ij}<N_{ij}$, for $1\leqslant i\leqslant m$, $1\leqslant j\leqslant n$, admits only finite parameter estimates.*

COROLLARY 3. *For a relative frequency matrix in blockwise-scaled form, finite row effects are estimated for at most the rows comprising one block. Likewise for column effects.*

COROLLARY 4. *If the rows and columns of a matrix of zeros and ones can be permuted so as to produce a scale with no variable blocks, this permutation can be obtained through the maximum likelihood estimation of row and column effects.*

The matrix of zeros and ones can be considered a relative frequency matrix where all $N_{ij}=1$ and $r_{ij}=0$ or $r_{ij}=1$. Then there is a unique scaled form associated with the maximum likelihood analysis.

Of course, if such a scale, known as a Guttman scale, exists, it can be found directly by ordering row and column sums. Nevertheless Corollary 4 provides an interesting connexion between combinatorial data analysis, such as Guttman scaling, and analyses based on parametric models such as (1). Note that according to Corollary 3, a square matrix containing only zeros above the diagonal and only ones below, with all diagonal terms equal to $\frac{1}{2}$, will have almost all infinite parameters. By setting the $m-1$ terms on the subdiagonal also equal to $\frac{1}{2}$, however, we are assured by Corollary 1 that no estimates are infinite.

The combinatorial structure inherent in the estimation problem for blockwise scalable data sets can also be seen in the following algorithm for determining the first and last few rows and columns in the scale without actually carrying out the maximization of the likelihood.

We first permute the rows so that any row which contains only zeros is at the top of the matrix and any row of ones is at the bottom. We then consider the reduced matrix consisting of the remaining rows. We permute any column containing only zeros to the left and those containing only ones to the right. We reduce the matrix further by eliminating these columns

and repeat the procedure until we arrive at a submatrix S which has no row or column consisting entirely of zeros or of ones.

The rows and columns of zeros or ones found during this procedure correspond to deterministic effects, so-called because the effect of any such row, say, found at step I of the algorithm, completely outweighs all column effects which have not been found to be deterministic at a previous step. In higher dimensions, a deterministic effect outweighs any combination of effects which it outranks in the hierarchy produced by the algorithm. Each set of deterministic rows found by the algorithm corresponds to a subset in the partition of the $\hat{a}_i$ defining the scaled form of $\hat{P}$. Note that step 1 may find no rows, and the algorithm may exhaust the matrix so that there is no submatrix S. At every other step, however, the algorithm must find at least one deterministic row or column, or else it stops.

When S is scaled by the maximum likelihood analysis, it will necessarily have variable blocks in the lower left and upper right-hand corners.

In solving the maximum likelihood problem by successive approximation, the hierarchy of deterministic effects becomes apparent in the different rates with which $\hat{\mu}^{(k)}$, the $\hat{a}_i^{(k)}$ and the $\hat{b}_j^{(k)}$ approach infinity with k.

COROLLARY 5. *Let $\hat{\mu}^{(k)}$, $\hat{a}_i^{(k)}$, $\hat{b}_j^{(k)}$ approach a maximum likelihood solution. If row i is found to be deterministic at step* 1 *of the algorithm, then*

$$\hat{b}_j^{(k)} = o(\hat{\mu}^{(k)} + \hat{a}_i^{(k)}) \tag{13}$$

for $1 \leqslant j \leqslant n$. A deterministic effect found at step I satisfies a set of relations of type (13) *only with those effects not previously found to be deterministic. For a row i passing through the submatrix S, there exists a column j passing through S which does not satisfy* (13), *and vice versa.*

REFERENCES

COX, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

GUTTMAN, L. (1944). A basis for scaling qualitative data. *Am. Soc. Rev.* **9**, 139–50.

HABERMAN, S. (1974). *The Analysis of Frequency Data*. University of Chicago Press.

HABERMAN, S. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815–41.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.

[*Received March* 1978. *Revised May* 1978]