7

A Stabilized Newton-Raphson Algorithm for Log-Linear Models for Frequency Tables Derived by Indirect Observation

*Shelby J. Haberman**

In a variety of problems involving models from genetics, latent-class analysis, and missing data, I apply a log-linear model to an indirectly observed frequency table. Current algorithms for computation of maximum likelihood estimates for such cases have often been unsatisfactory because they fail to converge at all or they converge at an unacceptable rate. I propose a new algorithm that converges both more quickly and more reliably than currently available alternatives. The algorithm assists in estimation of asymptotic variances of parameter estimates. It may be applied to both grouped and ungrouped data. I illustrate results in two examples from the literature on latent-class analysis.

1. INTRODUCTION

Log-linear models with indirectly observed frequency tables have been used by Haberman (1974*a*, 1976, 1979) in problems in genetics and in latent-class analysis. Unlike log-linear models for

Research for this paper was partially supported by National Science Foundation grant DMS 8607373 and U.S.-Israel Binational Fund grant 85-00014. This research has greatly benefitted from conversations with Christopher Winship and from extensive testing of the algorithm by James Coverdill. The computer programs used in this paper are available from the author on 5.25-inch floppy disks.

*Northwestern University

directly observed tables, log-linear models for indirectly observed tables do not require the likelihood function to be log-concave. Therefore, computation of maximum likelihood estimates is somewhat more difficult. In this paper, I propose a new algorithm based on the modified Newton-Raphson algorithm in Haberman (1974*b*, pp. 47–48). This algorithm preserves the stable convergence properties of the iterative scaling algorithm presented in Haberman (1976) and the rapid local convergence properties of the scoring algorithm in Haberman (1979, chap. 10). As in the scoring algorithm, an estimated asymptotic covariance matrix of the parameter estimates is produced as a by-product of computations. This matrix is not provided by the EM algorithm described by Dempster, Laird, and Rubin (1977), which is used for latent-class models by Clogg (1977), Goodman (1974*a*, 1974*b*), and Haberman (1977*a*). I describe the model under study and the conventional Newton-Raphson algorithm in section 2. I also describe conditions under which the estimated asymptotic covariance matrix produced via the proposed algorithm or via the Newton-Raphson algorithm is useful in computation of approximate confidence intervals for parameters. In section 3, I describe the proposed algorithm and demonstrate its properties of stability and rapid convergence. In section 4, I present some examples of its use.

2. THE LOG-LINEAR MODEL

In this section I present a general log-linear model for problems of indirect observation. Unlike the presentation in Haberman (1974*a*), this presentation emphasizes individual observations rather than tables; however, the model presented does not differ from the models in Haberman (1974*a*). In section 2.1, I provide the first, second, and third partial derivatives of the logarithm L of the likelihood function. In section 2.2, I summarize the relationship of the gradient vector ∇L and the Hessian matrix $\nabla^2 L$ of L to maximum likelihood estimates. In section 2.3, I show that ∇L and $\nabla^2 L$ may be used to construct the Newton-Raphson algorithm for computation of maximum likelihood estimates and that ∇L, $\nabla^2 L$, and the array $\nabla^3 L$ of third partial derivatives of L may be used to describe convergence properties. In section 2.4, I show that the Hessian $\nabla^2 L$ may be used under very general conditions to provide estimated asymptotic covariances of parameter estimates. In the problems considered in this paper, the

Hessian is easier to use than the Fisher information matrix, because the Hessian can be computed given the sets Z_i without reference to the classes G_i of possible Z_i. In section 2.6, I discuss the stability problem associated with the Newton-Raphson algorithm and justify the stabilized version of the algorithm, which is presented in section 3.

In the problems considered in this paper, a log-linear model is applied to N polytomous random variables Y_i, $1 \le i \le N$, but for each i, it is known only that Y_i has some value in a set Z_i. To increase the generality of results, the set J_i of possible values of Y_i may depend on i. In like manner, Z_i is a member of a collection G_i of disjoint nonempty subsets of J_i, where G_i may depend on i. For example, consider a household survey in which employment status is sought for every adult member of household i. Since the number of adults in the household varies, the set J_i of possible combinations of employment status varies from household to household. If data are complete for household i, then $Z_i = \{Y_i\}$. If the employment status for all household members is not known, then Z_i consists of all possible combinations of employment status consistent with the information provided.

To describe the distribution of the Y_i in terms of log-linear models, let $p_{ij} > 0$ be the probability that $Y_i = j$ for j in J_i. Corresponding to each subject i and each response j in J_i are a scale factor $z_{ij} > 0$ and a vector $\mathbf{x}_{ij}$ of associated fixed variables x_{ijk}, $1 \le k \le r$. Consider the log-linear model which assumes that for some unknown vector β with coordinates β_k, $1 \le k \le r$, and some unknown scalars α_i, $1 \le i \le N$,

$$\log(p_{ij}/z_{ij}) = \alpha_i + \sum_k \beta_k x_{ijk}. \tag{1}$$

To illustrate this class of models, I present two examples from the literature on latent-class analysis.

Example 1. Haberman (1979, p. 589) considers a latent-class model in which an observation Y_i consists of a vector $(U_i, A_i, B_i, C_i, D_i)$, where U_i, A_i, B_i, and C_i are 1 or 2 and D_i is 1, 2, or 3. The variable U_i is a latent variable reflecting the attitude of subject i toward legalized nontherapeutic abortion; A_i, B_i, and C_i are manifest variables that are responses of subject i to three questions concerning conditions under which such abortions should be legal; and D_i is a fixed variable that specifies the year in which subject i was interviewed. Thus, J_i consists

of all vectors (u, a, b, c, d) such that u, a, b, and c are 1 or 2 and d is 1, 2, or 3. The set $Z_i = \{(u, A_i, B_i, C_i, D_i): 1 \le z \le 2\}$. It is assumed that given U_i, the variables A_i, B_i, C_i, and D_i are conditionally independent. We can thus write

$$\log p_{iuabcd} = \alpha_d + \lambda_u^U + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ua}^{UA} + \lambda_{ub}^{UB} + \lambda_{uc}^{UC} + \lambda_{ud}^{UD},$$

where

$$\sum \lambda_u^U = \cdots = \sum_u \lambda_{ud}^{UD} = \sum_d \lambda_{ud}^{UD} = 0.$$

In (1), we can let $p = 9$, $z_{iuabcd} = 1$, $\beta_1 = \lambda_1^U$, $\beta_2 = \lambda_1^A$, $\beta_3 = \lambda_1^B$, $\beta_4 = \lambda_1^C$, $\beta_5 = \lambda_{11}^{UA}$, $\beta_6 = \lambda_{11}^{UB}$, $\beta_7 = \lambda_{11}^{UC}$, $\beta_8 = \lambda_{11}^{UD}$, $\beta_9 = \lambda_{12}^{UD}$, and let

$$\begin{aligned}
x_{iuabcd1} &= 1, && u = 1,\\
&= -1, && u = 2,\\
x_{iuabcd2} &= 1, && a = 1,\\
&= -1, && a = 2,\\
x_{iuabcd3} &= 1, && b = 1,\\
&= -1, && b = 2,\\
x_{iuabcd4} &= 1, && c = 1,\\
&= -1, && c = 2,\\
x_{iuabcd5} &= 1, && u = a,\\
&= -1, && u \ne a,\\
x_{iuabcd6} &= 1, && u = b,\\
&= -1, && u \ne b,\\
x_{iuabcd7} &= 1, && u = c,\\
&= -1, && u \ne c,\\
x_{iuabcd8} &= 1, && u = d = 1 \text{ or } u = 2 \text{ and } d = 3,\\
&= -1, && u = 2 \text{ and } d = 1 \text{ or } u = 1 \text{ and } d = 3,\\
&= 0, && d = 2,\\
x_{iuabcd9} &= 1, && u = 1 \text{ and } d = 2 \text{ or } u = 2 \text{ and } d = 3,\\
&= -1, && u = 2 \text{ and } d = 2 \text{ or } u = 1 \text{ and } d = 3,\\
&= 0, && d = 1.
\end{aligned}$$

If n_{abcd} is the number of subjects i with $A_i = a$, $B_i = b$, $C_i = c$, and $D_i = d$, then the values listed in Table 1 are obtained.

Example 2. Goodman (1974*a*) uses a two-variable latent-class model to analyze Table 2. This table, which is derived from Coleman (1964), cross-classifies self-perceived membership in and attitude toward the

TABLE 1
Responses to Three Questions on Abortion

Response to a[a]	Response to b[a]	Response to c[a]	Year[b]	Count
1	1	1	1	334
1	1	2	1	34
1	2	1	1	12
1	2	2	1	15
2	1	1	1	53
2	1	2	1	63
2	2	1	1	43
2	2	2	1	501
1	1	1	2	428
1	1	2	2	29
1	2	1	2	13
1	2	2	2	17
2	1	1	2	42
2	1	2	2	53
2	2	1	2	31
2	2	2	2	453
1	1	1	3	413
1	1	2	3	29
1	2	1	3	16
1	2	2	3	18
2	1	1	3	60
2	1	2	3	57
2	2	1	3	37
2	2	2	3	430

Source: Haberman 1979, pp. 399, 482.

Note: Respondents were asked whether or not they thought it should be possible for a pregnant woman to obtain a legal abortion (a) if she is married and does not want any more children, (b) if the family has a very low income and cannot afford any more children, (c) if she is not married and does not want to marry the father.

[a] 1 = yes, 2 = no.

[b] 1 = 1972, 2 = 1973, 3 = 1974.

"leading crowd" at two times. The variables under study for subject i are the six 0-1 variables: latent membership U_i, latent attitude V_i, membership A_i at time 1, attitude B_i at time 1, membership C_i at time 2, and attitude D_i at time 2. Thus, $Y_i = (U_i, V_i, A_i, B_i, C_i, D_i)$, and J_i consists of all vectors (u, v, a, b, c, d) with all coordinates equal to 0 or 1. The latent variables U_i and V_i are not observed, so Z_i contains the

TABLE 2
Subjects Classified by Identification with Leading Crowd

First Interview		Second Interview		
Self-Perceived Membership in Leading Crowd[a] (A)	Attitude Toward Leading Crowd[b] (B)	Self-Perceived Membership in Leading Crowd[a] (C)	Attitude Toward Leading Crowd[b] (D)	Count
1	1	1	1	458
1	1	1	0	140
1	1	0	1	110
1	1	0	0	49
1	0	1	1	171
1	0	1	0	182
1	0	0	1	56
1	0	0	0	87
0	1	1	1	184
0	1	1	0	75
0	1	0	1	531
0	1	0	0	281
0	0	1	1	85
0	0	1	0	97
0	0	0	1	338
0	0	0	0	554

Source: Goodman 1974*a*, p. 1183.
[a] 1 = membership.
[b] 1 = membership does not require going against one's principles sometimes.

four elements $(u, v, A_i, B_i, C_i, D_i)$, $u = 0$ or 1, $v = 0$ or 1, and G_i consists of the sets $\{(u, v, a, b, c, d)\colon 1 \le u \le 2,\ 1 \le v \le 2\}$, where a, b, c, and d may be 0 or 1. Goodman (1974*a*) considers a model with the standard local independence assumption that given U_i and V_i, the four manifest variables are conditionally independent. The model also assumes that given V_i, U_i and (B_i, D_i) are conditionally independent, and that given U_i, V_i and (A_i, C_i) are conditionally independent. As in Haberman (1979, p. 558), the probability $p_{iuvabcd}$ that $U_i = u$, $V_i = v$, $A_i = a$, $B_i = b$, $C_i = c$, and $D_i = d$ can be written

$$\log p_{iuvabcd} = \lambda + \lambda_U q_u + \lambda_V q_v + \lambda_A q_a + \lambda_B q_b + \lambda_C q_c + \lambda_D q_d + \lambda_{UV} q_u q_v + \lambda_{UA} q_u q_a + \lambda_{UC} q_u q_c + \lambda_{VB} q_v q_b + \lambda_{VD} q_v q_d,$$

so that (1) holds with $\beta_1 = \lambda_U$, $\beta_2 = \lambda_V$, etc.

2.1. *The Likelihood Function*

To use an algorithm based on the Newton-Raphson algorithm, we must consider the log-likelihood function and its derivatives. As in Haberman (1974*a*), if γ is the vector with coordinates γ_k, $1 \le k \le r$, then the logarithm $L(\gamma)$ of the likelihood function is

$$L(\gamma) = A(\gamma) - B(\gamma),$$

where

$$A(\gamma) = \sum \log Q_i(\gamma),$$

$$B(\gamma) = \sum \log R_i(\gamma),$$

$Q_i(\gamma)$ is the sum of $q_{ij}(\gamma) = \exp(\Sigma_k \gamma_k x_{ijk})$ over j in Z_i, and $R_i(\gamma)$ is the sum of $q_{ij}(\gamma)$ over j in J_i.

The Newton-Raphson algorithm itself requires the gradient ∇L and Hessian $\nabla^2 L$ of L. Analysis of convergence properties and of large-sample normal approximations requires $\nabla^3 L$, the array of third partial derivatives of L. Let $L_k(\gamma)$ denote the partial derivative of $L(\gamma)$ with respect to γ_k, evaluated at γ, $1 \le k \le r$. Then, the gradient $\nabla L(\gamma)$ is the vector with coordinates $L_k(\gamma)$. Similarly, let $L_{kl}(\gamma)$ be the second partial derivative of $L(\gamma)$ with respect to γ_k and γ_l, evaluated at γ. Then, the Hessian matrix $\nabla^2 L(\gamma)$ is the symmetric $r \times r$ matrix with elements $L_{kl}(\gamma)$, $1 \le k \le r$, $1 \le 1 \le r$. In like manner, $\nabla^3 L(\gamma)$ is the $r \times r \times r$ symmetric array of third partial derivatives $L_{klm}(\gamma)$ of $L(\gamma)$ with respect to γ_k, γ_l, and γ_m, $1 \le k \le r$, $1 \le 1 \le r$, $1 \le m \le r$. These derivatives are readily determined given the arguments in Haberman (1974*a*). To simplify notation, I use the term $\mathbf{a}^2$ to denote the $r \times r$ matrix with elements $a_k a_l$, $1 \le k \le r$, $1 \le 1 \le r$, and the term $\mathbf{a}^3$ to denote the $r \times r \times r$ array with elements $a_k a_l a_m$, $1 \le k \le r$, $1 \le 1 \le r$, $1 \le m \le r$.

The gradient $\nabla L(\gamma)$ of L is expressible as the difference for $\beta = \gamma$ between conditional expected values of $\Sigma \mathbf{X}_i$ given Y_i is in Z_i for $1 \le i \le N$ and the unconditional expected value of $\Sigma \mathbf{X}_i$, where $\mathbf{X}_i$ is the vector with coordinates x_{ijk}, $1 \le k \le r$, and $j = Y_i$. Thus,

$$\nabla L(\gamma) = \sum \mathbf{F}_i(\gamma) - \sum \mathbf{E}_i(\gamma),$$

where $\mathbf{F}_i(\gamma)$ is the sum of $\mathbf{x}_{ij} q_{ij}(\gamma)/Q_i(\gamma)$ for j in Z_i and $\mathbf{E}_i(\gamma)$ is the sum of $\mathbf{x}_{ij} q_{ij}(\gamma)/R_i(\gamma)$ for j in J_i.

In a similar manner, the Hessian matrix $\nabla^2 L(\gamma)$ of L at γ is the difference for $\gamma = \beta$ between the conditional covariance matrix of $\Sigma \mathbf{X}_i$ given Y_i is in Z_i for $1 \leq i \leq N$ and the unconditional covariance matrix of $\Sigma \mathbf{X}_i$. Thus,

$$\nabla^2 L(\gamma) = D(\gamma) - C(\gamma) = \sum D_i(\gamma) - \sum C_i(\gamma),$$

where $C(\gamma) = \Sigma C_i(\gamma)$, $C_i(\gamma)$ is the sum over j in Z_i of $[\mathbf{x}_{ij} - \mathbf{E}_i(\gamma)]^2 q_{ij}(\gamma)/Q_i(\gamma)$, $D(\gamma) = \Sigma D_i(\gamma)$, and $D_i(\gamma)$ is the sum over j in J_i of $[\mathbf{x}_{ij} - \mathbf{F}_i(\gamma)]^2 q_{ij}(\gamma)/R_i(\gamma)$.

In the case of $\nabla^3 L(\gamma)$,

$$\nabla^3 L(\gamma) = \sum H_i(\gamma) - \sum G_i(\gamma),$$

where $H_i(\gamma)$ is the sum over j in Z_i of $[\mathbf{x}_{ij} - \mathbf{E}_i(\gamma)]^3 q_{ij}(\gamma)/Q_i(\gamma)$ and $G_i(\gamma)$ is the sum over j in J_i of $[\mathbf{x}_{ij} - \mathbf{F}_i(\gamma)]^3 q_{ij}(\gamma)/R_i(\gamma)$.

In the case of direct observation, Z_i is always the set $\{Y_i\}$,

$$L = -B,$$

$$\nabla L(\gamma) = \sum \mathbf{X}_i - \sum \mathbf{E}_i(\gamma),$$

$$\nabla^2 L(\gamma) = -C(\gamma) = -\sum C_i(\gamma),$$

and

$$\nabla^3 L(\gamma) = -\sum G_i(\gamma).$$

2.2. *Maximum Likelihood Estimates*

In the problems of indirect observation considered in this paper, there may be zero, one, or more than one maximum likelihood estimate $\mathbf{b}$ of β (Haberman 1974*a*). Nonetheless, if $\mathbf{b}$ is a maximum likelihood estimate, $\nabla L(\mathbf{b}) = \mathbf{0}$ and $\nabla^2(\mathbf{b})$ is nonpositive definite. If $\nabla L(\mathbf{b})$ is $\mathbf{0}$ and $\nabla^2 L(\mathbf{b})$ is negative definite, then $\mathbf{b}$ is at least an isolated local maximum of the log likelihood L. As in Haberman (1977*c*), the proposed modification of the Newton-Raphson algorithm is designed to reduce the risk of convergence to a solution of the equation $\nabla L(\mathbf{b}) = \mathbf{0}$, which is not a maximum likelihood estimate.

2.3. *The Newton-Raphson Algorithm*

The Newton-Raphson algorithm has a simpler form for the model under study than the scoring algorithm. Let $\mathbf{b}_0$ be an initial

approximation to a maximum likelihood estimate $\mathbf{b}$ such that $\nabla^2(\mathbf{b}_0)$ is negative definite. Under the Newton-Raphson algorithm, approximations $\mathbf{b}_t$ of $\mathbf{b}$ are generated by the equation

$$\mathbf{b}_{t+1} = \mathbf{b}_t + \mathbf{s}(\mathbf{b}_t),$$

where

$$\mathbf{s}(\gamma) = -\left[\nabla^2 L(\gamma)\right]^{-1} \nabla L(\gamma).$$

This algorithm involves only the observed sets Z_i rather than all members of G_i. In scoring, the Hessian matrix is replaced by its estimated expected value given that β is $\mathbf{b}_t$. Thus, knowledge of all possible sets G_i is required.

The convergence properties of the Newton-Raphson algorithm are standard. In this paper, it is helpful to note results of Kantorovich and Akilov (1982, pp. 529–33). Let $\|\cdot\|$ be a norm on R^r, so that for any $r \times r$ matrix $\mathbf{M}$, $\|\mathbf{M}\|$ is the supremum of $\|\mathbf{Mc}\|/\|\mathbf{c}\|$ for $\mathbf{c}$ in R^r. For any $r \times r \times r$ array $\mathbf{S}$ with elements S_{ijk}, $1 \le i \le r$, $1 \le j \le r$, $1 \le k \le r$, let $\mathbf{Sc}$ be the $r \times r$ matrix with elements

$$\sum_k S_{ijk} c_k, \qquad 1 \le i \le r, \qquad 1 \le j \le r.$$

Let T be the initial step length $\|\mathbf{s}(\mathbf{b}_0)\|$, let $u > T$, and let V be the supremum of

$$\left\|\left[\nabla^2 L(\mathbf{b}_0)\right]^{-1} \nabla^3 L(\gamma)\mathbf{c}\right\| / \|\mathbf{c}\|$$

for $\mathbf{c}$ in R^r and $\|\gamma - \mathbf{b}_0\| \le u$. Thus, V is a normalized measure of the size of the third partial derivatives of the log likelihood. If $TV < 1/2$,

$$u \ge \left[1 - (1 - 2TV)^{1/2}\right]/V,$$

and

$$u < \left[1 + (1 - 2TV)^{1/2}\right]/V,$$

then $\mathbf{b}_t$ converges to the unique point $\mathbf{b}$ such that $\nabla L(\mathbf{b})$ is $\mathbf{0}$ and $\|\mathbf{b} - \mathbf{b}_0\| \le u$. We have

$$\|\mathbf{b}_t - \mathbf{b}\| \le T(2TV)^{\tau(n)-1}/\tau(n),$$

where $\tau(n) = 2^n$. It easily follows that if $\mathbf{b}$ is a maximum likelihood estimate of β such that $\nabla^2 L(\mathbf{b})$ is negative definite, then for $\mathbf{b}_0$ sufficiently close to $\mathbf{b}$, $\mathbf{b}_t$ converges to $\mathbf{b}$. Since the third partial

derivatives of L are uniformly bounded, it follows from the case $\mathbf{b}_0 = \mathbf{b}$ that if $V^* < \infty$ is the supremum of

$$\left\| \left[\nabla^2 L(\mathbf{b}) \right]^{-1} \nabla^3 L(\gamma)\mathbf{c} \right\| / \|\mathbf{c}\|$$

for $\mathbf{c}$ and γ in R^r, then no $\mathbf{c}$ exists such that $\nabla L(\mathbf{c})$ is $\mathbf{0}$ and $\|\mathbf{c} - \mathbf{b}\| < 2/V^*$. Thus, the conditions $\nabla L(\mathbf{b}) = \mathbf{0}$ and $\nabla^2 L(\mathbf{b})$ negative definite ensure that $\mathbf{b}$ is the only local maximum of the likelihood within an open ball of radius $2/V^*$.

In problems of indirect observation, this algorithm is often quite satisfactory. Example 1 provides a good example. Consider the crude starting values $b_{k0} = 0$, $1 \le k \le 4$, $b_{k0} = 1$, $5 \le k \le 7$, $b_{k0} = 0$, $8 \le k \le 9$. Convergence is rapid enough that no b_{k4} differs from b_{k5} by more than 0.00004. If we use Haberman's (1979, p. 549) starting values for the scoring algorithm, then no b_{k2} differs from b_{k3} by more than 0.00091, a result similar to that obtained via the scoring algorithm.

2.4. *Large-Sample Approximations*

A basic advantage of the Newton-Raphson algorithm over the EM algorithm is in the estimation of asymptotic standard deviations of parameters. The EM algorithm provides no information on asymptotic standard deviations. In the Newton-Raphson algorithm, estimates for asymptotic standard deviations of the maximum likelihood estimates are obtained as by-products. Let $\mathbf{c}$ be a nonzero constant vector in R^r. The approximation used states that the normalized difference

$$z = (\mathbf{c}'\mathbf{b} - \mathbf{c}'\beta) \Big/ \left\{ \mathbf{c}' \left[-\nabla^2 L(\mathbf{b}) \right]^{-1} \mathbf{c} \right\}^{1/2}$$

has an approximate standard normal distribution. The estimated asymptotic standard deviation (EASD) of $\mathbf{c}'\mathbf{b}$ is thus $\{\mathbf{c}'[\nabla^2 L(\mathbf{b})]^{-1}\mathbf{c}\}^{1/2}$. The estimates and estimated standard deviations of the coefficients b_k in example 1 are listed in Table 3. These results do not differ appreciably from those derived from Haberman (1979) via the scoring algorithm.

To appreciate the significance of the ability to compute estimated asymptotic standard deviations, it is helpful to note that the normal approximation for z is appropriate under a very wide variety of conditions. In particular, it is possible to use the normal approximation even when the vectors $\mathbf{x}_{ij}$ vary for each subject i or the dimension p of the parameter vector β depends on the sample size N.

TABLE 3
Parameter Estimates and Estimated Asymptotic Standard Deviations for Example 1

Parameter	Estimate	EASD
λ_1^U	−0.106	0.087
λ_1^A	−0.316	0.045
λ_1^B	0.327	0.049
λ_1^C	0.012	0.039
λ_{11}^{UA}	1.372	0.045
λ_{11}^{UB}	1.397	0.049
λ_{11}^{UC}	1.293	0.039
λ_{11}^{UD}	−0.105	0.026
λ_{12}^{UD}	0.045	0.026

To study the normal approximation for $\mathbf{b}$, consider the Newton-Raphson algorithm for the starting value $\mathbf{b}_0 = \beta$. Let the norm $\|\mathbf{a}\| = \{\mathbf{a}'[-\nabla^2 L(\beta)]\mathbf{a}\}^{1/2}$, and let $u = 2T$, so that

$$T = \left\{[\nabla L(\beta)]'[-\nabla^2 L(\beta)]^{-1}\nabla L(\beta)\right\}^{1/2},$$

and V is the supremum of

$$\mathbf{d}'[\nabla^3 L(\gamma)\mathbf{a}]\mathbf{d}/(\|\mathbf{a}\|^{1/2}\|\mathbf{d}\|)$$

for $\mathbf{a}$ and $\mathbf{d}$ in R^r and $\|\gamma - \beta\| \le 2T$. Let the Fisher information $I(\beta)$ be the expected value of $-\nabla^2 L(\beta)$. Let v be the maximum difference

$$\left|\mathbf{a}'[-\nabla^2 L(\beta)]\mathbf{a}/\mathbf{a}'I(\beta)\mathbf{a} - 1\right|$$

for $\mathbf{a}$ in R^r, and let

$$\kappa = \sum E\left|\mathbf{c}'[I(\beta)]^{-1}[\mathbf{F}_i(\beta) - \mathbf{E}_i(\beta)]\right|^3 \Big/ \left(\mathbf{c}'[I(\beta)]^{-1}\mathbf{c}\right)^{3/2}$$

For $TV < 1/2$, let $\mathbf{b}$ be the limit of $\mathbf{b}_t$. Since Taylor's theorem implies that

$$\mathbf{a}'\nabla^2 L(\mathbf{b})\mathbf{a} = \mathbf{a}'\nabla^2 L(\beta)\mathbf{a} + \mathbf{a}'[\nabla^3 L(\mathbf{d})(\mathbf{b} - \beta)]\mathbf{a}$$

for some $\mathbf{d}$ on the line segment between $\mathbf{b}$ and β, the definition of V and the condition $TV < 1/2$ imply that $\mathbf{b}$ is at least an isolated relative maximum of L, with no other critical point $\mathbf{b}^*$ such that $\|\mathbf{b}^* - \beta\| <$

$2T$. If T^2V converges in probability to 0, v converges in probability to 0, and κ converges to 0, then the distribution function of z converges to the distribution function Φ of the standard normal distribution $N(0,1)$. The arguments required differ little from those in Haberman (1977*b*, 1977*c*) and Friedman (1982).

In the special case of direct observation, results correspond to those in Haberman (1977*b*, 1977*c*). In this case, $I(\beta)$ and $-\nabla^2 L(\beta)$ are the same, an isolated local maximum of L is necessarily the maximum likelihood estimate, $\nabla^2 L$ is independent of the observations, and the condition on third absolute moments is redundant.

The conditions described here are trivial in the simple case of $\mathbf{x}_{ij}$, J_i, and G_i independent of i and $I(\beta)$ positive definite. As in Haberman (1977*b*, 1977*c*), the distribution of T^2 converges to a central chi square on p degrees of freedom, and V, v, and k are all of order $N^{-1/2}$. This situation applies in Table 1.

As in results of Haberman (1977*b*, 1977*c*), there is no need for $\mathbf{x}_{ij}$, J_i, and G_i to be constant over i for the normal approximation to apply. For example, if p is constant, $N[I(\beta)]^{-1}$ is bounded, and the x_{ijk} are uniformly bounded, then the distribution of T^2 still approaches that of a central chi square on p degrees of freedom, and k, V, and v are still of order $N^{-1/2}$, so that all conditions for asymptotic normality are satisfied. For related results for constant p, J_i, and G_i, see Fahrmeir and Kaufmann (1985). Arguments similar to those presented in Haberman (1977*b*, 1977*c*) can also be applied when the dimension p increases as the sample size N increases.

The problem of local maxima that are not unique global maxima is quite real, as is evident in the latent-class model for Table 1. We have $L(\mathbf{b}) = L(\mathbf{c})$ if $b_k = -c_k$ for $k = 1, 5, 6, 7, 8$, and 9 and $b_k = c_k$ for $2 \le k \le 4$. The asymptotic normality result for $\mathbf{b}$ holds only if we assume that b_4 and β_4 are both positive.

The following variation on the asymptotic normality conditions is occasionally relevant. Instead of $u = 2T$, let $u = WT$ for W a random variable at least 2. Let $\|\cdot\|$ be defined as before, and let the probability approach 1 that $TV < 1/2$. Then the probability approaches 1 that there is only one critical point $\mathbf{b}$ such that $\|\mathbf{b} - \beta\| < 1/V$. In the simple case of $\mathbf{x}_{ij}$, J_i, and G_i independent of i and $I(\beta)$ positive definite, it follows, as in Haberman (1977*a*), that for some open neighborhood $\mathbf{M}$ of β, the probability approaches 1 that there is exactly one critical point in $\mathbf{M}$. These conditions are met in example 1.

2.5. *Stability Problems*

As is evident from the example, the Newton-Raphson algorithm can work quite well in practice; nonetheless, it is not especially stable, even when $\nabla^2 L(\mathbf{b})$ is negative definite. For instance, in example 2, $\nabla^2 L(\mathbf{b}_0)$ is not negative definite if $b_{k0} = 0$ for $1 \le k \le 6$ and $b_{k0} = 1$ for $7 \le k \le 11$. This problem can be overcome by better initial values b_{k0}; however, as is evident from Goodman (1974*a*), good initial estimates cannot be constructed in a trivial fashion.

There are two basic problems that affect stability. The first problem, which is observed in example 2, is that the logarithm of the likelihood function need not be concave. Thus, the Hessian of L need not be negative definite at all points, and for fixed $\mathbf{b}$, $g(\lambda) = L(\mathbf{b} + \lambda \mathbf{s}(\mathbf{b}))$, which has derivative

$$g'(0) = f(\mathbf{b}) = [\nabla L(\mathbf{b})]'[\nabla^2 L(\mathbf{b})]^{-1} \nabla L(\mathbf{b}),$$

may be a decreasing function for $\mathbf{a}$ in the interval $[0, 1]$. This problem can clearly lead to cases in which $L(\mathbf{b}_t)$ exceeds $L(\mathbf{b}_{t+1})$ and no progress is made toward a maximum of the function L. In addition, computation of $\mathbf{s}(\mathbf{b}_t)$ is simpler if the Hessian $\nabla^2 L(\mathbf{b}_t)$ is negative definite. A related difficulty arises: Whenever the Hessian of L is not negative definite at all points, there are also points at which the Hessian is negative definite but nearly singular. This problem can lead to very large vectors $\mathbf{s}(\mathbf{b}_t)$. Instability results because $\mathbf{b}_{t+1}$ is then very far from $\mathbf{b}_t$.

The second problem involves deviation of L from a quadratic approximation. Let $\nabla^2 L(\mathbf{b})$ be negative definite. Were L quadratic, then L would have a unique maximum at $\mathbf{b} + \mathbf{s}(\mathbf{b})$. At this maximum, the value of L would be $L(\mathbf{b}) + (1/2)f(\mathbf{b})$. If the quadratic approximation is poor, then $L(\mathbf{b} + \mathbf{s}(\mathbf{b}))$ may be less than $L(\mathbf{b})$, even though $g'(0)$ is positive and $L(\mathbf{b} + \lambda \mathbf{s}(\mathbf{b}))$ exceeds $L(\mathbf{b})$ for sufficiently small positive λ. In the algorithm considered below, both of these issues are considered.

3. THE ALGORITHM

In the proposed algorithm, a fixed norm $\|\cdot\|$, a fixed α in $(0, 1/2)$, a fixed $\kappa > 0$, and a fixed τ in $(0, 1/2)/(1 - \alpha)$ are selected, and the ordinary Newton-Raphson algorithm is modified

whenever either $\nabla^2 L(\mathbf{b}_t)$ is not negative definite, $\|\mathbf{s}(\mathbf{b}_t)\| > k$, or

$$L(\mathbf{b}_t) + \alpha \mathbf{s}(\mathbf{b}_t)' \nabla L(\mathbf{b}_t) > L(\mathbf{b}_t + \mathbf{s}(\mathbf{b}_t)).$$

To describe the algorithm, assume for simplicity that $\mathbf{b}$ is a critical point of L such that $\nabla^2 L(\mathbf{b})$ is negative definite. Since $C(\mathbf{b})$ must be positive definite, it follows that if $\sum_k c_k x_{ijk}$ is constant over j for each i, then each c_k is 0. Therefore, $C(\gamma)$ is positive definite for all γ.

Let $\mathbf{b}_0$ be an initial approximation for $\mathbf{b}$. Recall the definitions of $\nabla L(\gamma)$, $\nabla^2 L(\gamma)$, $C(\gamma)$, and $D(\gamma)$ in section 2.1. Then, a sequence of approximations $\mathbf{b}_t$, $t \geq 0$, is constructed in the following fashion. The sequence satisfies

$$\mathbf{b}_{t+1} = \mathbf{b}_t + \lambda_t \mathbf{u}(\mathbf{b}_t),$$

where

$$\mathbf{u}(\gamma) = [C(\gamma) - m(\gamma) D(\gamma)]^{-1} \nabla L(\gamma),$$

$m(\gamma)$ is 1 if $-\nabla^2 L(\gamma) = C(\gamma) - D(\gamma)$ is positive definite and $\|\mathbf{s}(\gamma)\| \leq \kappa$ and $m(\gamma)$ is 0 otherwise, and λ_t is obtained by the following steps:

1. Let c_{t1} be $\max[1, \kappa / \|\mathbf{u}(\mathbf{b}_t)\|]$, let $k = 1$, and let $e_t = [\mathbf{u}(\mathbf{b}_t)]' \nabla L(\mathbf{b}_t)$.
2. If

$$L(\mathbf{b}_t + c_{tk} \mathbf{u}(\mathbf{b}_t)) \geq L(\mathbf{b}_t) + \alpha c_{tk} e_t,$$

   then let $\lambda_t = c_{tk}$. Otherwise, continue on to step 3.
3. Let

$$c_{t(k+1)} = \max\left[\tau c_{tk}, (1/2) c_{tk} e_t / \left\{ e_t - \left[L(\mathbf{b}_t + c_{tk} \mathbf{u}(\mathbf{b}_t)) - L(\mathbf{b}_t) \right] / c_{tk} \right\}\right],$$

   replace k by $k + 1$, and return to step 2.

3.1. *Relationship to the Newton-Raphson Algorithm*

The proposed algorithm is identical to the Newton-Raphson algorithm once $\mathbf{b}_t$ is sufficiently close to $\mathbf{b}$. Thus, the algorithm retains the rapid convergence properties of the Newton-Raphson algorithm. Verification of this identity for $\mathbf{b}_t$ near $\mathbf{b}$ reduces to demonstration that

$$\mathbf{u}(\gamma) = \mathbf{s}(\gamma) \text{ for } \gamma \text{ sufficiently close to } \mathbf{b} \tag{2}$$

and

$$L(\gamma + \mathbf{s}(\gamma)) \geq L(\gamma) + \alpha \mathbf{s}(\gamma)]' \nabla L(\gamma) \text{ for } \gamma \text{ sufficiently close to } \mathbf{b}. \tag{3}$$

Since $\nabla^2 L(\mathbf{b}) = D(\mathbf{b}) - C(\mathbf{b})$ is negative definite and $\nabla L(\mathbf{b})$ is $\mathbf{0}$, $\mathbf{u}(\gamma) = \mathbf{s}(\gamma)$ for γ sufficiently close to $\mathbf{b}$. Demonstration of (3) depends on standard use of Taylor expansions. Since $\nabla L(\mathbf{b})$ is $\mathbf{0}$ and $\nabla^2 L(\mathbf{b})$ is negative definite, $\mathbf{s}(\gamma) \to \mathbf{0}$ as $\gamma \to \mathbf{0}$. Standard use of Taylor expansions yields the following results:

$$\nabla L(\gamma) = \nabla^2 L(\mathbf{b})\gamma + \mathbf{o}(\gamma),$$

where $\mathbf{o}(\gamma)/\|\gamma - \mathbf{b}\| \to \mathbf{0}$ as $\gamma \to \mathbf{b}$;

$$[\mathbf{s}(\gamma)]'\nabla L(\gamma) = -[\nabla L(\gamma)]'[\nabla^2 L(\gamma)]^{-1}\nabla L(\gamma)$$
$$= -\gamma'\nabla^2 L(\mathbf{b})\gamma + o(\gamma),$$

where $o(\gamma)/\|\gamma - \mathbf{b}\|^2 \to 0$;

$$L(\gamma + \mathbf{s}(\gamma)) = L(\gamma) + [\mathbf{s}(\gamma)]'\nabla L(\gamma) + (1/2)[\mathbf{s}(\gamma)]'\nabla^2 L(\gamma)\mathbf{s}(\gamma) + o_1(\gamma)$$
$$= L(\gamma) - (1/2)\gamma'\nabla^2 L(\mathbf{b})\gamma + o_2(\gamma),$$

where for w equal 1 or 2, $o_w(\gamma)/\|\gamma - \mathbf{b}\|^2 \to 0$ as $\gamma \to \mathbf{b}$;

$$L(\gamma) + \alpha[\mathbf{s}(\gamma)]'\nabla L(\gamma) = L(\gamma) - \alpha\gamma'\nabla^2 L(\mathbf{b})\gamma + o(\gamma).$$

Since α is between 0 and 1/2, (3) holds.

The modifications of the Newton-Raphson algorithm have several purposes. As shown in section 3.2, the use of $\mathbf{u}(\mathbf{b}_t)$ rather than $\mathbf{s}(\mathbf{b}_t)$ ensures that $w_t(\lambda) = L(\mathbf{b}_t + \lambda\mathbf{u}(\mathbf{b}_t))$ is increasing in λ for small positive λ. To understand the choice of $\mathbf{u}(\mathbf{b}_t)$ for $\nabla^2 L(\mathbf{b}_t)$ not negative definite, consider the following hypothetical situation. Suppose that the EM algorithm were used at step t to define $\mathbf{b}_{t+1}$ and that the Newton-Raphson algorithm with starting value $\mathbf{b}_t$ were applied to the M (maximization) step of the EM algorithm. Then, the first iteration of this Newton-Raphson algorithm would yield $\mathbf{b}_t + \mathbf{u}(\mathbf{b}_t)$.

The scale factor λ_t is intended to approximate the maximum of $w_t(\lambda)$ over λ. The sequence of c_{tk} is normally selected so that $c_{t(k+1)}$ is the location of the maximum of the quadratic function $f_t(\lambda)$ such that at 0, $f_t(\lambda)$ equals $L(\mathbf{b}_t)$ and has derivative e_t, and at c_{tk}, $f_t(\lambda)$ equals $L(\mathbf{b}_t + c_{tk}\mathbf{u}(\mathbf{b}_t))$. However, to prevent unusually small values of c_{tk}, $c_{t(k+1)}$ is constrained to be at least τc_{tk}. To prevent extremely large changes of approximations to $\mathbf{b}$, c_{t1} is restricted to ensure that $\|\mathbf{b}_{t+1} - \mathbf{b}_t\|$ does not exceed κ.

3.2. *Stability Properties*

The algorithm has stability properties that the Newton-Raphson algorithm does not possess. The basic result is that $\mathbf{b}_t$ converges to $\mathbf{b}$ if $W = \{\gamma\colon L(\gamma) \geq L(\mathbf{b}_0)\}$ is bounded and if $\mathbf{b}$ is the only element γ of W such that $\nabla L(\gamma) = \mathbf{0}$.

Stability is ensured by the condition that $L(\mathbf{b}_t)$ is less than $L(\mathbf{b}_{t+1})$ unless $\nabla L(\mathbf{b}_t) = \mathbf{0}$. To verify this claim, assume that $\nabla L(\mathbf{b}_t) \neq \mathbf{0}$. We must verify that $e_t > 0$ and that λ_t is defined and positive. Since the matrix $C(\mathbf{b}_t) - m(\mathbf{b}_t)D(\mathbf{b}_t)$ is positive definite,

$$e_t = [\mathbf{u}(\mathbf{b}_t)]'\nabla L(\mathbf{b}_t) = [\nabla L(\mathbf{b}_t)]'[C(\mathbf{b}_t) - m(\mathbf{b}_t)D(\mathbf{b}_t)]^{-1}\nabla L(\mathbf{b}_t) > 0.$$

Since $e_t > 0$,

$$L(\mathbf{b}_t + \lambda\mathbf{u}(\mathbf{b}_t)) = L(\mathbf{b}_t) + \lambda e_t + o_3(\lambda),$$

where $o_3(\lambda)/\lambda \to 0$ as $\lambda \to 0$. Thus,

$$L(\mathbf{b}_t + \lambda\mathbf{u}(\mathbf{b}_t)) > L(\mathbf{b}_t) + \alpha\lambda e_t$$

for small enough positive λ. If λ_t is not set equal to c_{tk}, then $c_{t(k+1)}$ must be positive but less than $(1/2)c_{tk}/(1-\alpha)$. Thus, a k must eventually be encountered such that λ_t can be set equal to $c_{tk} > 0$.

Since the $L(\mathbf{b}_t)$, $t \geq 0$, form a nondecreasing sequence and since each $L(\mathbf{b}_t) \leq L(\mathbf{b})$, $L(\mathbf{b}_{t+1}) - L(\mathbf{b}_t) \to 0$ and $\lambda_t e_t \to 0$. The assumption that $W = \{\mathbf{g}\colon L(\mathbf{g}) \geq L(\mathbf{b}_0\}$ is bounded implies that $\{\mathbf{b}_t\colon t \geq 1\}$ is contained in the closed and bounded set W. Thus e_t, $\|\mathbf{u}(\mathbf{b}_t)\|$, and $\|\lambda_t\mathbf{u}(\mathbf{b}_t)\|$ are bounded above. Since

$$w_t(\lambda) = w_t(0) + \lambda e_t + (1/2)\lambda^2[\mathbf{u}(\mathbf{b}_t)]'\nabla^2 L(\mathbf{d}_t(\lambda))\mathbf{u}(\mathbf{b}_t)$$

for some $\mathbf{d}_t(\lambda) = \mathbf{b}_t + \rho\mathbf{u}(\mathbf{b}_t))$, $0 \leq \rho \leq \lambda$, we can use constants η and ρ to obtain bounds

$$w_t(0) + \lambda e_t + \eta\lambda^2 \leq w_t(\lambda) \leq w_t(0) + \lambda e_t + \rho\lambda^2$$

for $\lambda \leq c_{t1} = \max[1, \kappa/\|\mathbf{u}(\mathbf{b}_t)\|]$. If $\lambda < c_{t1}$, then

$$w_t(\lambda_t/\mu_t) < w_t(0) + \alpha\lambda_t e_t/\mu_t$$

for some μ_t between τ and 1 such that $\lambda_t/\mu_t \leq c_{t1}$. Thus, λ_t is bounded away from 0 and $e_t \to 0$. Since the $C(\mathbf{b}_t) - m(\mathbf{b}_t)D(\mathbf{b}_t)$ form a bounded sequence of matrices if W is bounded, it follows that $\nabla L(\mathbf{b}_t) \to \mathbf{0}$. Since $\mathbf{b}$ is the only element of W such that $\nabla L(\mathbf{b}) = \mathbf{0}$, $\mathbf{b}_t \to \mathbf{b}$.

One modest generalization is important in many problems involving latent-class analysis. Suppose that W contains a finite num-

ber of solutions of the equation $\nabla L(\mathbf{b}) = \mathbf{0}$. Then, $\mathbf{b}_t$ converges to one of these solutions.

4. EXAMPLES

The examples presented in this section are designed to illustrate the stability of the modified Newton-Raphson algorithm when quite crude starting values are used. In all examples, α is $1/16$, τ is 0.1, and $\kappa = 10$. The norm

$$\|\mathbf{c}\| = \max_i \max_j \left| \sum_k x_{ijk} c_k \right|.$$

Example 1. Results for the two-variable latent-class model of example 1 with the starting values previously used for the Newton-Raphson algorithm and results for the modified Newton-Raphson algorithm are exactly the same. Thus, the algorithm does not interfere in a case in which convergence is achieved by the conventional algorithm.

Example 2. Consider the starting values tried in the two-variable latent-class model of example 2. In this case, convergence is rapid; i.e., no value of $|b_{k6} - b_{k7}|$ is greater than 0.00048. In this particular example, $\mathbf{u}(\mathbf{b}_0)$ and $\mathbf{u}(\mathbf{b}_1)$ differ from $\mathbf{s}(\mathbf{b}_0)$ and $\mathbf{s}(\mathbf{b}_1)$, respectively. In all subsequent iterations t, $\mathbf{u}(\mathbf{b}_t) = \mathbf{s}(\mathbf{b}_t)$. The step size λ_t is one for all iterations. Final estimates and estimated asymptotic standard deviations are shown in Table 4. The parameters reported in this table can be produced from numbers reported in Goodman (1974*a*); however, the estimated asymptotic standard deviations cannot be obtained from this source.

Further insight into the behavior of the algorithm can be gained through some further experimentation with starting values. We can make the algorithm fail to converge to a maximum likelihood estimate if we let $\mathbf{b}_0$ be 0. In this case, the algorithm converges to the saddle point of L that corresponds to the maximum likelihood estimate of β subject to the independence and equiprobability restrictions that only λ_A, λ_B, λ_C, and λ_D may differ from 0. This problem of saddle points corresponding to very simple initial values can be expected in all latent-class models.

TABLE 4
Parameter Estimates and Asymptotic Standard Deviations for Example 2

Parameter	Estimate	EASD
λ_U	−0.025	0.279
λ_V	−0.087	0.221
λ_A	−0.239	0.082
λ_B	0.102	0.090
λ_C	−0.048	0.186
λ_D	0.191	0.092
λ_{UA}	0.800	0.073
λ_{UC}	1.204	0.163
λ_{UV}	0.304	0.035
λ_{VB}	0.608	0.064
λ_{VD}	0.611	0.066

Nonetheless, even very poor starting values lead to convergence to the maximum likelihood estimate **b** if we avoid accidental assumptions of independence. For example, the algorithm was used with all b_{k0} equal to 0, except for b_{80} and $b_{(11)0}$, which were set to 0.1. Satisfactory convergence to **b** was achieved after 15 iterations. Because of the poor starting values, $\mathbf{s}(\mathbf{b}_t)$ was not used until t was 8, and in two subsequent instances, λ_t was not 1. Nonetheless, the algorithm proved quite stable under the circumstances.

REFERENCES

Clogg, C. C. 1977. "Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users." Working Paper No. 1977-09. University Park: Penn State University, Population Issues Research Center.

Coleman, J. S. 1964. *Introduction to Mathematical Sociology*. New York: Free Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, ser. B, 39:1–39.

Fahrmeir, L., and H. Kaufmann. 1985. "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models." *Annals of Statistics* 13:342–68.

Friedman, M. 1982. "Piecewise Exponential Models for Survival Data with Covariates." *Annals of Statistics* 10:101–13.

Goodman, L. A. 1974*a*. "The Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. I. A Modified Latent Structure Approach." *American Journal of Sociology* 79:1179–1259.

_____. 1974*b*. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–31.

Haberman, S. J. 1974*a*. "Log-Linear Models for Frequency Tables Derived by Indirect Observation: Maximum Likelihood Equations." *Annals of Statistics* 2:911–24.

_____. 1974*b*. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

_____. 1976. "Iterative Scaling Procedures for Log-Linear Models for Frequency Tables Derived by Indirect Observation." Pp. 45–50 in *Proceedings of the Statistical Computing Section, American Statistical Association*. Washington, DC: American Statistical Association.

_____. 1977*a*. "Product Models for Frequency Tables Involving Indirect Observation." *Annals of Statistics* 5:1124–47.

_____. 1977*b*. "Log-Linear Models and Frequency Tables with Small Expected Cell Counts." *Annals of Statistics* 5:1148–69.

_____. 1977*c*. "Maximum Likelihood Estimates in Exponential Response Models." *Annals of Statistics* 5:815–41.

_____. 1979. *Analysis of Qualitative Data*. Vol. 2. New York: Academic Press.

Kantorovich, L. V., and G. P. Akilov. 1982. *Functional Analysis*. 2d ed. Oxford: Pergamon Press.