The Annals of Statistics
1997, Vol. 25, No. 1, 351–377

CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATORS IN GENERAL RANDOM EFFECTS MODELS FOR BINARY DATA

BY STEVEN M. BUTLER AND THOMAS A. LOUIS

University of Kentucky and University of Minnesota

We consider a general random effects model for repeated binary measures, assuming a latent linear model with any class of mixing distributions. The latent model is assumed to have the Laird–Ware structure, but the random effects may be from any specified class of multivariate distributions and the error vector may have any specified continuous distribution. Elementwise threshold crossing then gives the observed vector of binary outcomes. Special cases of this model include recently discussed mixed logistic regression and probit models, which have had either parametric (usually Gaussian) or nonparametric mixing distributions. We give sufficient conditions for identifiability of the mixing distribution and fixed effects and for convergence of maximum likelihood estimators for the mixing distribution and fixed effects. As expected, the conditions are much stronger for nonparametric mixing than for Gaussian mixing. We illustrate the conditions by applying them to a practical example.

1. Introduction. A number of authors have proposed random effects models for repeated binary measures. Some have assumed a parametric mixing distribution, usually Gaussian, while others have assumed nonparametric mixing. Maximum likelihood estimation for models with Gaussian mixing can be very cumbersome due to the integral form of the marginal likelihood, particularly for multivariate random effects. Several authors have used maximum likelihood estimation for a univariate Gaussian random intercept, including Mislevy (1985), Bock and Aitken (1981), Anderson and Aitken (1985) and Im and Gianola (1988), who maximize the marginal likelihood using the EM algorithm with numerical integration at each iteration. Conaway (1990) avoids numerical integration for the case of a random intercept by assuming a log–gamma mixing distribution and a log–log link function. For multivariate Gaussian random effects, a variety of methods for estimation have been proposed, for example, by Korn and Whittemore (1979), Stiratelli, Laird and Ware (1984), Harville and Mee (1984) for ordinal data, Gilmore, Anderson and Rae (1985), Zeger, Liang and Albert (1988) and Zeger and Karim (1991).

A nonparametric random intercept has been assumed by Feinberg, Bromet, Follman, Lambert and May (1985), Follman and Lambert (1989), Lindsay,

Received October 1992; revised October 1995.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62J02, 62J12.

Key words and phrases. Binary, repeated measures, nonparametric, semiparametric, random effects, maximum likelihood, consistency, identifiability.

Clogg and Grego (1991) and Butler and Louis (1992), who use maximum likelihood estimation with a discrete mixing distribution. The general results of Laird (1978) and Lindsay (1983) then imply that the marginal likelihood is maximized over all compact support mixing distributions, if the number of mass points in the mixing distribution is at least as large as the number of distinct observations. A lower limit is possible in special cases, as demonstrated by De Leeuw and Verhelst (1986), Follman and Lambert (1991) and Lindsay, Clogg and Grego (1991).

We consider a threshold crossing model for repeated binary measures, assuming a latent linear random effects model. This latent model has the Laird–Ware structure [Laird and Ware (1982)]. However, in place of their Gaussian assumptions, we allow any specified class of multivariate distributions for the random effects, and we allow the error vector to have any specified strictly increasing continuous c.d.f. This model includes most of the above models as special cases. We give sufficient conditions for identifiability of the mixing distribution and fixed effects and for consistency of maximum likelihood estimators for the mixing distribution and fixed effects. We define consistency for estimators of the mixing distribution in terms of convergence in distribution.

Conditions for consistency of maximum likelihood estimators for general mixing distributions have been given by Kiefer and Wolfowitz (1956), who assume identifiability, Pfanzagl (1988) and van der Vaart and Wellner (1992). However, for reasons discussed later, we must use another approach. We first demonstrate a type of convergence for the marginal distribution, relying on very specific properties of our model, and this is used along with identifiability conditions to prove consistency.

Conditions for the identifiability of general mixtures have been provided in a broad context by Tallis (1969) and Tallis and Chesson (1982). These are difficult to apply even in simple cases, as discussed in Maritz and Lwin (1989). Simpler conditions are possible for location parameter mixtures [Teicher (1961) and Maritz and Lwin (1989)], and these are extended as part of our proof of identifiability.

We describe the model in Section 2. In Section 3 we give some conditions for convergence of the marginal distribution with maximum likelihood estimation. In Section 4 we give conditions sufficient for identifiability and for consistency given marginal convergence. These conditions are quite strong when we allow a general nonparametric class of mixing distributions, and, as expected, they are much weaker for a specific parametric case such as Gaussian mixing. In Section 5 we discuss the application of our conditions to a practical situation. In Section 6 we discuss nonparametric maximum likelihood estimation using discrete mixtures.

2. The model. For each individual $i = 1, 2, \ldots, n$, we assume the existence of a latent linear model

$$T_i = X_i \alpha + Z_i b_i + \varepsilon_i,$$

where $T_i \in R^m$ is an $m \times 1$ vector of unobserved latent variables, $\alpha \in R^s$ is an $s \times 1$ vector of fixed effects, $b_i \in R^r$ is an $r \times 1$ vector of random effects, $\varepsilon_i \in R^m$ is an $m \times 1$ random error vector and $X_i \in R^{(m \times s)}$ and $Z_i \in R^{(m \times r)}$ are observed covariate matrices. The error vectors ε_i are i.i.d. among individuals with a specified continuous c.d.f. F. The random effects b_i are i.i.d. among individuals according to an unknown c.d.f. G. The vectors b_i and ε_i are independent, and both ε_i and b_i are independent of the covariates X_i and Z_i. Let Q be a class of probability measures on the Borel sets $\mathscr{B}^r$ in R^r. We suppose that the measure associated with G is contained in Q. For convenience, we often write this as "$G \in Q$." Two important examples for Q are the class of all probability measures on $\mathscr{B}^r$ and the class of Gaussian distributions on R^r.

Let $W_i = (X_i, Z_i)$, where $(\cdot, \cdot)$ indicates horizontal concatenation, and suppose that the W_i are restricted to some Borel set $S \subset R^{m \times (s+r)}$. The W_i may be selected randomly, systematically or by a combination of both types of procedure. Let $\rho^{(n)}$ be the empirical measure on S generated by $\{W_i,\ i = 1, 2, \ldots, n\}$. Let ρ be a probability measure on the Borel sets $\mathscr{S}$ in S. For example, in Theorem 3.1 we assume that $\rho^{(n)}$ converges in measure to ρ with probability 1. This assumption holds if the W_i are selected randomly and i.i.d. according to ρ, but it can also hold with systematic selection, say by assigning a fixed proportion of individuals to each of several treatment groups.

We observe the $m \times 1$ vector Y_i such that $Y_{ij} = 1$ if $T_{ij} \geq 0$ and $Y_{ij} = 0$ if $T_{ij} < 0$, $j = 1, 2, \ldots, m$. Let S_y be the set of all $m \times 1$ vectors with elements equal to 0 or 1. Then we have the following model: for $i = 1, 2, \ldots, n$,

$$(2.1)\quad P(Y_i = y | W_i, \alpha, b_i) = P\big((X_i\alpha + Z_i b_i + \varepsilon_i) * (2y - \mathbf{1}) \geq \mathbf{0} | W_i, b_i\big)$$

for each $y \in S_y$, where $\mathbf{1}$ is a vector of unit elements, $\mathbf{0}$ is a vector with all elements equal to 0 and "$*$" denotes elementwise multiplication. For example, if we let F^- be the c.d.f. for $-\varepsilon_i$, then $\Pr(Y_i = \mathbf{1} | W_i, \alpha, b_i) = F^-(X_i\alpha + Z_i b_i)$. Let $(\Omega, \mathscr{F}, P)$ be the probability space with σ-field $\mathscr{F}$ and measure P generated by the random components of $\{(Y_i, W_i, b_i, \varepsilon_i),\ i = 1, 2, \ldots\}$.

We write $W_i = w = (x, z)$ to indicate that $X_i = x$ and $Z_i = z$. Let $M(y|w, \alpha, G)$ be the marginal probability that $Y_i = y$ given that $W_i = w$, defined as

$$\begin{aligned} M(y|w, \alpha, G) &\equiv P(Y_i = y | W_i = w, \alpha, G) \\ &= E_G\big[P\big((x\alpha + zb + \varepsilon) * (2y - \mathbf{1}) \geq \mathbf{0} | b\big)\big] \end{aligned}$$

for each $y \in S_y$. For example, $M(\mathbf{1}|w, \alpha, G) = E_G[F^-(x\alpha + zb)]$. Notice that M is continuous in w on $R^{m \times (s+r)}$.

An important special case of model (2.1) is the "conditional independence" model, in which the elements of ε are independent. Then the elements of $Y|(W = w, b)$ are conditionally independent. If the elements of ε_i are identically distributed, as is usually assumed, and the elements of $-\varepsilon_i$ are dis-

tributed according to some c.d.f. F_1, then

$$\begin{aligned}&\Pr(Y_i = y|W_i = w, \alpha, b)\\ &\quad = \prod_{j=1}^{m}\left[F_1^{y_j}(x_j\alpha + z_j b)\big(1 - F_1(x_j\alpha + z_j b)\big)^{(1-y_j)}\right],\end{aligned} \tag{2.2}$$

where y_j is the jth element of y, and where x_j and z_j are the jth rows of x and z. This model includes most of the cases considered by the authors discussed in Section 1.

For the general model (2.1), we say that G_n and α_n are maximum likelihood estimators (m.l.e.'s) for G and α if they simultaneously maximize the joint marginal likelihood for $(Y_i|W_i)$, $i = 1, 2, \ldots, n$, over $G \in \mathcal{Q}$ and $\alpha \in R^s$. In other words, if

$$\prod_{i=1}^{n}\left[M(Y_i|W_i, \alpha_n, G_n)\right] = \max_{\phi, A}\prod_{i=1}^{n}\left[M(Y_i|W_i, \phi, A)\right],$$

where the maximum is over $\phi \in R^s$ and $A \in \mathcal{Q}$.

We give sufficient conditions on model (2.1) for identifiability of G and α, and we give sufficient conditions for consistency of the m.l.e.'s for G and α. We say that G and α are identifiable when we have the following implication: if $G' \in \mathcal{Q}$, $G \in \mathcal{Q}$, $\alpha' \in R^s$, $\alpha \in R^s$ and $M(\mathbf{1}|w, \alpha', G') = M(\mathbf{1}|w, \alpha, G)$ almost everywhere (ρ), then $G' = G$ and $\alpha' = \alpha$. We say that a sequence of estimators $\{G_n \in \mathcal{Q},\ \alpha_n \in R^s,\ n = 1, 2, \ldots\}$ is consistent for G and α if $G_n \to_{\mathscr{D}} G$ and $\alpha_n \to \alpha$, each with probability 1 (P).

The general conditions of Kiefer and Wolfowitz (1956), Pfanzagl (1988) and van der Vaart and Wellner (1992) for consistency of maximum likelihood estimators do not apply here, even assuming identifiability. Aside from the fact that their models do not include covariates, the most important reason is that we do not have continuity of the kernel density on a compact (or compactified) parameter space. In other words, we cannot compactify $R^{r\times s}$ so that $\Pr(Y = y|w, \alpha, b)$ is continuous in (α, b) on the resulting compact space. We circumvent this problem by first demonstrating a type of convergence for the marginal distribution, which is used together with our identifiability conditions to prove consistency.

3. Marginal convergence. In this section we give some conditions that imply two types of convergence for the marginal distribution, given that we have m.l.e.'s for G and α for each n. The first type is used in the proof of consistency. The second type is included for its immediate implications. These results and some of the conditions are presented in Theorem 3.1. The remaining conditions, which restrict S and the covariate selection process, are given afterwards. It is important to note that the conditions for marginal convergence do not depend on $\mathcal{Q}$. Let $(S_y, S) \equiv \{(y, w)\colon\ y \in S_y,\ w \in S\}$. Then we define η_n and μ to be the measures on the Borel sets in (S_y, S) such that $d\eta_n(y, w) = M(y|w, \alpha_n, G_n)\, d\rho(w)$ and $d\mu(y, w) = M(y|w, \alpha, G)\, d\rho(w)$, in

the obvious manner. Then we have the following result, proved in Appendix A.

THEOREM 3.1. *For each n, let G_n and α_n be maximum likelihood estimators for G and α. Suppose that $\rho^{(n)} \to_{\mathscr{D}} \rho$ with probability 1 (P), that F is strictly increasing on R^m and that we have the additional conditions given below. Then we have the following conclusions*:

1. *With probability* 1 (P), *for any subsequence* $\{n(j), j = 1, 2, \dots\}$ *there exists a further subsequence* $\{n(j(h)), h = 1, 2, \dots\}$ *such that*

$$M\big(y|w, \alpha_{n(j(h))}, G_{n(j(h))}\big) \to M(y|w, \alpha, G) \quad \text{as } h \to \infty$$

   for all y and for values of w almost everywhere (ρ).
2. *With probability* 1 (P), $\eta_n \to_{\mathscr{D}} \mu$ *as* $n \to \infty$.

For any ρ-continuity set $B \in \mathscr{S}$, conclusion 2 implies convergence of the expected marginal probabilities for Y over B according to ρ, in the sense that

$$\frac{1}{\rho(B)} \int_B M(y|W, \alpha_n, G_n)\, d\rho \to \frac{1}{\rho(B)} \int_B M(y|W, \alpha, G)\, d\rho$$

as $n \to \infty$ for all y.

Note. For the conditional independence model, conclusions 1 and 2 of Theorem 3.1 imply the corresponding conclusions for the marginal distribution of any $m' \le m$ elements of Y, conditional on the corresponding m' rows of W, if we replace ρ with the appropriate marginal probability measure.

Additional conditions. Roughly speaking, we allow discrete and continuous columns of W, we assume that the discrete columns have a finite number of possible values and we assume that the continuous columns are random with an absolutely continuous joint distribution. Denote the columns of W as $(W_1, W_2, \dots, W_{s+r})$. For some $t \le s + r$, let $W_{(1)}$ be a collection of t columns of W, so that $W_{(1)} = (W_{i_1}, W_{i_2}, \dots, W_{i_t})$ for some fixed $(i_1, i_2, \dots, i_t)$. The remaining $s + r - t$ columns of W are denoted $W_{(2)}$. We suppose that $W_{(2)}$ is restricted to a finite set $S_{(2)} = \{w_{(2),k}, k = 1, 2, \dots, K\} \subset R^{m \times (s+r-t)}$. Let $S = \{w \in R^{m \times (s+r)}: w_{(1)} \in R^{m \times t}, w_{(2)} \in S_{(2)}\}$.

For each k, let ρ_k be the measure on the Borel sets $\mathscr{B}^{m \times t}$ in $R^{m \times t}$ such that, if $D \in \mathscr{B}^{m \times t}$, then $\rho_k(D) = \rho(\{w \in S: w_{(1)} \in D, w_{(2)} = w_{(2),k}\})$. Let $q_k = \rho_k(R^{m \times t})$, and suppose that $q_k > 0$ for all k. Since $\rho^{(n)}$ converges in measure to ρ, then

$$\frac{1}{n} \sum_{i=1}^{n} I\{W_{i,(2)} = w_{(2),k}\} \to q_k \tag{3.1}$$

for each k, with probability 1 (P), where $I\{\cdot\}$ is equal to 1 if the argument is true, and equal to 0 if it is false. We suppose that the conditional random variables $W_{i(1)}|(W_{i(2)} = w_{(2),k})$ are i.i.d. according to the probability measure

ρ_k/q_k and that ρ_k is absolutely continuous with respect to the Lebesgue measure on $R^{m\times t}$.

4. Identifiability and consistency. In this section we assume the general model (2.1) and that some sequence $\{G_n \in \mathcal{Q},\ \alpha_n \in R^s,\ n = 1, 2, \ldots\}$ satisfies conclusion 1 of Theorem 3.1. Then we give some conditions that are sufficient for identifiability of G and α and that are also sufficient for consistency of G_n and α_n.

First, we need the following definitions. Let $\alpha = (\alpha_{(1)}, \alpha'_{(2)})'$ for a scalar $\alpha_{(1)}$ and an $(s-1)\times 1$ vector $\alpha_{(2)}$. Let $\alpha_{(2),j}$ be the jth element of $\alpha_{(2)}$. Similarly, let $\alpha_n = (\alpha_{n,(1)}, \alpha'_{n,(2)})'$ and let $\alpha_{n,(2),j}$ be the jth element of $\alpha_{n,(2)}$. Let $X = (X_1, X_2)$ for an $m\times 1$ vector X_1 and an $m\times(s-1)$ matrix X_2. Then $W = (X_1, X_2, Z)$. Similarly, denote any $w \in R^{m\times(s+r)}$ as $w = (x_1, x_2, z)$.

Recall that ρ is a probability measure on $\mathcal{S}$. Acting as if W were distributed according to ρ, let $\rho_{X(2),Z}$ be the marginal distribution of ρ for (X_2, Z), and let $\rho_{X_1|(x_2,z)}$ be the conditional distribution (where this exists) for X_1 given that $(X_2, Z) = (x_2, z)$. Let $S_2 \subset R^{m\times(s+r-1)}$ be a support for the marginal of ρ on (x_2, z). Let $X_{2,j}$ be the jth column of X_2, and let $X_{2,(j)}$ be the first j columns of X_2, $j = 1, 2, \ldots, s-1$. Let $(X_{2,(0)}, Z) = Z$.

For each $(x_2, z) \in S_2$, $\alpha \in R^s$ and $G \in \mathcal{Q}$, if $\alpha_{(1)} \neq 0$, then define $H_{\alpha,G}$ to be the c.d.f. for $(-1/\alpha_{(1)})(x_2\alpha_{(2)} + zb + \varepsilon)$. Let $\Gamma(x_2, z)$ be a class of c.d.f.'s on R^m that contains $\{H_{\alpha,G}\colon\ \alpha \in R^s,\ \alpha_{(1)} \neq 0,\ G \in \mathcal{Q}\}$. For example, if $\mathcal{Q}$ is the class of all distributions on $\mathcal{R}^r$, then we can choose $\Gamma(x_2, z)$ to be the class of all distributions on $\mathcal{R}^m$, although a smaller class is possible when $r < m$. For another example, if $\mathcal{Q}$ is the class of Gaussian distributions on $\mathcal{R}^r$ and F is Gaussian, then we can take $\Gamma(x_2, z)$ to be the class of all Gaussian distributions on R^m.

We say that the members of a class Γ of c.d.f.'s on R^m are uniquely determined by their values on a set $C \subset R^m$ if the following implication holds: when $H \in \Gamma$, $G \in \Gamma$ and $H = G$ on C, then $H = G$ on R^m. For example, if Γ is the class of all distributions on $\mathcal{R}^m$, then the smallest such C would be R^m. If Γ is a parametric class such as the class of Gaussian distributions on R^m, then a much smaller C would suffice. The following theorem is proved in Appendix B.

THEOREM 4.1. *Let $G' \in \mathcal{Q}$ and $\alpha' \in R^s$, and suppose that $M(\mathbf{1}|W, \alpha', G') = M(\mathbf{1}|W, \alpha, G)$ almost everywhere (ρ). Suppose that conclusion 1 of Theorem 3.1 holds for some sequence $\{G_n \in \mathcal{Q},\ \alpha_n \in R^s,\ n = 1, 2, \ldots\}$ and that we have the following conditions*:

1. *For any $q\times m$ matrix T of rank $q \le m$, the characteristic function of $T\varepsilon$ is nonzero on some set of points dense in R^q.*
2. *$s \ge 1$ and $\alpha_{(1)} \neq 0$.*
3. *There exists a collection of sets $\{A_l \subset S_2,\ l = 1, 2, \ldots\}$ with $\rho_{X_2,Z}(A_l) > 0$ for all l, such that, if $(x_2, z) \in \bigcup_l A_l$, then there exists a countable or open set $C(x_2, z) \subset R^m$ with the following properties. Let $x_1 \in C(x_2, z)$. Then*

any open neighborhood $N(x_1)$ *of* x_1 *satisfies* $\rho_{X_1|(x_2,z)}(N(x_1) \cap C(x_2,z)) > 0$, *the members of* $\Gamma(x_2,z)$ *are uniquely determined by their values on* $C(x_2,z)$ *and there exist some* ν^* *and* ν *in* $C(x_2,z)$ *such that* $\nu_j^* > \nu_j$ *for some* j *and* $\nu_i^* \geq \nu_i$ *for all* i, *where* ν_i *and* ν_i^* *are the* ith *elements of* ν *and* ν^*, *respectively*.

Then we have the following results, which depend on the properties of individual sets in the collection $\{A_l \subset S_2,\ l = 1,2,\ldots\}$.

4. *For some* $l \in \{1,2,\ldots,L\}$, *suppose that each* $(x_2,z) \in A_l$ *has* $\operatorname{rank}(x_2, z) < m$.

Then $\alpha_{(1)}^* = \alpha_{(1)}$ *and* $\alpha_{(1),n} \to \alpha_{(1)}$.

5. *Suppose that* $\alpha_{(1)}^* = \alpha_{(1)}$ *and* $\alpha_{(1),n} \to \alpha_{(1)}$. *For* $j \in \{1,2,\ldots,s-1\}$ *and for some* $l(j) \in \{1,2,\ldots\}$, *suppose that we have either* (a) *or* (b) *below*:
   (a) *For each* $(x_2,z) \in A_{l(j)}$, $\operatorname{rank}(x_{2,(j-1)}, z) < m$ *and* $\operatorname{rank}(x_{2,(j)}, z) = \operatorname{rank}(x_{2,(j-1)}, z) + 1$.
   (b) *For some* $k(j) \in \{1,2,\ldots,L\}$, *for every* $(x_2,z) \in A_{l(j)}$ *and* $(x_2^*, z^*) \in A_{k(j)}$, *we have* $(x_{2,(j-1)}, z) = (x_{2,(j-1)}^*, z^*)$ *and* $(x_{2,j}, z) \neq (x_{2,j}^*, z^*)$.

Then $\alpha_{(2),j}^* = \alpha_{(2),j}$ *and* $\alpha_{n,(2),j} \to \alpha_{(2),j}$.

6. *Suppose that* $\alpha^* = \alpha$ *and* $\alpha_n \to \alpha$. *Suppose that, for some* $l \in \{1,2,\ldots,L\}$, *each* $(x_2,z) \in A_l$ *has* $\operatorname{rank}(z) = r$.

Then $G^* = G$ *and* $G_n \to_{\mathscr{D}} G$.

4.1. *Discussion and special cases.* In this section we discuss some implications of these conditions, and we consider the cases of discrete covariates, general mixing and Gaussian mixing. Much of the discussion centers on how the key condition 3 depends on Q and on the role of condition 3 in the proof. We discuss each condition in turn, after a general note on discrete covariates.

Discrete covariates. The conditions of Theorem 4.1 are simpler when some or all columns of (X_2, Z) contain discrete covariates. For example, if all columns of (X_2, Z) are discrete, then it may be possible to choose A_l to be a single element $(x_2^l, z_l) \in S_2$ for each l. If only some columns of (X_2, Z) are discrete, then it may be possible for all elements of A_l to have the same values in the discrete columns for each l. This can cause conditions 3 through 6 to be easier to satisfy, as discussed below and illustrated in Section 5.

Conditions 1 *through* 6. Condition 1 is very weak, and it holds for any common choice of F. For example, this condition is satisfied by the multivariate Gaussian distribution.

Conditions 2 and 3 compensate for the loss of information in observing Y rather than T. Condition 3 depends strongly on the choice of Q. For example, suppose that we let Q be the class of all distributions on $\mathscr{R}^r$ and that, for each $(x_2,z) \in A_l$, we take $\Gamma(x_2,z)$ to be the class of all distributions on $\mathscr{R}^m$.

Then condition 3 requires that $C(x_2, z) = R^m$. This holds, for example, when $X_1|((X_2, Z) = (x_2, z))$ has a continuous distribution with a positive density on R^m. This strong condition is used, in our proof, to uniquely identify $H_{\alpha,G}$ in $\Gamma(x_2, z)$ given the values of $H_{\alpha,G}$ on $C(x_2, z)$. For an illustrative example, consider the case $m = 1$, $r = 1$, $s = 1$, $\alpha_{(1)} = 1$, $Z \equiv 1$ and a general Q, so that $(-1/\alpha_{(1)})(x_2\alpha_{(2)} + zb + \varepsilon) = -(b + \varepsilon)$. Then $H_{\alpha,G}$ is simply the marginal c.d.f. for a location mixture with a general mixing distribution. Then $\Gamma(x_2, z)$ must be the class of all distributions on $\mathscr{R}^1$, and $C(x_2, z)$ must be equal to R^1. In fact, our condition 1 is the same as the condition given by Teicher (1961) and Maritz and Lwin (1989) for identifiability of the mixing distribution, given the marginal distribution for a general location mixture.

If we restrict Q to a parametric family, then condition 3 clearly becomes much weaker because we can choose a much smaller class $\Gamma(x_2, z)$. For example, suppose that Q is the class of Gaussian distributions and F is a Gaussian c.d.f., so that we can take $\Gamma(x_2, z)$ to be the class of all Gaussian distributions on R^m. Then condition 3 requires that the Gaussian c.d.f.'s on R^m are uniquely determined by their values on $C(x_2, z)$. An example of this is discussed in Section 5.

Condition 4 identifies $\alpha_{(1)}$. This condition will always hold if $m > s + r - 1$. However, for obvious reasons this can hold for smaller values of m when some covariates are discrete or when some covariates are constant within individuals.

Condition 5 identifies $\alpha_{(2)}$. Option (a) implies that $m \geq r + 1$. This option holds, for example, when $m > s + r - 1$ and $\text{rank}(x_2, z) = s + r - 1$ for every (x_2, z) in A_l. Option (a) can be satisfied with a much smaller m when some covariates are discrete. Option (b) requires for some $l(j)$ and $k(j)$, that $(x_{2,(j-1)}, z)$ be constant over $A_{l(j)}$ and that $(x^*_{2,(j-1)}, z^*)$ be constant over $A_{k(j)}$. This is only reasonable when $(X_{2,(j-1)}, Z)$ are discrete covariates. Compared to option (a), this option can allow much smaller values of m and more variables to be constant within individuals.

Condition 6 identifies G once the fixed effects are identified. This condition implies that $r \leq m$, so that the number of repeated observations is at least as large as the number of random effects.

Other practical notes. As illustrated in Section 5, for different elements of α we may rearrange the columns in W to use different columns as X_1. Also, following the note after Theorem 3.1, for the conditional independence model it is sufficient that the conditions hold for the marginal of ρ on any subset of m' rows of w. In fact, we may choose different subsets of m' rows for each l of conditions 4, 5 and 6.

5. An application. We now illustrate the conditions by applying them to a particular situation. This is a doubly nested study of interviewer variability in a binary response, with covariates. Anderson and Aitken (1985) assume a conditional independence model, a probit link function and Gaussian mixing.

They analyze the responses on a binary item from a consumer attitude survey carried out by the Social and Community Planning Research (SCPR) Institute. Interviews were conducted by 64 interviewers, two at each of 32 locations, with a total of 1265 respondents (all heads of households). The average number of respondents per interviewer was 19.8. Due to the large number of locations and interviewers, Anderson and Aitken argue that it is appropriate to model these as having additive random effects, which they assume to be independent and Gaussian. The covariates used by Anderson and Aitken include six categories for interviewer age, five for interviewer marital status and two for interviewer experience. Respondent age and marital status were also included, with similar categories.

We now discuss the application of our conditions to this situation. First, we suppose that the model assumed by Anderson and Aitken is used. We refer to this as case A. We must assume a slightly different use of the available data because we have not considered the case of unequal numbers of repeated measures. We suppose that only those locations in which each interviewer has 10 or more interviews are included and that only the first 10 interviews are used for each interviewer at these locations. The number 10 is arbitrarily chosen, and our discussion would be similar for any other choice. Otherwise, we assume the same data used by Anderson and Aitken. In case B we assume general nonparametric mixing, and we must make additional changes discussed below.

Let Y_i be the 20×1 vector of binary responses for the ith location, listed so that the first 10 elements of Y_i are responses to "interviewer 1" and the last 10 elements are responses to "interviewer 2." (The labels "1" and "2" are arbitrarily assigned to the interviewers at each location.) Let W_i be the corresponding 20×21 design matrix for the ith location, organized as follows: let X_i be the first 19 columns of W_i, which contains 19 dummy variables for the five categorical covariates. Dropping the subscript "i," let the first nine columns of X contain the dummy variables for respondent characteristics, out of which the first five columns contain the dummy variables for respondent age. Then the last 10 columns of X contain the dummy variables for interviewer characteristics (notice that these are constant over the first 10 rows and over the last 10 rows). Let Z_i be the last two columns of W_i. Let the first column of Z_i be $\mathbf{1}_{20\times 1}$, defined as a 20×1 vector of 1's, and let the second column of Z_i be $(\mathbf{1}'_{10\times 1}, \mathbf{0}'_{10\times 1})'$, or the 20×1 vector in which the first 10 elements are 1's and the last 10 elements are 0's.

Then α is a 19×1 vector of fixed effects, and b is a 2×1 vector of random effects. The first element of b is the sum of the location effect and the interviewer 1 effect, while the second element of b is the interviewer 2 effect. Then in case A, following Anderson and Aitken, we take Q to be the class of independent bivariate Gaussian distributions. In case B we take Q to be the class of all distributions on $\mathscr{R}^2$. In case B we suppose that "respondent age" is used directly as a continuous variable and that this is the covariate in the column X_1. Then the first five columns of X contain respondent characteristics, and the total number of columns in X is 15.

We suppose that the number of locations n approaches ∞ and that the locations, interviewers and respondents are selected independently and at random. Then the (X_i, Z_i) are i.i.d. according to a measure ρ. We suppose that every combination of the possible values for the categorical covariates will be observed with some positive probability.

First, we discuss the conditions of Theorem 3.1, which hold in both cases A and B. Recall that these conditions are independent of Q. The link function F is the c.d.f. for the m-dimensional Gaussian distribution with the identity covariance matrix, which is continuous, increasing and never equal to 0 or 1, as required. In case A let $W = W_{(2)} = (X, Z)$, and let $S_{(2)}$ be the finite set of possible values for (X, Z). For each $(x, z) \in S_{(2)}$, the proportion of observations with $(X, Z) = (x, z)$ converges almost surely to a positive constant, due to i.i.d. sampling. Therefore, Theorem 3.1 holds for this case. In case B let $W_{(1)} = X_1$ and $W_{(2)} = (X_2, Z)$. Let $S_{(2)}$ be the finite set of possible values for (X_2, Z). For each $(x_2, z) \in S_{(2)}$, the proportion of observations with $(X_2, Z) = (x_2, z)$ converges almost surely to a positive constant. If the conditional distribution $X_1|(X_2, Z) = (x_2, z)$ is continuous on R^m for each (x_2, z), then Theorem 3.1 holds in this case as well.

We now check the conditions of Theorem 4.1. Notice that condition 1 holds because F is Gaussian. Let S_2 be the finite set of possible values for (X_2, Z). First, consider case A, and denote the respondent age category associated with X_1 as category 1. The first four columns of X contain dummy variables for respondent age categories other than category 1, and we denote these as categories 2, 3, 4 and 5. Category 6 is the reference category. Condition 2 requires that the true effect associated with category 1, relative to the reference category, is not equal to 0. This condition seems technically moot, since the true effect would never be exactly 0. However, a large absolute effect size could reduce the range of the distributions in $\Gamma(x_2, z)$ for each (x_2, z), which could help in satisfying condition 3.

The relationship between X_1 and the first four columns of X_2 necessitates the following approach. We begin by considering only those values of X for which all the respondent ages are either in category 1 or the reference category 6. In other words, let S_2' be the subset of S_2 consisting of all members (x_2, z) of S_2 for which $x_{2,(4)}$ is a 20×4 matrix of 0's, or $x_{2,(4)} = \mathbf{0}_{20\times 4}$. We take each A_l to consist of one point $(x_2, z) \in S_2'$, each of which has positive $\rho_{x(2),z}$ measure. For each $A_l \equiv \{(x_2, z)\} \subset S_2'$, we take $\Gamma(x_2, z)$ of condition 3 to be the class of all Gaussian distributions on R^m. Then condition 3 holds because $C(x_2, z)$ includes every $m \times 1$ vector with elements equal to 0 or 1 and the Gaussian c.d.f.'s on R^m are uniquely determined by their values on these points. We choose $\nu^* = \mathbf{1}_{m\times 1}$ and $\nu^* = \mathbf{0}_{m\times 1}$.

Condition 4 holds for each $(x_2, z) \in S_2'$, as follows. Recall that x_2 has 18 columns, and the last 10 contain interviewer characteristics. Then the last 10 columns must be linear combinations of the columns of Z. Since $x_{2,(4)} = \mathbf{0}_{20\times 4}$, then $\text{rank}(x_2, z) \leq 7$. In fact, for particular choices of $(x_2, z) \in S_2'$, the rank of (x_2, z) can be as low as 2, for example, when all elements of x_2 are equal to 0. Therefore, $\alpha_{(1)}^* = \alpha_{(1)}$ and $\alpha_{(1),n} \to \alpha_{(1)}$. Now, in order to demonstrate

identifiability and consistency for the effects associated with respondent age categories 2, 3, 4 and 5, we can exchange each of these categories with category 1 and repeat the above discussion.

We use option (b) of condition 5 for the last 14 elements of α. For any (x_2, z) and (x_2^*, z^*) in S_2', we have that $z = z^*$ and $x_{2,(4)} = x_{2,(4)}^*$, so that for each $j = 5, 6, \ldots, 18$ we must find some (x_2, z) and (x_2^*, z^*) in S_2' such that $x_{2,(j-1)} = x_{2,(j-1)}^*$ and $x_{2,j} \neq x_{2,j}^*$. There are many such choices. For example, for each j we may take $x_{2,(j-1)} = x_{2,(j-1)}^* = \mathbf{0}_{20\times(j-1)}$, $x_{2,j} = \mathbf{1}_{20\times 1}$, $x_{2,j}^* = \mathbf{0}_{20\times 1}$ and $x_{2,k} = x_{2,k}^* = \mathbf{0}_{20\times 1}$ for $k = j+1, j+2, \ldots, 18$, where $\mathbf{1}$ is a vector with unit elements. Finally, condition 6 holds for every $(x_2, z) \in S_2'$.

Now consider case B. Condition 2 holds if the linear effect of age is not 0. We take each A_l to consist of one point (x_2, z) in S_2, each point of which has positive $\rho_{x(2),z}$ measure. For each $A_l \equiv \{(x_2, z)\} \subset S_2$, we take $\Gamma(x_2, z)$ of condition 3 to be the class of all distributions on $\mathscr{R}^m$. Then condition 3 will hold for each (x_2, z) such that $\rho_{X_1|(x_2,z)}$ has a positive density on R^m, so that $C(x_2, z) = R^m$. This is technically impossible, since the range of respondent ages is finite. Some further limitations on the classes $\Gamma(x_2, z)$ are therefore necessary, in practice, so that the range of X_1 will be "large enough." However, it is clear that a continuous "respondent age" allows a larger class Q than does a categorical "respondent age." If condition 3 holds for some Q, then condition 4 holds because $\text{rank}(x_2, z) \leq 7$, as before. Then option (b) of condition 5 holds due to arguments similar to (but simpler than) those used for case A, and condition 6 holds for any (x_2, z).

6. Discrete mixtures. We now discuss some practical implications of the general results due to Laird (1978) and Lindsay (1983) on the existence of discrete nonparametric maximum likelihood estimators. Let Q be the set of all distributions on the Borel sets in R^r such that each distribution has a compact support. Let N be the number of distinct observations of (Y, W), and let Q^N be the class of discrete distributions in Q with N mass points. Suppose that G_n and α_n are maximum likelihood estimators (m.l.e.'s) for G over Q^N and for α over R^s. Then we can show that these also maximize the marginal likelihood over Q and R^s, as follows. Let $B \subset R^r$ and $C \subset R^s$ be compact sets. Define Q^B to be the class of distributions in Q that have support B. Then the conditions of Lindsay (1983) guarantee the existence of m.l.e.'s G_n and α_n over Q^B and C such that G_n has at most N mass points. But this implies that G_n and α_n are m.l.e.'s over Q^B and C for *any* compact B and C, and the claim follows.

APPENDIX A

Proof of Theorem 3.1. For this proof we combine α and b so that b is an $(s + r) \times 1$ random vector in which the first s elements are degenerate. Redefine M, G and Q accordingly. Define μ as in Section 3, and let $\mu^{(n)}$ be the empirical c.d.f. on the Borel sets in (S_y, S) generated by $\{(Y_i, W_i),\ i =$

$1, 2, \ldots, n\}$. We now present some inequalities upon which the proof is founded. First, we have a standard inequality based on the nonnegativity of the Kullback–Liebler distance. For all $(y, w) \in (S_y, S)$,

$$\text{(A.1)} \quad \sum_y \left[M(y|w,G) \ln\left(M(y|w,G) \right) \right] \geq \sum_y \left[M(y|w,G) \ln\left(M(y|w,G_n) \right) \right].$$

Therefore, for any Borel set $H \subset S$,

$$\text{(A.2)} \quad E_\mu \left[\ln\left(M(Y|W,G) \right) I\{W \in H\} \right] \geq E_\mu \left[\ln\left(M(Y|W,G_n) \right) I\{W \in H\} \right].$$

Because $\ln M(Y|W,G_n) \geq 0$, for any Borel set $H \subset S$ we have

$$\text{(A.3)} \quad E_{\mu^{(n)}} \left[\ln\left(M(Y|W,G_n) \right) I\{W \in H\} \right] \geq E_{\mu^{(n)}} \left[\ln\left(M(Y|W,G_n) \right) \right],$$

where $I\{\cdot\}$ is equal to 1 if the argument is true, and 0 otherwise. Finally, from the definition of G_n,

$$\text{(A.4)} \quad E_{\mu^{(n)}} \left[\ln\left(M(Y|W,G_n) \right) \right] \geq E_{\mu^{(n)}} \left[\ln\left(M(Y|W,G) \right) \right].$$

The conditions of Theorem 3.1 imply that

$$\text{(A.5)} \quad E_{\mu^{(n)}} \left[\ln\left(M(Y|W,G) \right) \right] \to E_\mu \left[\ln\left(M(Y|W,G) \right) \right],$$

with probability 1 (P) (hereafter "w.p.1"). This follows from (3.1) and the SLLN, which together imply for each k that

$$\begin{aligned} &\frac{1}{n} \sum_i \ln\left(M(Y_i|W_i,G) \right) I\{W_{i,(2)} = w_{(2),k}\} \\ &\quad \to q_k E_{\rho_k/q_k} \left[\sum_y \ln\left(M(y|W,G) \right) M\left(y|W,G,W_{(2)} = w_{(2),k} \right) \right] \\ &\quad = q_k E_\mu \left[\ln\left(M(Y|W,G) \right) | W_{(2)} = w_{(2),k} \right] \end{aligned}$$

as $n \to \infty$, w.p.1, where the expectation E_{ρ_k/q_k} is over $W_{(1)}$ according to ρ_k/q_k and where we define $M(y|W,G,W_{(2)} = w_{(2),k}) = P(Y = y|W,G,W_{(2)} = w_{(2),k})$. Finally, we may take the sum of each side over k in the above expression, and the convergence holds because the sum is finite. This gives (A.5).

We now give two lemmas. Together with (A.5), these lemmas will allow us to link inequalities (A.1) to (A.4) as outlined later in (A.7) and (A.8).

LEMMA A.1. *For any subsequence* $\{n(j), j = 1, 2, \ldots\}$, *there exists a further subsequence* $\{n(j(h)), h = 1, 2, \ldots\}$ *and a function* $M^*(y, w)$ *on* (S_y, S) *such that, for every* y *and for* w *almost everywhere* ρ *on* S,

$$\text{(A.6)} \quad M\left(y|w,G_{n(j(h))} \right) \to M^*(y,w) \quad \text{as } h \to \infty.$$

PROOF. Let $b_{(1)}$ be the $t \times 1$ random vector consisting of the elements of b that are associated with the columns of $W_{(1)}$. Let $\text{sign}(b_{(1)})$ be the $t \times 1$ random vector with lth element $\text{sign}(b_{(1)})_l$ defined as follows: let $\text{sign}(b_{(1)})_l = 1$ if $b_{(1)l} \geq 0$, and let $\text{sign}(b_{(1)})_l = -1$ if $b_{(1)l} < 0$ for $l = 1, 2, \ldots, t$. Let $\boldsymbol{\nu}$ be the random diagonal matrix with $\text{sign}(b_{(1)})$ on the diagonal, or, in other words,

with $\text{diag}(\boldsymbol{\nu}) = \text{sign}(b_{(1)})$. We denote any particular value of $\boldsymbol{\nu}$ as ν. Also, for each $y \in S_y$, let $\text{sign}(y)$ be the $m \times 1$ random vector with lth element $\text{sign}(y)_l$ defined as follows: let $\text{sign}(y)_l = 1$ if $y_l = 1$, and let $\text{sign}(y)_l = -1$ if $y_l = 0$ for $l = 1, 2, \ldots, m$. Let $\tau(y)$ be the diagonal matrix with diagonal $\text{diag}(\tau(y)) = \text{sign}(y)$.

To simplify the notation, we define $G'_j = G_{n(j)}$ in the proof of this lemma. For each k, we make the following definitions for all w in S such that $w_{(2)} = w_{(2),k}$:

Let $M_{j,k}(y, w_{(1)}) = M(y|w, G'_j)$.

For each ν, let $M_{j,k,\nu}(y, w_{(1)}) = P(Y = y|W = w, G'_j, \text{sign}(b_{(1)}) = \text{diag}(\nu))$, and let $p_{j,\nu} = P(\text{sign}(b_{(1)}) = \text{diag}(\nu)|G'_j)$, so that $M_{j,k}(y, w_{(1)}) = \sum_\nu p_{j,\nu} M_{j,k,\nu}(y, w_{(1)})$.

Then $M_{j,k,\nu}(y, \tau(y)\gamma\nu)$ is nondecreasing in γ on $R^{m\times t}$, which follows directly from the definition of $M(y|w, G'_j)$.

This allows us to proceed exactly as, for example, in the proof of Theorem 25.9 in Billingsley (1986), which does not require that $M_{j,k,\nu}(y, \tau(y)\gamma\nu))$ be a c.d.f. in γ, to construct a function $M^\dagger_{k,\nu}(y, \gamma)$ with the following properties:

1. For a subsequence $j(h)$, we have $M_{j(h),k,\nu}(y, \tau(y)\gamma\nu) \to M^\dagger_{k,\nu}(y, \gamma)$ on continuity points of $M^*_{k,\nu}(y, \tau(y)\gamma\nu)$ in $R^{m\times t}$ as $h \to \infty$.
2. The function $M^\dagger_{k,\nu}(y, \gamma)$ is right continuous and nondecreasing on $\gamma \in U_{y,\nu}$.

By choosing successive subsequences on the finite set of values for y, k and ν, this construction may be done so that the convergence holds for all y, k and ν on the same subsequence. Also, by a successive construction using Helly's selection theorem for a tight sequence of measures, we may construct p^*_ν with $\sum_\nu p^*_\nu = 1$ such that $p_{j(h),\nu} \to p^*_\nu$ for all ν.

Notice that the transformation $\gamma \to \tau(y)\gamma\nu$ from $R^{m\times t}$ to $R^{m\times t}$ is invertible for any y and ν, so that we can define $M^*_{k,\nu}(y, \tau(y)\gamma\nu) = M^\dagger_{k,\nu}(y, \gamma)$. Then property 1 implies that $M_{j(h),k,\nu}(y, w_{(1)}) \to M^*_{k,\nu}(y, w_{(1)})$ on continuity points of $M^*_{k,\nu}(y, w_{(1)})$ in $R^{m\times t}$. Also, property 2 implies that $M^*_{k,\nu}(y, \tau(y)\gamma\nu)$ is continuous for γ almost everywhere (λ), where λ is the Lebesgue measure on $R^{m\times t}$. [*Note:* The set of discontinuities in $M^*_{k,\nu}(y, \tau(y)\gamma\nu)$ is measurable, as discussed for a more general class of functions in Billingsley (1986), pages 343 and 391. It is straightforward to show that the Lebesgue measure of this set must be 0, since $M^*_{k,\nu}(y, \tau(y)\gamma\nu)$ is nondecreasing in γ.] Therefore, $M^*_{k,\nu}(y, w_{(1)})$ is continuous for $w_{(1)}$ almost everywhere (λ) on $R^{m\times t}$.

Finally, then, for all y and k, we have that $\sum_\nu p_{j(h),\nu} M_{j(h),k,\nu}(y, w_{(1)}) \to \sum_\nu p^*_\nu M^*_{k,\nu}(y, w_{(1)})$ for $w_{(1)}$ almost everywhere (λ). This convergence holds almost everywhere (ρ_k) on $R^{m\times t}$, since ρ_k is absolutely continuous with respect to λ. This gives result (A.6) with $M^*(y, w) = \sum_\nu p^*_\nu M^*_{k,\nu}(y, w_{(1)})$ for all $w \in S$ such that $w_{(2)} = w_{(2),k}$, $k = 1, 2, \ldots, K$. [The fact that this convergence holds for $w_{(1)}$ almost everywhere (λ) will be used later in the proof of this theorem.] □

Denote the mt elements of any $w_{(1)} \in R^{m\times t}$ as $e_1, e_2, e_3, \ldots, e_{mt}$ (according to any arbitrary one-to-one allocation scheme). Let B be a positive integer, and let $C \subset R^{m\times t}$ be the half open hypercube $C = \{w_{(1)} \in R^{m\times t}: e_j \in [-B, B), j = 1, 2, \ldots, mt\}$. Let A be a positive integer, and partition C into disjoint half open cells of width $1/A$ in each dimension, as follows. Let $(l_1, l_2, l_3, \ldots, l_{mt})$ be a list of mt integers between 1 and $2BA$. In other words, let $l_j \in \{1, 2, \ldots, 2BA\}$ for $j = 1, 2, \ldots, mt$. Then we define the cell

$$I(l_1, l_2, l_3, \ldots, l_{mt}) = \Big\{w_{(1)} \in R^{m\times t}: e_j \in \big[-B + (l_j - 1)/A, -B + l_j/A\big), \quad j = 1, 2, \ldots, mt\Big\}.$$

There are $(2BA)^{mt}$ such cells, and, in order to reduce the notation, we reindex them as follows. With each integer $\psi \in \{1, 2, \ldots, (2BA)^{mt}\}$ we associate a unique list $(l_1^\psi, l_2^\psi, l_3^\psi, \ldots, l_{mt}^\psi)$ with $l_j^\psi \in \{1, 2, \ldots, 2BA\}$ for $j = 1, 2, \ldots, mt$. (This may be done according to any arbitrary one-to-one allocation scheme). Then we define $I_\psi = I(l_1^\psi, l_2^\psi, l_3^\psi, \ldots, l_{mt}^\psi)$. Define the collection $\mathscr{I} = \{I_\psi,\ \psi = 1, 2, \ldots, (2BA)^{mt}\}$. Then C is the union of all the cells in $\mathscr{I}$. For each y, k and ψ, define

$$v_n(y, k, \psi) = \sup\Big\{\big|M(y|w, G_n) - M(y|\zeta, G_n)\big|: w_{(1)} \in I_\psi,\ \zeta_{(1)} \in I_\psi \text{ and } w_{(2)} = \zeta_{(2)} = w_{(2),k}\Big\},$$

where $\zeta_{(1)}$ and $\zeta_{(2)}$ are defined for ζ as $w_{(1)}$ and $w_{(2)}$ are defined for w. Similarly, define $v^*(y, k, \psi)$ as above but with $M^*(y, w)$ in place of $M(y|w, G_n)$. Then we will prove the following lemma.

LEMMA A.2. *For each y and k, we have that $(1/(2BA)^{mt})\Sigma_\psi[v_n(y, k, \psi)] \le mt/(2BA)$ and that $(1/(2BA)^{mt})\Sigma_\psi[v^*(y, k, \psi)] \le mt/(2BA)$.*

Before the proof, we first note the following consequences. Let $\delta > 0$, and let $\mathscr{I}_n$ be the set of cells $I_\psi \in \mathscr{I}$ such that both $v_n(y, k, \psi) < \delta$ and $v^*(y, k, \psi) < \delta$ for all y and k (simultaneously). Let $C_{n,\delta} \subset C$ be the union of the cells in $\mathscr{I}_n$. For each y and k, Lemma A.2 provides the upper bound $mt/(2BA)$ for the average of the $v_n(y, k, \psi)$ over all ψ. This implies that the proportion of the cells in $\mathscr{I}$ that have $v_n(y, k, \psi) \ge \delta$ is less than or equal to $mt/(\delta 2BA)$. Similarly, the proportion of the cells in $\mathscr{I}$ that have $v^*(y, k, \psi) \ge \delta$ is less than or equal to $mt/(\delta 2BA)$. Hence, the proportion of the cells in $\mathscr{I}$ that have either $v_n(y, k, \psi) \ge \delta$ or $v^*(y, k, \psi) \ge \delta$ for some y or k is less than or equal to $2^m K 2mt/(\delta 2BA) = 2^m Kmt/(\delta BA)$. This implies that $\lambda(C_{n,\delta})/\lambda(C) \le 2^m Kmt/(\delta BA)$, so that $\lambda(C) - \lambda(C_{n,\delta}) \ge \lambda(C)(1 - 2^m Kmt/(\delta BA))$. Notice that this bound is independent of n.

Therefore, for fixed values of B and δ, we have $\lambda(C_{n,\delta}) \to \lambda(C)$ uniformly over n as $A \to \infty$, where λ is the Lebesgue measure on $R^{m\times t}$. Since ρ_k is absolutely continuous with respect to λ on $R^{m\times t}$, this implies that $\rho_k(C_{n,\delta}) \to \rho_k(C)$ uniformly over n as $A \to \infty$. [This follows from a general result: for any $\tau > 0$, there exists $\varsigma > 0$ such that any Borel set $U \subset R^{m\times t}$

with $\lambda(U) < \varsigma$ also has $\rho_k(U) < \tau$. This result follows by contradiction: otherwise, for some $\tau > 0$ and every $\varsigma > 0$, there exists $U_\varsigma \subset R^{m\times t}$ such that $\lambda(U_\varsigma) < \varsigma$ and $\rho_k(U_\varsigma) \geq \tau$. Take $\varsigma(\eta) = 1/2^\eta$ for $\eta = 1, 2, \ldots,$ and let $D_q = \cup_{(\eta \geq q)} U_{\varsigma(\eta)}$. Then $\lambda(D_q) < 2^{-q}/(1 - 1/2) \to 0$ as $q \to \infty$, while $\rho_k(D_q) \geq \rho_k(U_{\varsigma(q)}) \geq \tau$. But D_q decreases with q to some limiting Borel set $D_\infty \subset S$ with $\lambda(D_\infty) = 0$. Therefore, $\rho_k(D_q) \to \rho_k(D_\infty) = 0$ as $q \to \infty$, giving a contradiction.]

Since K is finite, for any $\xi > 0$, we may choose B large enough so that $\rho_k(C)/q_k > 1 - \xi$ for all k. Also, for any $\delta > 0$, we may then choose A large enough so that $\rho_k(C)/q_k \geq \rho_k(C_{n,\delta})/q_k > 1 - \xi$ for all n and k. Let $T = \{w \in S\colon w_{(1)} \in C\}$, and let $T(n, \delta) = \{w \in S\colon w_{(1)} \in C_{n,\delta}\}$. Then, for A and B as above, we have that $\rho(T) \geq \rho(T(n, \delta)) > 1 - \xi$ for all n.

PROOF OF LEMMA A.2. Consider the same definitions as in the first paragraph of the proof of Lemma A.1. Then, for each k, we make the following definitions for all w in S such that $w_{(2)} = w_{(2),k}$: let $M_{n,k}(y, w_{(1)}) = M(y|w, G_n)$. For each ν, let $M_{n,k,\nu}(y, w_{(1)}) = P(Y = y|W = w, G_n, \text{sign}(b_{(1)}) = \text{diag}(\nu))$, and let $p_{n,\nu} = P(\text{sign}(b_{(1)}) = \text{diag}(\nu)|G_n)$, so that $M_{n,k}(y, w_{(1)}) = \sum_\nu p_{n,\nu} M_{n,k,\nu}(y, w_{(1)})$.

Then it follows that

$$
\begin{aligned}
&\sum_\psi v_n(y, k, \psi) \\
&\quad = \sum_\psi \sup\left\{\left|M_{n,k,\nu}(y, w_{(1)}) - M_{n,k,\nu}(y, \zeta_{(1)})\right|\colon w_{(1)} \in I_\psi,\ \zeta_{(1)} \in I_\psi\right\} \\
&\quad = \sup\left\{\sum_\psi \left(\left|M_{n,k}(y, w_{(1),\psi}) - M_{n,k}(y, \zeta_{(1),\psi})\right|\right)\colon \right. \\
&\qquad\qquad \left. w_{(1),\psi} \in I_\psi,\ \zeta_{(1),\psi} \in I_\psi\ \forall\, \psi\right\} \\
&\quad \leq \sup\left\{\sum_\psi \left(E_{\boldsymbol{\nu}}\left[\left|M_{n,k,\boldsymbol{\nu}}(y, w_{(1),\psi}) - M_{n,k,\boldsymbol{\nu}}(y, \zeta_{(1),\psi})\right|\right]\right)\colon \right. \\
&\qquad\qquad \left. w_{(1),\psi} \in I_\psi,\ \zeta_{(1),\psi} \in I_\psi\ \forall\, \psi\right\} \\
&\quad = \sup\left\{E_{\boldsymbol{\nu}}\left[\sum_\psi \left(\left|M_{n,k,\boldsymbol{\nu}}(y, w_{(1),\psi}) - M_{n,k,\boldsymbol{\nu}}(y, \zeta_{(1),\psi})\right|\right)\right]\colon \right. \\
&\qquad\qquad \left. w_{(1),\psi} \in I_\psi,\ \zeta_{(1),\psi} \in I_\psi\ \forall\, \psi\right\},
\end{aligned}
$$

where the expectation $E_{\boldsymbol{\nu}}$ is over the distribution for $\boldsymbol{\nu}$ according to the probabilities $p_{n,\nu}$. We will show that $\sum_\psi |M_{n,k,\nu}(y, w_{(1),\psi}) - M_{n,k,\nu}(y, \zeta_{(1),\psi})| \leq mt(2BA)^{mt-1}$ for every ν, for any choice of $w_{(1),\psi} \in I_\psi$ and $\zeta_{(1),\psi} \in I_\psi$ for each ψ. Given this result, the first conclusion of Lemma A.2 follows immediately.

To show this, we use the original indexing of the cells in $\mathscr{I}$ according to $(l_1, l_2, l_3, \ldots, l_{mt})$, $l_j \in \{1, 2, \ldots, 2BA\}$, $j = 1, 2, \ldots, mt$. Recall that $I_\psi = I(l_1^\psi, l_2^\psi, l_3^\psi, \ldots, l_{mt}^\psi)$. Consider the cells for which $l_j = 1$ for at least one j. In other words, consider the collection denoted $\mathscr{I}^* = \{I(l_1, l_2, l_3, \ldots, l_{mt}) \in \mathscr{I}: l_j = 1 \text{ for some } j \in \{1, 2, \ldots, mt\}\}$. These are the cells at the lower boundaries of C. Then $\mathscr{I}^*$ has at most $mt(2BA)^{(mt-1)}$ elements. For each $I_\psi \in \mathscr{I}^*$, define an associated diagonal subset $\mathscr{L}(\psi) \subset \mathscr{I}$ as follows: let

$$\mathscr{L}(\psi) = \{I_\theta \in \mathscr{I}: (l_1^\theta, l_2^\theta, l_3^\theta, \ldots, l_{mt}^\theta) = (l_1^\psi + u, l_2^\psi + u, l_3^\psi + u, \ldots, l_{mt}^\psi + u), \quad u = 0, 1, 2, \ldots\}.$$

Because $M_{n,k,\nu}(y, \tau(y)\gamma\nu)$ is bounded between 0 and 1 and nondecreasing on $\gamma \in U_{y,\nu}$, it follows that

$$\sum_{\{\theta:\, I_\theta \in \mathscr{L}(\psi)\}} \left| M_{n,k,\nu}(y, w_{(1),\theta}) - M_{n,k,\nu}(y, \zeta_{(1),\theta}) \right| \le 1,$$

with any choice of $w_{(1),\theta} \in I_\theta$ and $\zeta_{(1),\theta} \in I_\theta$ for each θ. Because every cell in $\mathscr{I}$ belongs to $\mathscr{L}(\psi)$ for some $I_\psi \in \mathscr{I}^*$, we have the desired result.

Recall the properties of $M^*(y, w)$ and $M^*_{k,\nu}(y, w_{(1)})$ given in the proof of Lemma A.1. Then the second conclusion in Lemma A.2 follows in the same way as the first, substituting p^*_ν for $p_{n,\nu}$, substituting $M^*_{k,\nu}(y, w_{(1)})$ for $M_{n,k,\nu}(y, w_{(1)})$ and substituting $M^*(y, w)$ for $M(y|w, G_n)$ in the above argument. □

We now need a series of definitions, after which we will present the general structure of the proof. First, for the subsequence $n(j(h))$ defined in Lemma A.1 let $G'_h = G_{n(j(h))}$, let $\mu'_h = \mu^{n(j(h))}$, let $\rho'^{(h)} = \rho^{(n(j(h)))}$ and let $T'(h, \delta) = T(n(j(h)), \delta)$. For each $k \in \{1, 2, \ldots, K\}$ and $\psi \in \{1, 2, \ldots, (2BA)^{mt}\}$, let $\Psi(k, \psi) = \{w \in S: w_{(1)} \in I_\psi, w_{(2)} = w_{(2),k}\}$. If we fix A and B, then for each (k, ψ) we have $\rho'^{(h)}(\Psi(k, \psi)) \to \rho(\Psi(k, \psi))$, w.p.1. For each (y, k, ψ), let $\Pi(y, k, \psi) = \{(\varpi, w) \in (S_y, S): \varpi = y, w \in \Psi(k, \psi)\}$. Then $\mu'_h(\Pi(y, k, \psi)) \to \mu(\Pi(y, k, \psi))$, w.p.1. This convergence is uniform over (y, k, ψ), because the set $\{(y, k, \psi): y \in S_y, k \in \{1, 2, \ldots, K\}, \psi \in \{1, 2, \ldots, (2BA)^{mt}\}\}$ is finite.

Let $\phi = \inf\{M(y|w, G): y \in S_y, w \in C\}$. Recall that F is a continuous and strictly increasing c.d.f. on R^m, so that $F > 0$ on R^m. Therefore, $M(y|w, G)$ is continuous in w, and $M(y|w, G) > 0$ for all w in the closure $\overline{C}$ of C for every y. This implies that $\phi > 0$. For each h, let

$$\varphi(h) = \min\{\mu'_h(\Pi(y, k, \psi))/\rho'^{(h)}(\Psi(k, \psi)): y \in S_y, k \in \{1, 2, \ldots, K\}, I_\psi \in \mathscr{I} \text{ and } \rho'^{(h)}(\Psi(k, \psi)) > 0\}$$

[where the minimum is over all (y, k, ψ) such that the stated conditions hold]. Then $\varphi(h) \to \varphi(\infty)$ as $k \to \infty$, where

$$\varphi(\infty) = \min\{\mu(\Pi(y, k, \psi))/\rho(\Psi(k, \psi)): y \in S_y, k \in \{1, 2, \ldots, K\}, I_\psi \in \mathscr{I} \text{ and } \rho(\Psi(k, \psi)) > 0\}.$$

Then $\varphi(\infty) \geq \phi$. Let θ be such that $\phi > \theta > 0$. Then there exists some h_θ such that $\varphi(h) > \phi - \theta > 0$ for all $h > h_\theta$.

Let $L = E_\mu[\ln(M(Y|W, G))]$, which is a finite, negative constant. Consider the same $\theta > 0$ as above. Then, using (A.4) and (A.5), we have that, w.p.1, there exists an $h'_\theta > h_\theta$ such that $E_{\mu'_h}[\ln(M(Y|W, G'_h))] > L - \theta$ for all $h > h'_\theta$.

Denote the set of observed data as $\mathscr{O}_h = \{(Y_i, W_i),\ i = 1, 2, \ldots, n(j(h))\}$. Let $1 > \varepsilon > 0$, and, for each k and h, let

$$\mathscr{J}(h, \varepsilon, k) = \big\{I_\psi \in \mathscr{I}\text{: for each } y \in S_y \text{ there exists some } (Y_i, W_i) \in \mathscr{O}_h \text{ with } Y_i = y,\ W_i \in \Psi(k, \psi) \text{ and } M(y|W_i, G'_h) > \varepsilon\big\}.$$

By definition, if $I_\psi \in \mathscr{J}(h, \varepsilon, k)$, then for each y there exists some $\zeta(h, \varepsilon, y, k, \psi) \in \Psi(k, \psi)$ such that $M(y|\zeta(h, \varepsilon, y, k, \psi), G'_h) > \varepsilon$. Let $V(h, \varepsilon, k)$ be the union of the $\Psi(k, \psi)$ over all ψ such that $I_\psi \in \mathscr{J}(h, \varepsilon, k)$. Let $V(h, \varepsilon)$ be the union of the $V(h, \varepsilon, k)$ over all k, and let $H_{h, \varepsilon, \delta} = V(h, \varepsilon) \cap T'(h, \delta)$.

Let $\mathscr{J}^0(h, \varepsilon, k)$ be the set of elements in $\mathscr{I}$ that are not contained in $\mathscr{J}(h, \varepsilon, k)$. Suppose that $I_\psi \in \mathscr{J}^0(h, \varepsilon, k)$. Then there exists some $y^0(k, \psi) \in S_y$ such that $M(y^0(k, \psi)|w, G'_h) \leq \varepsilon$ on $\Psi(k, \psi)$, and therefore $\ln(M(y^0(k, \psi)|w, G'_h)) \leq \ln(\varepsilon)$ on $\Psi(k, \psi)$. Let $V^0(h, \varepsilon, k)$ be the union of the $\Psi(k, \psi)$ over all ψ such that $I_\psi \in \mathscr{J}^0(h, \varepsilon, k)$, and let $V^0(h, \varepsilon)$ be the union of the $V^0(h, \varepsilon, k)$ over all $k \in \{1, 2, \ldots, K\}$. Then $V^0(h, \varepsilon) = V^c(h, \varepsilon) \cap T$, where "$c$" denotes "complement." Now, if $h > h'_\theta$, then

$$\begin{aligned}
&E_{\mu'_h}\big[\ln\big(M(Y|W, G'_h)\big)\big] \\
&\quad = \sum_{\{k, \psi\}} \big[E_{\mu'_h}\big[\ln\big(M(Y|W, G'_h)\big) I\{W \in \Psi(k, \psi)\}\big]\big] \\
&\quad \leq \sum_{\{k, \psi:\, I_\psi \in \mathscr{J}^0(h, \varepsilon, k)\}} \big[E_{\mu'_h}\big[\ln\big(M(Y|W, G'_h)\big) I\{W \in \Psi(k, \psi),\ Y = y^0(k, \psi)\}\big]\big] \\
&\quad \leq \sum_{\{k, \psi:\, I_\psi \in \mathscr{J}^0(h, \varepsilon, k)\}} \big[\ln(\varepsilon)\, \varphi(h)\, \rho'^{(h)}(\Psi(k, \psi))\big] \\
&\quad = \ln(\varepsilon)\, \varphi(h)\, \rho'^{(h)}\big(V^0(h, \varepsilon)\big).
\end{aligned}$$

Therefore, since $\ln(\varepsilon) < 0$,

$$\begin{aligned}
\rho'^{(h)}(V^c(h, \varepsilon) \cap T) = \rho'^{(h)}\big(V^0(h, \varepsilon)\big) &< (L - \theta)/(\varphi(h)\ln(\varepsilon)) \\
&< (L - \theta)/((\phi - \theta)\ln(\varepsilon)).
\end{aligned}$$

Recall that $\rho'^{(h)}(\Psi(k, \psi)) \to \rho(\Psi(k, \psi))$ uniformly over all (k, ψ), w.p.1. Because this convergence is uniform, there exists, w.p.1, some $h_\varepsilon > h'_\theta$ such that

$$\rho(V^c(h, \varepsilon) \cap T) = \rho\big(V^0(h, \varepsilon)\big) < (L - \theta)/((\phi - \theta)\ln(\varepsilon))$$

for all $h > h_\varepsilon$. For fixed values of A and B, this bound can be made arbitrarily small by choosing small enough ε.

For any $\xi > 0$, we can choose B large enough so that $\rho(T) > 1 - \xi/3$, and then ε small enough so that $(L - \theta)/((\phi - \theta)\ln(\varepsilon)) < \xi/3$. We may then take $\delta < \varepsilon^2/3$, and, as a result of Lemma A.2 (discussed after the statement of the lemma), we can choose A large enough so that $\rho(T'(h, \delta)) > 1 - \xi/3$ for all h. Then there exists some h_ε such that $\rho(H_{h,\varepsilon,\delta}) = \rho(T'(h, \delta) \cap V(h, \varepsilon)) > 1 - (\xi/3 + \xi/3 + \xi/3) = 1 - \xi$ for all $h > h_\varepsilon$. Recall that the probability space is denoted $(\Omega, \mathscr{F}, P)$, and let $\Omega_0 \subset \Omega$ be the set on which h_ε exists.

We now present the general structure of the proof, which is outlined in expressions (A.7) and (A.8) below. The three inequalities in (A.7) follow from inequalities (A.3), (A.4) and (A.2), respectively. In both (A.7) and (A.8) we use "$\simeq$" to indicate "within any given distance, w.p.1, for large enough B, A, $1/\varepsilon$ and h" in a sense that will be formalized and proved below. Then we demonstrate the following:

$$
\begin{aligned}
&E_\mu\big[\ln(M^*(Y,W))\,I\{W \in H_{h,\varepsilon,\delta}\}\big] \\
&\quad \simeq E_{\mu'_h}\big[\ln\big(M(Y|W,G'_h)\big)\,I\{W \in H_{h,\varepsilon,\delta}\}\big] \\
&\quad \geq E_{\mu'_h}\big[\ln\big(M(Y|W,G'_h)\big)\big] \geq E_{\mu'_h}\big[\ln(M(Y|W,G))\big] \\
&\quad \simeq E_\mu\big[\ln(M(Y|W,G))\big] \qquad \text{(A.7)} \\
&\quad \simeq E_\mu\big[\ln(M(Y|W,G))\,I\{W \in H_{h,\varepsilon,\delta}\}\big] \\
&\quad \geq E_\mu\big[\ln\big(M(Y|W,G'_h)\big)\,I\{W \in H_{h,\varepsilon,\delta}\}\big] \\
&\quad \simeq E_\mu\big[\ln(M^*(Y,W))\,I\{W \in H_{h,\varepsilon,\delta}\}\big]
\end{aligned}
$$

(notice that the first and last terms are the same), and

$$
\begin{aligned}
&E_\mu\big[\ln(M^*(Y,W))\,I\{W \in H_{h,\varepsilon,\delta}\}\big] \\
&\qquad \simeq E_\mu\big[\ln(M(Y|W,G))\,I\{W \in H_{h,\varepsilon,\delta}\}\big]. \qquad \text{(A.8)}
\end{aligned}
$$

The formalization of (A.8) will be used to show that, w.p.1, for each $y \in S_y$ we have $M^*(y,w) = M(y|w,G)$ for w almost everywhere (ρ).

We now address each of the four relations denoted as "$\simeq$" in (A.7). We begin with the third "$\simeq$." We will follow with the second, fourth and first "$\simeq$," in that order.

The third "$\simeq$." First, notice that $E_\mu[\ln(M(Y|W,G))] > -\infty$, which follows from the fact that $M(y|w,G)\ln(M(y|w,G)) > -\exp(-1)$ for all y and w. Let $\pi > 0$. Then there exists some ξ_π such that, for any Borel set $H \in S$ with $\rho(H) > 1 - \xi_\pi$,

$$
\big|E_\mu[\ln(M(Y|W,G))] - E_\mu[\ln(M(Y|W,G))\,I\{W \in H\}]\big| < \pi.
$$

[This follows by contradiction: if not, then for every $\xi > 0$ there exists some Borel set $K_\xi \subset S$ such that $\rho(K_\xi) < \xi$ and $|E_\mu[\ln(M(Y|W,G))I\{W \in K_\xi\}]| \geq \pi$. Take $\xi(\eta) = 1/2^\eta$ for $\eta = 1, 2, \ldots,$ and let $D_q = \bigcup_{(\eta > q)} K_{\xi(\eta)}$. Then $\rho(D_q) < 2^{-q}/(1 - 1/2) \to 0$ as $q \to \infty$. Because $\ln(M(y|w,G)) \leq 0$, we have

$$\left|E_\mu\left[\ln(M(Y|W,G))I\{W \in D_q\}\right]\right| \geq \left|E_\mu\left[\ln(M(Y|W,G))I\{W \in K_{\xi(q)}\}\right]\right| \geq \pi$$

for all q. But D_q converges to some Borel set $D_\infty \subset S$ with $\rho(D_\infty) = 0$, and therefore $E_\mu[\ln(M(Y|W,G))I\{W \in D_q\}] \to E_\mu[\ln(M(Y|W,G))I\{W \in D_\infty\}] = 0$ by the dominated convergence theorem, giving a contradiction.]

The second "$\simeq$." Expression (A.5) implies that, w.p.1, there exists some h_δ such that $|E_{\mu'_h}[\ln(M(Y|W,G))] - E_\mu[\ln(M(Y|W,G))]| < \delta$ for all $h > h_\delta$. Let $\Omega_1 \subset \Omega$ be the set on which h_δ exists.

The fourth "$\simeq$." Consider B, ε, δ and A to be fixed. Recall from the proof of Lemma A.1 that, for each k, the convergence result (A.6) holds for w with $w_{(2)} = w_{(2),k}$ and $w_{(1)}$ almost everywhere (λ) on $R^{m\times t}$, where λ is the Lebesgue measure. Therefore, for each y and k, we have $M(y|w, G'_h) \to M^*(y,w)$ for $w_{(2)} = w_{(2),k}$ and $w_{(1)}$ almost everywhere (λ) on $R^{m\times t}$. Then, for each (y,k,ψ), there exists some $w(y,k,\psi) \in \Psi(k,\psi)$ such that $M(y|w(y,k,\psi), G'_h) \to M^*(y, w(y,k,\psi))$. The set of all values for (y,k,ψ) is finite, so that we have uniform convergence over all (y,k,ψ). Therefore, there exists some h'_δ such that, if $h > h'_\delta$, then $|M(y|w(y,k,\psi), G'_h) - M^*(y, w(y,k,\psi))| < \delta$ for all (y,k,ψ).

Consider any (y,k,ψ) such that $I_\psi \in \mathcal{I}_{n(j(h))} \cap \mathcal{J}(h,\varepsilon,k)$. Then there exists some $\zeta(h,\varepsilon,y,k,\psi) \in \Psi(k,\psi)$ such that $M(y|\zeta(h,\varepsilon,y,k,\psi), G'_h) > \varepsilon > \varepsilon^2 > 3\delta$. Also, from the discussion after the statement of Lemma A.2, we have $v_{n(j(h))}(y,k,\psi) < \delta$ and $v^*(y,k,\psi) < \delta$. Suppose that $h > h'_\delta$, and consider any $w \in \Psi(k,\psi)$. Then, using the triangle inequality, we have

$$\begin{aligned}
&\left|M(y|w, G'_h) - M^*(y,w)\right| \\
&\quad \leq \left|M(y|w, G'_h) - M(y|w(y,k,\psi), G'_h)\right| \\
&\qquad + \left|M(y|w(y,k,\psi), G'_h) - M^*(y, w(y,k,\psi))\right| \\
&\qquad + \left|M^*(y, w(y,k,\psi)) - M^*(y,w)\right| < 3\delta.
\end{aligned}$$

Also, using the fact that $M(y|\zeta(h,\varepsilon,y,k,\psi), G'_h) > \varepsilon$, we have $M(y|w, G'_h) > \varepsilon - \delta > 0$. Therefore, for all $w \in K(h,\varepsilon,\delta)$, we have that $|M(y|w, G'_h) - M^*(y,w)| < 3\delta$ and $M(y|w, G'_h) > \varepsilon - \delta > \varepsilon - 3\delta > 0$ for each y. Also, again using the triangle inequality and the fact that $M(y|\zeta(h,\varepsilon,y,k,\psi), G'_h) > \varepsilon$, we have $M^*(y,w) > \varepsilon - 3\delta$. Then since $\ln(\cdot)$ is monotonic and differentiable with maximum derivative $1/(\varepsilon - 3\delta)$ on $[\varepsilon - 3\delta, 1]$ and because $\delta < \varepsilon^2/3$, we have

$$\begin{aligned}
\left|\ln(M(y|w, G'_h)) - \ln(M^*(y,w))\right| &< 3\delta/(\varepsilon - 3\delta) \\
&< \varepsilon^2/(\varepsilon - \varepsilon^2) = \varepsilon/(1 - \varepsilon).
\end{aligned}$$

Finally, then,

$$\begin{aligned}\left|E_{\mu}\left[\ln\left(M(Y|W,G'_h)\right)I\{W\in K_{h,\varepsilon,\delta}\}\right]-E_{\mu}\left[\ln\left(M^*(Y,W)\right)I\{W\in K_{h,\varepsilon,\delta}\}\right]\right|\\ \leq E_{\mu}\left[\left|\ln\left(M(Y|W,G'_h)\right)I\{W\in K_{h,\varepsilon,\delta}\}-\ln\left(M^*(Y,W)\right)I\{W\in K_{h,\varepsilon,\delta}\}\right|\right]\\ <\varepsilon/(1-\varepsilon).\end{aligned}$$

The first "$\simeq$." We continue the above discussion with B, ε, δ and A fixed, and we now demonstrate a relationship of the following type:

$$E_{\mu'_h}\left[\ln\left(M(Y|W,G'_h)\right)I\{W\in K_{h,\varepsilon,\delta}\}\right]\simeq E_{\mu}\left[\ln\left(M(Y|W,G'_h)\right)I\{W\in K_{h,\varepsilon,\delta}\}\right].$$

Combined with the fourth "$\simeq$," this will imply the first "$\simeq$," w.p.1, as we will show. First, since $\mu'_h(\Pi(y,k,\psi))\to\mu(\Pi(y,k,\psi))$ uniformly over (y,k,ψ) w.p.1, there exists some h''_δ such that, if $h>h''_\delta$, then

$$\left|\mu'_h(\Pi(y,k,\psi))-\mu(\Pi(y,k,\psi))\right|<\delta\left(\left|\ln(\varepsilon-3\delta)\right|2^mK(2BA)^{mt}\right)^{-1}$$

for all (y,k,ψ). The reason for choosing this particular bound will become clear shortly. Let $\Omega_2\subset\Omega$ be the set on which h''_δ exists.

Suppose that $\omega\in\Omega_0\cap\Omega_2$ and that $h>h_\varepsilon$, $h>h'_\delta$ and $h>h''_\delta$. Consider any (y,k,ψ) such that $I_\psi\in\mathscr{I}_{n(j(h))}\cap\mathscr{J}(h,\varepsilon,k)$. Then, for any $w\in\Psi(k,\psi)$ and $\zeta\in\Psi(k,\psi)$, we have that $\ln(M(y|w,G'_h))>\ln(\varepsilon-3\delta)$, that $\ln(M(y|\zeta,G'_h))>\ln(\varepsilon-3\delta)$ and that $|M(y|w,G'_h)-M(y|\zeta,G'_h)|<\delta$. Since $\ln(\cdot)$ is monotonic and differentiable with maximum derivative $1/(\varepsilon-3\delta)$ on $[\varepsilon-3\delta,1]$, we have

$$\begin{aligned}\left|\ln\left(M(y|w,G'_h)\right)-\ln\left(M(y|\zeta,G'_h)\right)\right|&<\delta/(\varepsilon-3\delta)<3\delta/(\varepsilon-3\delta)\\ &<\varepsilon^2/(\varepsilon-\varepsilon^2)=\varepsilon/(1-\varepsilon).\end{aligned}$$

In the following, let $I(y,k,\psi)$ be shorthand for $I\{(Y,W)\in\Pi(y,k,\psi)\}$. We now show that

$$\begin{aligned}&\left|E_{\mu'_h}\left[\ln\left(M(Y|W,G'_h)\right)I(y,k,\psi)\right]-E_{\mu}\left[\ln\left(M(Y|W,G'_h)\right)I(y,k,\psi)\right]\right|\\ &\leq\mu(\Pi(y,k,\psi))\varepsilon/(1-\varepsilon)+\left(\delta/\left(2^mK(2BA)^{mt}\right)\right)\\ &\quad\times\left(1+\varepsilon\left((1-\varepsilon)\left|\ln(\varepsilon-3\delta)\right|\right)^{-1}\right).\end{aligned}\tag{A.9}$$

Suppose that $\mu(\Pi(y,k,\psi))\leq\mu'_h(\Pi(y,k,\psi))$. Then we create a new measure $\hat{\mu}$ by adding a mass of weight $\hat{p}=(\mu'_h(\Pi(y,k,\psi))-\mu(\Pi(y,k,\psi)))$ to μ at some point $(y,w)\in\Pi(y,k,\psi)$. Since $\hat{\mu}(\Pi(y,k,\psi))=\mu'_h(\Pi(y,k,\psi))$, then

$$\begin{aligned}&\left|E_{\mu'_h}\left[\ln\left(M(Y|W,G'_h)\right)I(y,k,\psi)\right]+E_{\hat{\mu}}\left[\ln\left(M(Y|W,G'_h)\right)I(y,k,\psi)\right]\right|\\ &\leq\hat{\mu}(\Pi(y,k,\psi))\varepsilon/(1-\varepsilon).\end{aligned}$$

Therefore, since $\hat{p} < \delta(|\ln(\varepsilon - 3\delta)|2^m K(2BA)^{mt})^{-1}$ and $\ln(M(y|w, G'_h)) > \ln(\varepsilon - 3\delta)$ on $\Pi(y, k, \psi)$, we have

$$\begin{aligned}
&\left|E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right] - E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right| \\
&\quad \le \left|E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right] - E_{\hat{\mu}}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right| \\
&\qquad + \left|E_{\hat{\mu}}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right] - E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right| \\
&\quad \le \hat{\mu}(\Pi(y, k, \psi))\,\varepsilon/(1 - \varepsilon) + \hat{p}(|\ln(\varepsilon - 3\delta)|) \\
&\quad = \mu(\Pi(y, k, \psi))\,\varepsilon/(1 - \varepsilon) + \hat{p}((\varepsilon/(1 - \varepsilon)) + |\ln(\varepsilon - 3\delta)|) \\
&\quad \le \mu(\Pi(y, k, \psi))\,\varepsilon/(1 - \varepsilon) + \left(\delta/\left(2^m K(2BA)^{mt}\right)\right) \\
&\qquad \times\left(\varepsilon((1 - \varepsilon)|\ln(\varepsilon - 3\delta)|)^{-1} + 1\right).
\end{aligned}$$

Alternatively, suppose that $\mu(\Pi(y, k, \psi)) > \mu'_h(\Pi(y, k, \psi))$. Then we create a new measure $\hat{\mu}$ by adding a mass of weight $\hat{p} = (\mu(\Pi(y, k, \psi)) - \mu'_h(\Pi(y, k, \psi)))$ to μ'_h at some point $(y, w) \in \Pi(y, k, \psi)$. Then $\hat{\mu}(\Pi(y, k, \psi)) = \mu(\Pi(y, k, \psi))$, $\hat{p} < \delta(|\ln(\varepsilon - 3\delta)|2^m K(2BA)^{mt})^{-1}$ and $\ln(M(y|w, G'_h)) > \ln(\varepsilon - 3\delta)$ on $\Pi(y, k, \psi)$. Therefore, we have

$$\begin{aligned}
&\left|E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right] - E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right| \\
&\quad \le \left|E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right] - E_{\hat{\mu}}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right| \\
&\qquad + \left|E_{\hat{\mu}}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right] - E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right| \\
&\quad \le \hat{p}(|\ln(\varepsilon - 3\delta)|) + \mu(\Pi(y, k, \psi))(\varepsilon/(1 - \varepsilon)) \\
&\quad \le \left(\delta/\left(2^m K(2BA)^{mt}\right)\right) + \mu(\Pi(y, k, \psi))(\varepsilon/(1 - \varepsilon)) \\
&\quad \le \left(\delta/\left(2^m K(2BA)^{mt}\right)\right)\left(1 + \varepsilon((1 - \varepsilon)|\ln(\varepsilon - 3\delta)|)^{-1}\right) \\
&\qquad + \mu(\Pi(y, k, \psi))(\varepsilon/(1 - \varepsilon)).
\end{aligned}$$

Hence, we have (A.9). As a consequence, where the following summations are over (y, k, ψ) such that $I_\psi \in \mathcal{I}_{n(j(h))} \cap \mathcal{J}(h, \varepsilon, k)$, we have

$$\begin{aligned}
&\left|E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I\{W \in H_{h,\varepsilon,\delta}\}\right] - E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I\{W \in H_{h,\varepsilon,\delta}\}\right]\right| \\
&\quad = \left|\sum\left[E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right]\right. \\
&\qquad \left. - \sum\left[E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right]\right| \\
&\quad \le \sum\left[\left|E_{\mu'_h}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right.\right. \\
&\qquad \left.\left. - E_{\mu}\left[\ln\left(M(Y|W, G'_h)\right)I(y, k, \psi)\right]\right|\right]
\end{aligned}$$

$$\begin{aligned}
&\le \sum \Big[\mu(\Pi(y,k,\psi))\,\varepsilon/(1-\varepsilon) + \big(\delta/\big(2^m K(2BA)^{mt}\big)\big) \\
&\qquad\qquad \times\big(1+\varepsilon((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1}\big)\Big] \\
&\le \varepsilon/(1-\varepsilon) + 2^m K(2BA)^{mt}\big(\delta/\big(2^m K(2BA)^{mt}\big)\big) \\
&\quad \times\big(1+\varepsilon((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1}\big) \\
&= \varepsilon/(1-\varepsilon) + \delta\big(1+\varepsilon((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1}\big).
\end{aligned}$$

Recall that $3\delta < \varepsilon^2 < \varepsilon$, and notice that $((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1} \to 0$ as $\varepsilon \to 0$. Incorporating the fourth "$\simeq$," we have

$$\begin{aligned}
&\big|E_{\mu'_h}\big[\ln(M(Y|W,G'_h))I\{W\in H_{h,\varepsilon,\delta}\}\big] \\
&\quad -E_{\mu}\big[\ln(M^*(Y|W,G))\,I\{W\in H_{h,\varepsilon,\delta}\}\big]\big| \\
&\qquad \le \big|E_{\mu'_h}\big[\ln(M(Y|W,G'_h))I\{W\in H_{h,\varepsilon,\delta}\}\big] \\
&\qquad\quad -E_{\mu}\big[\ln(M(Y|W,G'_h))I\{W\in H_{h,\varepsilon,\delta}\}\big]\big| \\
&\qquad\quad +\big|E_{\mu}\big[\ln(M(Y|W,G'_h))I\{W\in H_{h,\varepsilon,\delta}\} \\
&\qquad\qquad -\ln(M^*(Y|W,G))\,I\{W\in H_{h,\varepsilon,\delta}\}\big]\big| \\
&\qquad < 2\varepsilon/(1-\varepsilon) + \delta\big(1+\varepsilon((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1}\big),
\end{aligned}$$

which completes the formalization of (A.7).

Now let $\pi > 0$, let $\xi < \xi_\pi$, let $\varepsilon > 0$ and choose B, δ and A as discussed above. Then $\rho(H_{h,\varepsilon,\delta}) > 1-\xi > 1-\xi_\pi$. Suppose that $\omega \in \Omega_0 \cap \Omega_1 \cap \Omega_2$, and suppose that $h > h_\varepsilon$, $h > h_\delta$, $h > h'_\delta$ and $h > h''_\delta$. Then, using the formalization of (A.7), we have

$$\begin{aligned}
&E_{\mu}\big[\ln(M^*(Y,W))\,I\{W\in H_{h,\varepsilon,\delta}\}\big] \\
&\qquad \ge E_{\mu}\big[\ln(M(Y|W,G))\,I\{W\in H_{h,\varepsilon,\delta}\}\big] + 2\varepsilon/(1-\varepsilon) \\
&\qquad\quad + \delta\big(1+\varepsilon((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1}\big) + \delta + \pi
\end{aligned}$$

and

$$\begin{aligned}
&E_{\mu}\big[\ln(M(Y|W,G))\,I\{W\in H_{h,\varepsilon,\delta}\}\big] \\
&\qquad \ge E_{\mu}\big[\ln(M^*(Y,W))\,I\{W\in H_{h,\varepsilon,\delta}\}\big] + \varepsilon/(1-\varepsilon),
\end{aligned}$$

so that

$$\begin{aligned}
&\big|E_{\mu}\big[\ln(M(Y|W,G))\,I\{W\in H_{h,\varepsilon,\delta}\}\big] \\
&\quad -E_{\mu}\big[\ln(M^*(Y,W))\,I\{W\in H_{h,\varepsilon,\delta}\}\big]\big| \\
&\qquad \le 2\varepsilon/(1-\varepsilon) + \delta\big(1+\varepsilon((1-\varepsilon)|\ln(\varepsilon-3\delta)|)^{-1}\big) + \delta + \pi.
\end{aligned} \tag{A.10}$$

This finally formalizes (A.8), since, for any $\chi > 0$, we can choose π, ξ, B, ε, δ and A so that the right-hand side of (A.10) is less than χ on an ω set of probability 1 (P).

Returning to Lemma A.1, since $\sum_y M(y|w, G'_h) = 1$ for all w and h, the convergence (A.6) implies that $\sum_y M^*(y, w) = 1$ for w almost everywhere ρ on S. Therefore, for such w,

$$\sum_y \left[M(y|w, G) \ln(M(y|w, G)) \right] \geq \sum_y \left[M(y|w, G) \ln(M^*(y, w)) \right]$$

by the nonnegativity of the Kullback–Liebler distance, where equality holds only if $M(y|w, G) = M^*(y, w)$ for all y. Let

$$D_0 = \left\{ w \in S: M(y|w, G) \neq M^*(y, w) \text{ for some } y \in S_y \right\}.$$

Then, for $w \in D_0$,

$$\sum_y \left[M(y|w, G) \ln(M(y|w, G)) \right] > \sum_y \left[M(y|w, G) \ln(M^*(y, w)) \right].$$

Suppose that $\rho(D_0) > 0$ for all ω in some set $\Omega_3 \in \Omega$, with $P(\Omega_3) > 0$. If $\omega \in \Omega_3$, then for some $\vartheta > 0$ there exists a set $D_\vartheta \subset D_0$ with $\rho(D_\vartheta) = d > 0$ such that

$$\sum_y \left[M(y|w, G) \ln(M(y|w, G)) \right] > \sum_y \left[M(y|w, G) \ln(M^*(y, w)) \right] + \vartheta$$

for all $w \in D_\vartheta$.

Now, in the argument leading to (A.10), we may choose π, ξ, B, ε, δ and A so that $\xi < d/2$ and $2\varepsilon/(1 - \varepsilon) + \delta(1 + \varepsilon((1 - \varepsilon)|\ln(\varepsilon - 3\delta)|)^{-1}) + \delta + \pi < \vartheta d/2$. Then take $\omega \in \Omega_3 \cap \Omega_0 \cap \Omega_1 \cap \Omega_2$, where this set has probability $P(\Omega_3 \cap \Omega_0 \cap \Omega_1 \cap \Omega_2) = P(\Omega_3) > 0$. Take $h > h_\varepsilon$, $h > h_\delta$, $h > h'_\delta$ and $h > h''_\delta$. Then we have (A.10), but we also have $\rho(H_{h, \varepsilon, \delta}) > 1 - d/2$ for all h, so that $\rho(H_{h, \varepsilon, \delta} \cap D_\vartheta) \geq d/2$ for all h. Then

$$\begin{aligned} &E_\mu\left[\ln(M(Y|W, G)) I\{W \in H_{h, \varepsilon, \delta}\} \right] \\ &\quad \geq E_\mu\left[\ln(M^*(Y, W)) I\{W \in H_{h, \varepsilon, \delta}\} \right] + \vartheta d/2 \end{aligned}$$

for all h, which contradicts (A.10). Therefore, $\rho(D_0) = 0$ w.p.1, giving conclusion 1.

Conclusion 1 implies that for any subsequence $n(j)$ there exists a further subsequence $n(j(h))$ such that $\eta_{n(j(h))} \to_{\mathscr{D}} \mu$, w.p.1. This implies tightness of η_n, for example from Theorem 25.10 of Billingsley (1986). Also, if $\eta_{n(j)}$ converges to some probability measure μ^*, then $\eta_{n(j(h))} \to_{\mathscr{D}} \mu^*$, so that $\mu = \mu^*$. It follows that $\eta_n \to_{\mathscr{D}} \mu$, for example from the corollary to Theorem 25.10 in Billingsley (1986). This gives conclusion 2. □

APPENDIX B

Proof of Theorem 4.1. Consider any $\Pi \in S$ with $\rho(\Pi) = 1$. Then for each A_l of condition 3 there exists some $(x_2^l, z_l) \in A_l$ such that $C(x_2^l, z_l)$ satisfies the following: for each $x_1 \in C(x_2^l, z_l)$ and for any open neighborhood $N(x_1)$ of x_1, the set $N(x_1) \cap C(x_2^l, z_l)$ contains a sequence $\{\nu_k, k = 1, 2, \ldots\}$ converging to x_1, and $\Pi_l(x_2^l, z_l) \equiv \{(\nu, x_2, z) \in S: (x_2, z) = (x_2^l, z_l), \nu \in$

$\{\nu_k,\ k = 1, 2, \ldots\}\} \subset \Pi$. Otherwise, for each $(x_2, z) \in A_l$ and for some $x_1^* \in C(x_2, z)$, there exists an open interval $N^*(x_1^*)$ such that $\{(x, z) \in S\colon (x_2, z) = (x_2^l, z_l),\ x_1 \in N^*(x_1^*)\} \cap \Pi$ is empty. But $\rho(\{(x, z) \in S\colon (x_2, z) \in A_l,\ x_1 \in N^*(x_1^*)\}) > 0$, which then contradicts the assumption that $\rho(\Pi) = 1$.

We first prove identifiability. For this part of the proof, let Π be the set on which $M(\mathbf{1}|w, \alpha, G) = M(\mathbf{1}|w, \alpha^*, G^*)$. Because M is a continuous function in w, this equality holds for all w in the set $\Psi_l = \{(x, z) \in S\colon (x_2, z) = (x_2^l, z_l),\ x_1 \in C(x_2^l, z_l)\}$ for each l. We now describe a procedure that identifies α and G, given the values of $M(\mathbf{1}|w, \alpha, G)$ on $\Psi = \bigcup_l \Psi_l$.

For any l, the sign of $\alpha_{(1)}$ is determined by the value of $M(\mathbf{1}|(x_1, x_2^l, z_l), \alpha, G)$ on ν and ν^*. We may assume, without loss of generality, that $\alpha_{(1)} > 0$. Then, using condition 3 for each l, the values of

$$M\left(\mathbf{1}\left(x_1, x_2^l, z_l\right), \alpha, G\right) = \Pr\left(\left(-1/\alpha_{(1)}\right)\left(x_2^l \alpha_{(2)} + z_l b + \varepsilon\right) \le x_1 G\right)$$

on Ψ_l determine the distribution of $(-1/\alpha_{(1)})(x_2^l \alpha_{(2)} + z_l b + \varepsilon)$. Let $U_l = x_2^l \alpha_{(2)} + z_l b$. Then for any $q \times m$ matrix T, this determines the distribution of $(1/\alpha_{(1)})T(U_l + \varepsilon)$. For l as in condition 4, there exists a $1 \times m$ vector T_0 with rank 1 such that $T_0 U_l = 0$, so that the distribution of $(1/\alpha_{(1)})T_0 \varepsilon$ is determined. Because F is continuous, $T_0 \varepsilon$ is not degenerate at 0, and therefore $\alpha_{(1)}$ is determined. Now consider any $q \times m$ matrix T of rank q. Using condition 1 and the continuity of characteristic functions, from the characteristic function of $TU_l + T\varepsilon$ we can solve for the characteristic function of TU_l for every l. This argument is of a type that extends simpler proofs for the identifiability of general location parameter mixtures in Teicher (1961) and Maritz and Lwin (1989).

Let $\alpha_{(2),j}$ be the jth element of $\alpha_{(2)}$. Using $l(j)$ from condition 5 with $j = s - 1, s - 2, \ldots, 1$, and using option (a), we can choose T_j such that $T_j z_l = \mathbf{0}$, $T_j x_{2,j-1}^l = \mathbf{0}$ when $j \ge 2$, and $T_j x_{2,j}^l \ne \mathbf{0}$ (where $\mathbf{0}$ is a $q \times 1$ vector with elements equal to 0), so that we successively determine the values of $\alpha_{2,j}$ $j = s - 1, s - 2, \ldots, 1$. Using option (b) with T as the identity matrix, then

$$\begin{aligned} &U_{l(j)} \sim x_2^{l(j)}\alpha_{(2)} + z_{l(j)}b, \qquad U_{k(j)} \sim x_2^{k(j)}\alpha_{(2)} + z_{k(j)}b, \\ &U_{l(j)} \sim U_{k(j)} + \left(x_{2,j}^{l(j)} - x_{2,j}^{k(j)}\right)\alpha_{(2),j} + \left(x_{2,j+1}^{l(j+1)} - x_{2,j+1}^{k(j+1)}\right)\alpha_{(2),j+1} \\ &\qquad + \left(x_{2,s-1}^{l(s-1)} - x_{2,s-1}^{k(s-1)}\right)\alpha_{(2),s-1}, \end{aligned} \tag{B.1}$$

so that we successively determine the values of $\alpha_{(2),j}$, $j = s - 1, s - 2, \ldots, 1$. [The two options (a) and (b) may clearly be used in combination.] Finally, for l as in condition 6, we may choose T to be a left inverse T_l for z_l. Then $T_l U_l = T_l x_2^l \alpha_{(2)} + b$, and the distribution G for b is determined.

We now prove consistency. We do so by showing that for $n(j(h))$ in conclusion 1 of Theorem 3.1 there exists a further subsequence $n(j(h(\tau)))$ such that $G_{n(j(h(\tau)))} \to_{\mathscr{D}} G$ and $\alpha_{n(j(h(\tau)))} \to \alpha$ as $\tau \to \infty$. It follows that $G_n \to_{\mathscr{D}} G$ and $\alpha_n \to \alpha$ as $n \to \infty$ (e.g., by contradiction at a continuity point of G). For

this proof, let Π be the set of values for w on which conclusion 1 of Theorem 3.1 holds. For notational convenience we replace $n(j(h))$ with h in the remainder of this proof. Without loss of generality, let $\alpha_1 > 0$. Then, for each l and for every $(x_1, x_2^l, z_l) \in \Psi_l$, we have that

$$
\begin{aligned}
&M\big(\mathbf{1}|(x_1, x_2^l, z_l), \alpha_h, G_h\big) \\
&\quad = \Pr\Big(\big(-1/\alpha_{h,(1)}\big)\big(x_2^l \alpha_{h,(2)} + z_l b + \varepsilon\big) \\
&\qquad\qquad \le x_1 | G_h, \alpha_{h,(1)} > 0\Big) I\{\alpha_{h,(1)} > 0\} \\
&\quad + \Pr\Big(\big(x_1 \alpha_{h,(1)} + x_2^l \alpha_{h,(2)} + z_l b + \varepsilon\big) \\
&\qquad\qquad \ge \mathbf{0} | G_h, \alpha_{h,(1)} \le 0\Big) I\{\alpha_{h,(1)} \le 0\},
\end{aligned}
\tag{B2}
$$

where $I\{\cdot\}$ is an indicator function taking the value 1 if the argument is true and 0 if the argument is false, and where, if $I\{\alpha_{h,(1)} > 0\}$ or $I\{\alpha_{h,(1)} \le 0\}$ is equal to 0, then the corresponding term on the right-hand side of (B.2) is defined to be equal to 0.

We may use Helly's selection theorem for a not necessarily tight sequence of measures to obtain a further subsequence $h(\tau)$, $\tau = 1, 2, \ldots,$ such that the following hold: $I\{\alpha_{h(\tau),(1)} > 0\} = \delta$ for all τ, where δ is equal to 0 or 1. If $\delta = 1$, then for each l there exists a (possibly improper) probability measure μ_l with a (possibly improper) c.d.f. D_l on R^m such that

$$
\begin{aligned}
&\Pr\Big(\big(-1/\alpha_{h(\tau),(1)}\big)\big(x_2^l \alpha_{h(\tau),(2)} + z_l b + \varepsilon\big) \le x_1 | G_{h(\tau)}, \alpha_{h(\tau),(1)} > 0\Big) \\
&\qquad \to D_l(x_1)
\end{aligned}
\tag{B.3}
$$

on continuity points of D_l as $\tau \to \infty$. This construction may be done so that (B.3) holds on a countable dense subset $C'(x_2^l, z_l)$ of $C(x_2^l, z_l)$, such that for any $x_1 \in C(x_2^l, z_l)$ the set $C'(x_2^l, z_l)$ contains a sequence converging to x_1 from above. (The sequence may be constant at x_1.) This may be done by finding a countable set $C'(x_2^l, z_l)$ which contains, for each $x_1 \in C(x_2^l, z_l)$, a sequence converging from above to x_1. Such a set must exist because $C(x_2^l, z_l)$ is countable or open in R^m. Then $C'(x_2^l, z_l)$ can be included as part of a countable dense subset of R^m, which can then be used in the same way as the rational numbers are used in Billingsley (1986) for the preliminary construction.

Now suppose that $\delta = 0$. Then for $x_1 \in C'(x_2^l, z_l)$ the second term on the right in (B.2), which is nonincreasing in x_1, must converge to $M(\mathbf{1}|(x_1, x_2^l, z_l), \alpha, G)$ as $\tau \to \infty$. However, as x_1 approaches ν and x_1^* approaches ν^* of condition 3, then eventually $M(\mathbf{1}|(x_1^*, x_2^l, z_l), \alpha, G) > M(\mathbf{1}|(x_1, x_2^l, z_l), \alpha, G)$ because F is strictly increasing. This contradicts (B.2). Therefore, $\delta = 1$, and from (B.2) we have for each l that

$$
D_l(x_1) = \Pr\Big(\big(-1/\alpha_{(1)}\big)\big(x_2^l \alpha_{(2)} + z_l b + \varepsilon\big) \le x_1 | G\Big)
\tag{B.4}
$$

for every $x_1 \in C'(x_2^l, z_l)$. Because the term on the right-hand side of (B.4) is continuous in x_1 and because D_l is right continuous, (B.4) holds for every

$x_1 \in C(x_2^l, z_l)$. Condition 3 then implies that $D_l(x_1)$ is the c.d.f. for $(-1/\alpha_{(1)})(x_2^l \alpha_{(2)} + z_l b + \varepsilon)$.

From this point the proof parallels that of identifiability. Let $U_l \sim (x_2^l \alpha_2 + z_l b)|G$, and let $U_{l,h(\tau)} \sim (x_2^l \alpha_{h(\tau),2} + z_l b)|G_{h(\tau)}$, where " $\sim$ " indicates "is distributed as." For any $q \times m$ matrix T, then $(1/\alpha_{h(\tau),(1)})(TU_{l,h(\tau)} + T\varepsilon) \to_{\mathscr{D}} (1/\alpha_{(1)})(TU_l + T\varepsilon)$. For l as in condition 4 and with T_0 as above, then $T_0 U_l \sim 0$ and $T_0 U_{l,h(\tau)} \sim 0$. Because $T_0 \varepsilon$ is not degenerate at 0, it follows that $\alpha_{h(\tau),(1)} \to \alpha_1$.

Because $TU_{l,h(\tau)} + T\varepsilon \to_{\mathscr{D}} TU_l + T\varepsilon$ for each l, it follows that the distributions for $TU_{l,h(\tau)}$ are a tight sequence. Also, using condition 1 as in the proof of identifiability, if $TU_{l,h(\tau)}$ converges in distribution on some further subsequence, then the limiting distribution must be that of TU_l. These two facts together imply that $TU_{l,h(\tau)} \to_{\mathscr{D}} TU_l$, as can be shown by contradiction at a continuity point of the c.d.f. for TU_l.

Let $\alpha_{h(\tau),(2),j}$ be the jth element of $\alpha_{h(\tau),(2)}$. For $l(j)$ as in condition 5, using option (a), we obtain T_j as above. Then, with $T = T_j$, $j = s-1, s-2, \ldots, 1$, we successively demonstrate that $\alpha_{h(\tau),(2),j} \to \alpha_{(2),j}$, $j = s-1, s-2, \ldots, 1$. Using option (b) and the analogs to (B.1) for $U_{l,h(\tau)}$ and U_l, then

$$\begin{aligned}
&\left(x_{2,j}^{l(j)} - x_{2,j}^{k(j)}\right)\alpha_{h(\tau),(2),j} + \left(x_{2,j+1}^{l(j+1)} - x_{2,j+1}^{k(j+1)}\right)\alpha_{h(\tau),(2),j+1} + \cdots \\
&\quad + \left(x_{2,s-1}^{l(s-1)} - x_{2,s-1}^{k(s-1)}\right)\alpha_{h(\tau),(2),s-1} \\
&\qquad \to \left(x_{2,j}^{l(j)} - x_{2,j}^{l^*(j)}\right)\alpha_{(2),j} + \left(x_{2,j+1}^{l(j+1)} - x_{2,j+1}^{k(j+1)}\right)\alpha_{(2),j+1} + \cdots \\
&\qquad\quad + \left(x_{2,s-1}^{l(s-1)} - x_{2,s-1}^{k(s-1)}\right)\alpha_{(2),s-1}.
\end{aligned}$$

Therefore, we demonstrate successively that $\alpha_{h(\tau),(2),j} \to \alpha_{(2),j}$ for $j = s-1, s-2, \ldots, 1$. Finally, for l as in condition 6, we choose $T = T_l$ as above, so that $(T_l x_2^l \alpha_{h(\tau),(2),j} + b)|G_{h(\tau)} \to_{\mathscr{D}} (T_l x_2^l \alpha_{(2)} + b)|G$ and $G_{h(\tau)} \to_{\mathscr{D}} G$. □

REFERENCES

ANDERSON, D. A. and AITKEN, M. (1985). Variance component models with binary response: interviewer variability. *J. Roy. Statist. Soc. Ser. B* **47** 203–210.

BILLINGSLEY, P. (1986). *Probability and Measure*, 2nd ed. Wiley, New York.

BOCK, R. D. and AITKEN, M. (1981). Marginal maximum likelihood estimation of item parameters: applications of an EM algorithm. *Psychometrika* **46** 443–459.

BUTLER, S. M. and LOUIS, T. (1992). Random effects models with nonparametric priors. *Statist. in Medicine* **11** 1981–2000.

CONAWAY, M. R. (1990). A random effects model for binary data. *Biometrics* **46** 317–328.

DE LEEUW, J. and VERHELST, N. (1986). Maximum likelihood estimation in generalized Rasch models. *J. Ed. Statist.* **11** 183–196.

FEINBERG, S. E., BROMET, E. J., FOLLMAN, D. L., LAMBERT, D. and MAY, S. M. (1985). Longitudinal analysis of categorical epidemiological data: a study of Three Mile Island. *Environmental Health Perspectives* **63** 241–248.

FOLLMAN, D. A. and LAMBERT, D. (1989). Generalizing logistic regression by nonparametric mixing. *J. Amer. Statist. Assoc.* **84** 295–300.

FOLLMAN, D. A. and LAMBERT, D. (1991). Identifiability of finite mixtures of logistic regression models. *J. Statist. Plann. Inference* **84** 295–300.

GILMORE, A. R., ANDERSON, R. D. and RAE, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72** 593–599.

HARVILLE, D. A. and MEE, R. W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics* **40** 393–408.

IM, S. and GIANOLA, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Appl. Statist.* **37** 196–204.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

KORN, E. L. and WHITTEMORE, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35** 795–804.

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

LAIRD, N. M. and WARE, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38** 963–974.

LINDSAY, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86–94.

LINDSAY, B. G., CLOGG, C. C. and GREGO, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model in item analysis. *J. Amer. Statist. Assoc.* **86** 96–107.

MARITZ, J. S. and LWIN, T. (1989). *Empirical Bayes Methods*, 2nd ed. Chapman and Hall, New York.

MISLEVY, R. J. (1985). Estimation of latent group effects. *J. Amer. Statist. Assoc.* **80** 993–997.

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain parametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137–158.

STIRATELLI, R., LAIRD, N. M. and WARE, J. H. (1984). Random effects models for serial observations with binary response. *Biometrics* **40** 961–971.

TALLIS, G. M. (1969). The identifiability of mixtures of distributions. *J. Appl. Probab.* **6** 389–398.

TALLIS, G. M. and CHESSON, P. (1982). Identifiability of mixtures. *J. Austral. Math. Soc. Ser. A* **32** 339–348.

TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.

VAN DER VAART, A. W. and WELLNER, J. A. (1992). Existence and consistency of maximum likelihood in upgraded mixture models. *J. Multivariate Anal.* **43** 133–146.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

ZEGER, S. L., LIANG, K. and ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44** 1049–1060.

DEPARTMENT OF BIOSTATISTICS
GENENTECH, INC.
1600 GRANDVIEW DRIVE
SOUTH SAN FRANCISCO, CALIFORNIA 94080
E-MAIL: butler.steve@gene.com

DIVISION OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA