PSYCHOMETRIKA—VOL. 59, NO. 1, 21–47
MARCH 1994

THE STATISTICAL ANALYSIS OF GENERAL PROCESSING TREE MODELS WITH THE EM ALGORITHM

XIANGEN HU AND WILLIAM H. BATCHELDER

UNIVERSITY OF CALIFORNIA, IRVINE

Multinomial processing tree models assume that an observed behavior category can arise from one or more processing sequences represented as branches in a tree. These models form a subclass of parametric, multinomial models, and they provide a substantively motivated alternative to loglinear models. We consider the usual case where branch probabilities are products of nonnegative integer powers in the parameters, $0 \le \theta_s \le 1$, and their complements, $1 - \theta_s$. A version of the EM algorithm is constructed that has very strong properties. First, the E-step and the M-step are both analytic and computationally easy; therefore, a fast PC program can be constructed for obtaining MLEs for large numbers of parameters. Second, a closed form expression for the observed Fisher information matrix is obtained for the entire class. Third, it is proved that the algorithm necessarily converges to a local maximum, and this is a stronger result than for the exponential family as a whole. Fourth, we show how the algorithm can handle quite general hypothesis tests concerning restrictions on the model parameters. Fifth, we extend the algorithm to handle the Read and Cressie power divergence family of goodness-of-fit statistics. The paper includes an example to illustrate some of these results.

Key words: EM algorithm, multinomial models, processing trees, power divergence family.

Introduction

Batchelder and Riefer (1986, 1990; Batchelder, Hu, & Riefer, in press; Riefer & Batchelder, 1988, 1991; Riefer & Rouder, 1992) define and provide many examples of a class of substantive multinomial models for cognitive psychology. Members of this class are called multinomial processing tree models, and other researchers have studied models that fit into this class (e.g., Bäuml, 1991; Chechile & Meyer, 1976; Humphreys & Bain, 1983; Ross & Bower, 1981). A large subclass consists of discrete-state Markov learning models, for finite segments of error and success trials (e.g., Wickens, 1982). Riefer and Batchelder (1988) argue that models in this class are useful as measurement tools in cognitive psychology. They occupy a position on an ordering of theoretical specificity between general purpose, off-the-shelf approaches, like ANOVA and loglinear models, and stronger theoretical models that are designed to explain data from many different paradigms, for example, artificial intelligence models such as SOAR (Rosenbloom, Laird, Newell, & McCarl, 1991) or neural network models (Rumelhart & McClelland, 1986).

Processing tree models assume that observed category counts arise from processing branches consisting of separate conditional links or stages. Each branch probability is the product of its conditional link probabilities, and more than one branch can terminate in the same observed category. Because of their structure, processing tree models are similar to the class of tree models in statistical genetics used to infer gene frequencies from phenotypic category frequencies, such as the well-known multinomial

This research was supported by National Science Foundation Grant BNS-8910552 to William H. Batchelder and David M. Riefer. We are grateful to David Riefer for his useful comments, and to the Institute for Mathematical Behavior Sciences for its support.

Requests for reprints and computer programs should be sent to William H. Batchelder, School of Social Sciences, University of California, Irvine, CA 92717.

0033-3123/94/0300-92433$00.75/0

© 1994 The Psychometric Society

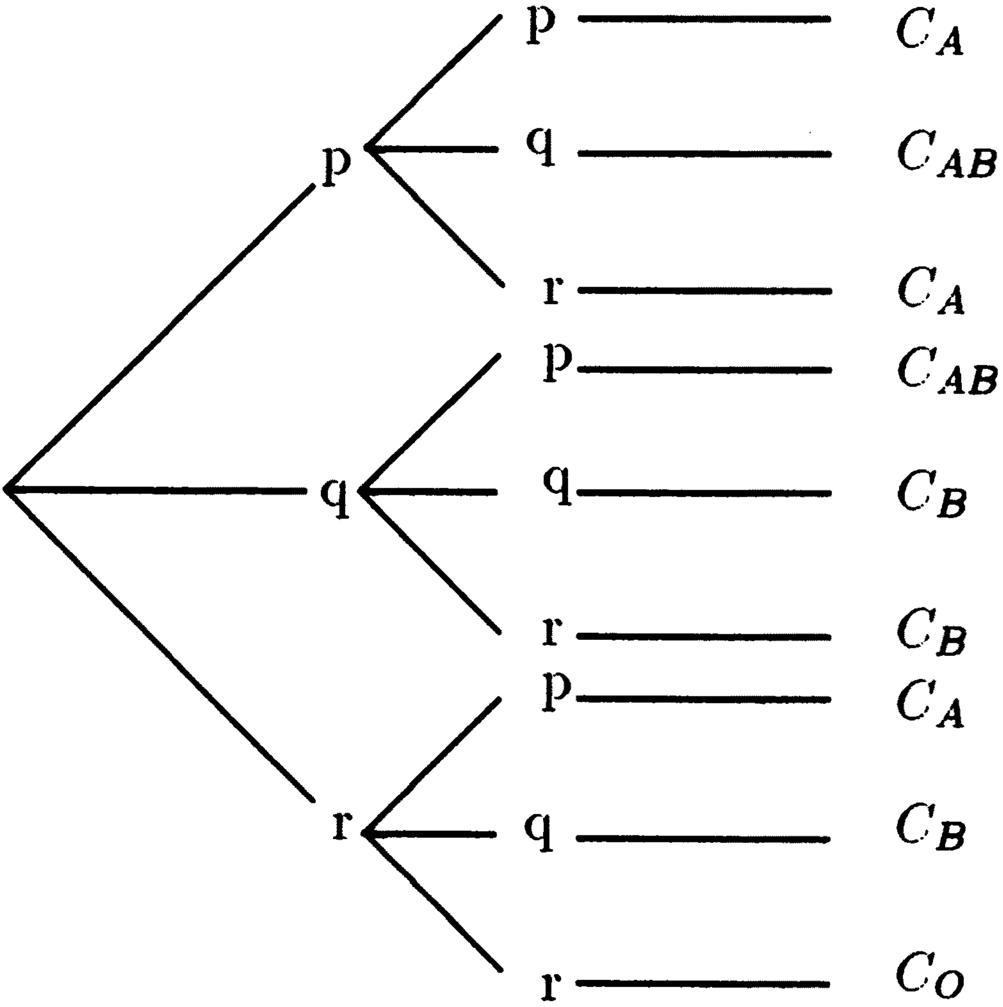

FIGURE 1.
Tree representation of the ABO blood group.

model for the ABO blood group (Bernstein, 1925; Landsteiner, 1901) described in Figure 1. In the ABO blood group model of Figure 1, there are four phenotypic categories, C_A, C_B, C_{AB}, C_O, corresponding to the four blood types that can be detected by medical procedures. The tree has nine two-link branches corresponding to the nine combinations of mother's gene (A, B or O) with father's gene. In the model, O is a recessive gene and A, B are dominant. The parameters p, q, and r refer to the equilibrium proportions of A, B, and O genes, respectively, in the population, where $p + q + r = 1$. For more detail and other examples of multinomial tree models in statistical genetics, see Elandt-Johnson (1971) and Weir (1990).

The purpose of this article is to advance the statistical analysis of processing tree models by showing how parameter estimation and hypothesis testing based on the likelihood function can be accomplished simply by adapting the so-called EM (expectation-maximization) algorithm (Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987; Rubin, 1991). The adaptation utilizes special properties of the class of models that make it easy to program a small PC to accomplish rapidly all stages of statistical inference. Also strong convergence results are obtained, and we show how the algorithm can be extended to cover estimation based on the Read and Cressie (1988) power divergence family of goodness-of-fit measures.

The EM algorithm provides an iterative method of obtaining maximum likelihood estimates (MLEs) for a model where some data may be regarded as "missing". The algorithm is particularly useful when three circumstances hold: (a) the process of obtaining MLEs for the model with the observed data is computationally intensive; (b) analytic expressions exist for obtaining the conditional expectation of the missing data given the observed data as a function of the model parameters (the E-step); and (c) given values of both the observed and missing data, analytic expressions exist for the MLEs of the model parameters (the M-step). The algorithm starts with initial parameter estimates and then alternates the E-step and M-step until, hopefully, the sequence of resulting parameter estimates converges to an MLE.

The general idea of applying the EM algorithm to categorical data where some observations can be only partially categorized goes back at least to Hartley (1958). Before Hartley and others developed the general approach to categorical data, statistical geneticists (e.g., Ceppellini, Siniscalco, & Smith, 1955; Smith, 1957) described a related method called "gene counting" for estimating gene frequency parameters for certain multinomial tree models (see Weir, 1990).

In the first section of the paper, we define a general class of parametric multinomial models motivated by processing tree representations. The models involve observed categories that may be subdivided into unobserved categories. Next, we set up the likelihood function in terms that treat the unobserved category counts as missing data subject to constraints in terms of the observed category counts. This representation permits us to develop a special version of the EM algorithm. The analytic specificity of our class of models permits us to prove stronger convergence results and obtain better computational efficiency than the established results with the EM algorithm that hold for arbitrary members of the exponential family. Computationally easy methods of obtaining an estimate of the variance-covariance matrix and of testing a wide class of hypotheses are provided. Finally, our methods are extended to handle general members of the Read and Cressie (1988) power divergence family of goodness-of-fit measures.

The EM Algorithm for General Processing Tree Models

Notation

Any processing tree model consists of a set of branches that are partitioned into observable categories. Let $C_1, \ldots, C_J$ denote the observable categories and $B_{1j}, \ldots, B_{ij}, \ldots, B_{I_j j}$ denote the collection of branches that terminate in category C_j. Let $\mathbf{N}_j$ and $\mathbf{M}_{ij}$ be random variables denoting the counts in category C_j and branch B_{ij}, respectively, and let n_j, m_{ij} be their corresponding realizations, where

$$\sum_{j=1}^{J} n_j = n, \qquad \sum_{i=1}^{I_j} m_{ij} = n_j. \tag{1}$$

Denote the model parameters by $\Theta = (\theta_1, \ldots, \theta_S) \in \Omega$, where Ω is the parameter space and the θ_s are functionally independent with $0 \leq \theta_s \leq 1$, $s = 1, \ldots, S$. The class of models under investigation requires that each branch in the tree consists of a sequence of links, where the branch probability is a product of the corresponding link probabilities. To achieve a wide application of our methods, we require that link probabilities be written as a product of a positive constant and nonnegative integer powers in the model's parameters and their complements (the θ_s and $1 - \theta_s$, respectively). In this case the branch probabilities take the form

$$p_{ij}(\Theta) = \Pr\,(B_{ij};\, \Theta) = c_{ij} \prod_{s=1}^{S} \theta_s^{a_{ijs}}(1 - \theta_s)^{b_{ijs}}, \tag{2}$$

where c_{ij} is the product of positive constants on the links and the a_{ijs} and b_{ijs} are the sum over links of nonnegative integer exponents on θ_s and $(1 - \theta_s)$, respectively. The possibility of $c_{ij} \neq 1$ can arise from the model itself, as in the Batchelder and Riefer (1986) model for storage and retrieval, or it can arise from hypothesis restrictions on some parameters or from combining trees (discussed later).

The category probabilities can be obtained from (2) as

$$p_j(\Theta) = \Pr\,(C_j;\, \Theta) = \sum_{i=1}^{I_j} p_{ij}(\Theta) = \sum_{i=1}^{I_j} c_{ij} \prod_{s=1}^{S} \theta_s^{a_{ijs}}(1 - \theta_s)^{b_{ijs}}, \tag{3}$$

where

$$\sum_{j=1}^{J} p_j(\Theta) = 1, \tag{4}$$

for all $\Theta \in \Omega = [0,\, 1]^S$.

The requirements in (2) that link probabilities be in the form of

$$c \prod_{s=1}^{S} \theta_s^a(1 - \theta_s)^b, \tag{5}$$

and that parameters have the full range [0, 1] may seem overly restrictive. For example, the ABO model in Figure 1 violates both of these conditions, since the functionally independent parameters are p and q, with $0 \leq p + q \leq 1$, and further, $r = 1 - p - q$ is not in the form of (5). However, it is easy to reparameterize this model with $\theta_A = p$, $\theta_B = q/(1 - p)$, to satisfy these constraints. Then $0 \leq \theta_A, \theta_B \leq 1$ and $r = 1 - p - q = (1 - \theta_A)(1 - \theta_B)$, and it is a simple exercise to show that the model fits the form of (2), (3), and (4). It is easy to generalize this type of reparameterization in the ABO model to cases where, at each choice point, all but one link probabilities are single parameters; however, we do not develop the issue of permissible reparameterizations to its fullest extent in this paper.

It should be noted that due to the sum in (3), models in this class are not, in general, loglinear models. While some of the models represented by (3) can be reparameterized to be equivalent to loglinear models without increasing the number of parameters, others can not be so reparameterized. One consequence of this is that current software packages for categorical data like GLIM and SAS can not be used conveniently to analyze this class. This is one of the main motivations for presenting the methods in this paper.

The motivation for the representation leading to (2) and (3) comes from the processing tree representation; however, the methods we develop based on the EM algorithm apply to any multinomial model that can be written in that form. We call these models general processing tree (GPT) models, and the next definition formalizes this class.

Definition 1. Let

$$\mathcal{M}(\Theta;\ \langle c_{ij}\rangle;\ \langle a_{ijs}\rangle;\ \langle b_{ijs}\rangle) \tag{6}$$

be a parametric multinomial model defined over J observable categories with S functionally independent parameters, $\Theta = (\theta_1, \dots, \theta_S)$. Then $\mathcal{M}$ is a general processing tree model in case there are positive integers I_j; nonnegative integers a_{ijs} and b_{ijs}; and positive reals c_{ij} so that the category probabilities $p_j(\Theta)$ can be written in the form of (3) and (4), where $i = 1, \dots, I_j$, $j = 1, \dots, J$, and $s = 1, \dots, S$.

Trivially, any GPT model can be written as a tree with $\sum_{j=1}^{J} I_j$ one-link branches, partitioned appropriately into J equivalence classes corresponding to the observable categories. This particular tree representation is uninteresting, but it may admit to equivalent, structurally interpretable GPT models such as those cited earlier from the literature. The issue of equivalent GPT models is interesting, but we do not investigate it here because our results depend only on the representation in (3). The GPT representation includes a number of models not normally thought of as tree structures. For example, finite latent class models for dichotomous questions (see Lazarsfeld & Henry, 1968) can be parameterized to fit Definition 1.

For the remainder of the paper, we avoid triviality by adapting two conventions regarding Definition 1. First, we will require that each parameter θ_s has at least one branch, (i, j), where $a_{ijs} + b_{ijs} > 0$; thus, the collection of $p_{ij}(\Theta)$ will not be constant in any parameter. Second, we include fixed probability distributions over J categories as GPT models with no parameters.

The class of GPT models has several nice properties that we can utilize in constructing an EM algorithm. Observation 1 shows that the class is closed under setting arbitrary subsets of the parameters to constants in [0, 1].

Observation 1. Let

$$\mathcal{M}(\Theta;\ \langle c_{ij}\rangle;\ \langle a_{ijs}\rangle;\ \langle b_{ijs}\rangle)$$

be a GPT model with parameters $\Theta = (\theta_1, \dots, \theta_S)$. Suppose a subset of the parameters are set to particular constants in [0, 1]. Then the resulting submodel is itself a GPT model.

(All proofs will be in Appendix A.)

The class of GPT models has some strong identifiability properties that we need in constructing our EM algorithm.

Definition 2. A GPT model $\mathcal{M}$ is identifiable if $p_j(\Theta) = p_j(\Theta')$, for all $j = 1, \dots, J$, implies $\Theta = \Theta'$, for all $\Theta, \Theta' \in (0, 1)^S$, that is, the interior of Ω.

The next observation shows that if the GPT model has observable branches ($I_j = 1$ for all categories), then it necessarily has fewer than J parameters, which is a necessary condition for identifiability.

Observation 2. Suppose $\mathcal{M}$ is a GPT model on J categories with S functionally independent parameters, $\Theta = (\Theta_1, \dots, \Theta_S) \in [0, 1]^S$. Then if the model has observable branches, that is $I_j = 1$, for all j, it follows that $S < J$.

Observation 2 has deep implications because it shows that the structure of (2) and (4) impose considerable limitations on the constructability of GPT models. In fact, when coupled with Observation 3 to follow, it shows that a GPT model with observable branches is necessarily identifiable. Of course, this result is not true for more general classes of parametric models. For example, consider a model for a single flip of a coin where the probability of a head is given by $p = ab^2(1 - b)$, for all $0 \leq a, b \leq 1$. In this case, the branches are observable and a and b are functionally independent, with $\Omega = [0, 1]^2$; however, the model is nonidentifiable, and of course, it is not a GPT model since $q = 1 - p = 1 - ab^2(1 - b)$ does not fit the structural form of (2), and it cannot be reparameterized as such.

A direct corollary of Observation 2 is that the number of parameters in any GPT model is less than $\sum_{j=1}^{J} I_j$. Thus, even when branches are combined, there are limitations on the number of parameters for a GPT model.

Likelihood Functions

Suppose one observes the category counts n_j. Then the likelihood function is given by

$$L(\Theta; \langle n_j \rangle_{j=1}^{J}) = N! \prod_{j=1}^{J} \frac{[p_j(\Theta)]^{n_j}}{n_j!}, \tag{7}$$

where the $p_j(\Theta)$ are given by (3). Because of the sum in (3), it often happens that analytic methods cannot be used to obtain MLEs, so that some sort of iterative algorithm must be employed.

Suppose one has the "missing" branch frequencies,

$$\mathbf{D} = \langle \langle m_{ij} \rangle_{i=1}^{I_j} \rangle_{j=1}^{J},$$

then the complete likelihood function becomes

$$L(\Theta; \mathbf{D}) = N! \prod_{j=1}^{J} \prod_{i=1}^{I_j} \frac{[p_{ij}(\Theta)]^{m_{ij}}}{m_{ij}!}, \tag{8}$$

where the $p_{ij}(\Theta)$ are given by (2).

Observation 3. Equation (8) and

$$\sum_{j=1}^{J} \sum_{i=1}^{I_j} a_{ijs} m_{ij} \neq 0, \quad \sum_{j=1}^{J} \sum_{i=1}^{I_j} b_{ijs} m_{ij} \neq 0, \tag{9}$$

for all $s = 1, \ldots, S$, has an unique maximum in the interior of $[0, 1]^S$, given by

$$\hat{\theta}_s = \frac{\sum_{j=1}^{J} \sum_{i=1}^{I_j} a_{ijs} m_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{I_j} (a_{ijs} + b_{ijs}) m_{ij}}. \tag{10}$$

A direct corollary of Observation 3 is that a GPT model with observable branches is necessarily identifiable. This is implied by the uniqueness in (10) coupled with the consistency property of MLEs.

The EM Algorithm

In the remaining sections, unless otherwise stated, we assume that Θ is in the interior of Ω. To construct an EM algorithm for a GPT model, we consider the $\langle n_j \rangle$ as the observed data and $\mathbf{D}$, subject to (1), as the missing data. Then (10) provides the M-step for the algorithm. The E-step is also straightforward, because for any value of Θ, the conditional expected value of m_{ij} is given by

$$m_{ij}(\Theta) = E(\mathbf{M}_{ij} | n_j;\ \Theta) = \frac{n_j p_{ij}(\Theta)}{p_j(\Theta)}, \tag{11}$$

where $p_{ij}(\Theta)$ and $p_j(\Theta)$ are given by (2) and (3), respectively. Any EM algorithm defines a function $\Phi = (\phi_1, \ldots, \phi_S) = M(\Theta)$, where Θ is a parameter vector entering a cycle of the algorithm and Φ is the revised estimate after the cycle. The function $M(\Theta)$ can be obtained by substituting $m_{ij}(\Theta)$ from (11) for the m_{ij} in (10), then the result is

$$\phi_s = \frac{\sum_{j=1}^{J} \sum_{i=1}^{I_j} a_{ijs} m_{ij}(\Theta)}{\sum_{j=1}^{J} \sum_{i=1}^{I_j} (a_{ijs} + b_{ijs}) m_{ij}(\Theta)}, \tag{12}$$

for $s = 1, \ldots, S$. So, the EM algorithm can be easily summarized as

$$\Theta^{(n+1)} = M(\Theta^{(n)}). \tag{13}$$

Dempster et al. (1977) show that a cycle of the EM algorithm does not decrease the likelihood function, that is,

$$L(M(\Theta);\ \langle n_j \rangle) \geq L(\Theta;\ \langle n_j \rangle), \tag{14}$$

for all Θ, so that there is reason to hope that repeated applications may maximize (7). A later section is devoted to the issue of the convergence of our EM algorithm. For now, we assume that we have an identifiable model and that the algorithm obtains the unique MLE $\hat{\Theta}$, and we proceed to obtain confidence intervals and confidence regions for the parameter estimators.

Estimating Confidence Intervals

Confidence intervals for each parameter estimator can be obtained from an estimator of its variance. In the literature, efforts have been made to get the variance-covariance matrix or Fisher information matrix for the EM algorithm in general settings. However, these earlier approaches require additional numerical computations, for example, Louis (1982) and Meng and Rubin (1991). However, for GPT models it is possible to derive a closed-form expression for the *observed* Fisher information matrix in the context of the EM algorithm. To simplify the notation, we define the random variables,

$$\mathbf{A}_s(\Theta) = \sum_{j=1}^{J} \mathbf{N}_j \sum_{i=1}^{I_j} \left[a_{ijs} \frac{p_{ij}(\Theta)}{p_j(\Theta)} \right],$$

$$\mathbf{B}_s(\Theta) = \sum_{j=1}^{J} \mathbf{N}_j \sum_{i=1}^{I_j} \left[b_{ijs} \frac{p_{ij}(\Theta)}{p_j(\Theta)} \right],$$

and their realizations $\alpha_s(\Theta)$ and $\beta_s(\Theta)$, which are obtained by substituting n_j for the $\mathbf{N}_j$.

By definition, the observed Fisher information matrix is

$$\left. \left(\frac{\partial^2 \ln L(\Theta; \langle n_j \rangle_{j=1}^{J})}{\partial \theta_s \partial \theta_r} \right)_{S \times S} \right|_{\Theta = \hat{\Theta}}, \tag{15}$$

(see Efron & Hinkley, 1978). From (3) and (7), it is easy to obtain

$$\frac{\partial \ln L(\hat{\Theta}; \langle n_j \rangle_{j=1}^{J})}{\partial \theta_s} = \sum_{j=1}^{J} n_j \sum_{i=1}^{I_j} \left[\frac{a_{ijs} p_{ij}(\hat{\Theta})}{\hat{\theta}_s p_j(\hat{\Theta})} - \frac{b_{ijs} p_{ij}(\hat{\Theta})}{(1 - \hat{\theta}_s) p_j(\hat{\Theta})} \right],$$

and using $\alpha_s(\Theta)$ and $\beta_s(\Theta)$,

$$\frac{\partial \ln L(\hat{\Theta}; \langle n_j \rangle)}{\partial \theta_s} = \left[\frac{\alpha_s(\hat{\Theta})}{\hat{\theta}_s} - \frac{\beta_s(\hat{\Theta})}{1 - \hat{\theta}_s} \right].$$

Since

$$\begin{aligned} \frac{\partial \alpha_s(\hat{\Theta})}{\partial \theta_r} = -\sum_{j=1}^{J} n_j \Bigg[& \sum_{i=1}^{I_j} a_{ijs} \frac{\displaystyle\sum_{k=1}^{I_j} \left[\frac{a_{kjr}}{\hat{\theta}_r} - \frac{b_{kjr}}{1 - \hat{\theta}_r} \right] p_{kj}(\hat{\Theta})}{p_j^2(\hat{\Theta})} \\ & - \sum_{i=1}^{I_j} a_{ijs} \frac{\left[\frac{a_{ijr}}{\hat{\theta}_r} - \frac{b_{ijr}}{1 - \hat{\theta}_r} \right] p_{ij}(\hat{\Theta})}{p_j(\hat{\Theta})} \Bigg], \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \beta_s(\hat{\Theta})}{\partial \theta_r} = -\sum_{j=1}^{J} n_j \Bigg[& \sum_{i=1}^{I_j} b_{ijs} \frac{\displaystyle\sum_{k=1}^{I_j} \left[\frac{a_{kjr}}{\hat{\theta}_r} - \frac{b_{kjr}}{1 - \hat{\theta}_r} \right] p_{kj}(\hat{\Theta})}{p_j^2(\hat{\Theta})} \\ & - \sum_{i=1}^{I_j} b_{ijs} \frac{\left[\frac{a_{ijr}}{\hat{\theta}_r} - \frac{b_{ijr}}{1 - \hat{\theta}_r} \right] p_{ij}(\hat{\Theta})}{p_j(\hat{\Theta})} \Bigg], \end{aligned}$$

element (r, s) of (15) becomes

$$\frac{\partial^2 \ln L(\hat{\Theta}; \langle n_j \rangle)}{\partial \theta_s \partial \theta_r} = - \begin{cases} \dfrac{\dfrac{\partial \alpha_s(\hat{\Theta})}{\partial \theta_r}}{\hat{\theta}_s} - \dfrac{\dfrac{\partial \beta_s(\hat{\Theta})}{\partial \theta_r}}{1 - \hat{\theta}_s}, & \text{if } r \neq s \\ \dfrac{\dfrac{\partial \alpha_s(\hat{\Theta})}{\partial \theta_s}}{\hat{\theta}_s} - \dfrac{\dfrac{\partial \beta_s(\hat{\Theta})}{\partial \theta_s}}{1 - \hat{\theta}_s} - \left[\dfrac{\alpha_s(\hat{\Theta})}{\hat{\theta}_s^2} + \dfrac{\beta_s(\hat{\Theta})}{(1 - \hat{\theta}_s)^2} \right], & \text{if } r = s. \end{cases} \tag{16}$$

By inverting the observed Fisher information matrix, one obtains an estimate of the variance-covariance matrix of the estimators; therefore, confidence intervals or confidence regions for the estimators can be obtained as well.

Convergence of the EM Algorithm for GPT Models

In this section we explore convergence for the class of GPT models from a theoretical perspective. There are many discussions about the convergence of the EM algorithm (e.g., Boyles, 1983; Dempster et al., 1977; Little & Rubin, 1987; Wu, 1983). These papers have obtained some theorems about the convergence of the EM algorithm in a general setting, for example (14), as well as some useful theorems for the exponential family. Processing tree models are members of the exponential family; however, from the general theorems, even a convergent EM sequence only guarantees a stationary point of the likelihood function. This makes the convergence properties of the EM algorithm not different from ordinary optimization algorithms for finding MLEs (e.g., conjugate-gradient). The guaranteed convergence of the EM sequence to a local maximum can be obtained by adding some restricted conditions, such as Theorem 4 of Dempster et al. (1977) or the condition in Theorem 4 of Wu (1983). However, these conditions are not easy to verify in general. Nevertheless, when we restrict consideration to the class of GPT models, we obtain the stronger convergence result in the next observation.

Observation 4. If Θ^* is the limit of some EM sequence $\langle \Theta^{(n)} \rangle$, then Θ^* is a (at least local) maxima of (7).

The results of Observation 4 guarantee that any convergent EM sequence will converge to a local maxima of (7). Later in this paper, we will introduce a simple trick that guarantees a convergent sequence regardless of the starting value of the EM algorithm. To assure a global maxima, different starting values should be tried. However, in practice we have found that for most identifiable models and data sets, the same (global) maximum is reached for a given data set regardless of the starting value. It is clear from (2) and (3) that (7) defines a polynomial over the entire S-dimensional reals, and for most models in practice, it is a polynomial of low degree. Thus it does not seem surprising to us that in most cases the polynomial would have only a single maxima in $(0, 1)^S$. We consider it an interesting open problem to try to obtain some precise results on this point.

Testing Hypotheses

Testing statistical hypotheses is an important part in any modeling effort. By using the EM algorithm developed in previous sections, some hypothesis testing problems become very easy. In particular, any hypothesis testing task involving two nested models can be done within our EM framework, if each model can be formulated in terms of a GPT model, where the parameters of the restricted model are a subset of the parameters of the general model. In this case, the EM algorithm can be used to obtain the MLEs for each model, and the rest is accomplished by standard methods such as the likelihood ratio criterion. In general, when parameter restrictions are imposed on a GPT model to create a nested model, link probabilities may no longer satisfy (5) in the unrestricted parameters; yet it may be possible to redesign the tree to satisfy this property. Observation 5 establishes some very general conditions when this is possible.

Observation 5. Let $\mathcal{M}(\Theta; \langle c_{ij} \rangle; \langle a_{ijs} \rangle; \langle b_{ijs} \rangle)$ be a GPT model as described in Definition 1, with $\Theta = (\theta_1, \ldots, \theta_S)$. Suppose the first z parameters, $\Theta' = (\theta_1, \ldots, \theta_z)$, are specified by the remaining $S - z$ parameters, $\Theta'' = (\theta_{z+1}, \ldots, \theta_S)$, by a function $\Theta' = (f_1(\Theta''), \ldots, f_z(\Theta''))$. Then the restricted model with parameter Θ'' is a GPT model if each $f_s(\Theta'')$, $s = 1, \ldots, z$, is in one of the following forms:

i. $f_s(\Theta'') = \alpha$, $\alpha \in (0, 1)$.
ii. $f_s(\Theta'') = \Pi_{k=z+1}^{S} \theta_k^{\mu_k}$, $\mu_k \in \{0, 1\}$, $k = z + 1, \ldots, S$.
iii. $f_s(\Theta'') = \Sigma_{k=z+1}^{S} \nu_k \theta_k^{\rho_k}$,

where the ρ_k are positive integers and the ν_k nonnegative reals, with $\Sigma_{k=z+1}^{S} \nu_k = 1$.

Observation 5 covers many of the ways that nested models are created from a general model. For example, the case of equating a parameter to a constant is covered by Item i and constraining different parameters to equal each other is covered as a special case of Item ii with all but one of the μ_k zero. Also Items ii and iii are quite general restrictions based on products and weighted sums respectively. In the unlikely case that the nested model of interest falls outside the scope of Observation 5, it will be necessary to use conventional iterative methods to get MLEs of the restricted model.

Another important case of hypothesis testing arises in the case of a between-group design, where the same multinomial model is thought to apply to each group separately with possibly different parameter values. The resulting data structure is a product multinomial, where the experimenter controls the group marginals. The extension of our EM algorithm to joint multinomial models (discussed next) is required to handle this situation.

Joint Multinomial Models

In Riefer and Batchelder (1988), the class of multinomial models includes so-called joint multinomial models. These models have several category systems, with preset, fixed total counts for each system that are distributed over the categories within the system. Moreover, all the category probabilities are functions of the same parameter vector Θ. Such a system covers a within-subject design where, for example, different types of items result in different category systems (e.g., Batchelder & Riefer, 1986). A special case of a joint multinomial model occurs when data from a between-subject experiment exist and one assumes that the same processing tree model fits each group, with possible restrictions on the parameters across trees. If the underlying model has only a single category system, then the joint model for the between group case is technically called a product multinomial model (Read & Cressie, 1988).

It is straightforward to adapt the EM algorithm, designed earlier, to the case of joint multinomial models. Suppose there are K category systems, each arising from a GPT model, and each having probabilities in the form of (3) over the same parameter

Θ. Let C_{jk} denote category j in tree k, and let B_{ijk} be branch i that terminates in C_{jk}, where $i = 1, \ldots, I_{jk}$, $j = 1, \ldots, J_k$, and $k = 1, \ldots, K$. Denote the model parameters by $\Theta = (\theta_1, \ldots, \theta_S)$. For each category system, (2), (3), and (4) become

$$p_{ijk}(\Theta) = \Pr(B_{ijk}; \Theta) = c_{ijk} \prod_{s=1}^{S} \theta_s^{a_{ijks}} (1 - \theta_s)^{b_{ijks}}, \tag{17}$$

$$p_{jk}(\Theta) = \Pr(C_{jk}; \Theta) = \sum_{i=1}^{I_{jk}} p_{ijk}(\Theta), \tag{18}$$

and

$$\sum_{j=1}^{J_k} p_{jk}(\Theta) = 1, \tag{19}$$

for all $\Theta \in \Omega$ and $k = 1, \ldots, K$.

It is easy to join the K systems into a single GPT model with $\sum_{k=1}^{K} J_k$ categories. Let $\underline{c}_{ijk} = \tau_k c_{ijk}$ for positive real numbers τ_k such that $\sum_{k=1}^{K} \tau_k = 1$. Then (17) becomes

$$\underline{p}_{ijk}(\Theta) = \underline{c}_{ijk} \prod_{s=1}^{S} \theta_s^{a_{ijks}} (1 - \theta_s)^{b_{ijks}}, \tag{20}$$

which defines a GPT model with

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} \underline{p}_{jk}(\Theta) = 1,$$

where $\underline{p}_{jk}(\Theta) = \tau_k p_{jk}(\Theta)$ for all $\Theta \in \Omega$.

The construction leading to (20) can be used to analyze joint multinomial models with preset marginals. Define τ_k as the proportion of total observations in category system k. Then if the GPT model in (20) is analyzed with the EM algorithm, the resulting MLEs will be MLEs of the joint multinomial model. This is because the E-step involves only expected frequencies. Of course the joint multinomial model with preset category frequencies is not a GPT model; nevertheless, the next observation shows that both the MLE and Fisher information matrix of the constructed model in (20) apply to the joint multinomial model.

Observation 6. Assume that for $k = 1, \ldots, K$, $\mathcal{M}_k(\Theta; \langle c_{ijk} \rangle; \langle a_{ijks} \rangle; \langle b_{ijks} \rangle)$, $\Theta = (\theta_1, \ldots, \theta_S)$, is a GPT model with likelihood functions $L_k(\Theta; \langle n_{jk} \rangle_{j=1}^{J_k})$. Then the following two likelihood functions in (21) and (22) will give the same MLE $\hat{\Theta}$ and Fisher information matrix for Θ:

$$\prod_{k=1}^{K} L_k(\Theta; \langle n_{jk} \rangle_{j=1}^{J_k}), \tag{21}$$

where $L_k(\Theta; \langle n_{jk} \rangle_{j=1}^{J_k})$ is defined as in (8) for each k, and

TABLE 1
Group 3 × 3 Data Tables Constructed from Harvey (1985)

	Manic						Schizophrenic						Normal		
	NTD			TD			NTD			TD			subjects		
	(Group 1)			(Group 2)			(Group 3)			(Group 4)			(Group 5)		
item	S	T	N	S	T	N	S	T	N	S	T	N	S	T	N
Say	22	27	31	43	6	31	13	21	46	44	10	26	23	22	35
Think	7	54	19	20	15	45	4	42	34	32	8	40	9	45	26
New	4	26	50	5	9	66	6	20	54	24	7	49	7	10	63

Note: NTD=non-thought disordered; TD=thought disordered; responses are as follows: S=Say; T=think; N=new. Each entry is the frequency of the column response to the row type of item Row marginals are preset by the experimenter.

$$\binom{N}{N_1 \cdots N_K} \prod_{k=1}^{K} L_k(\Theta; \langle n_{jk}\rangle_{j=1}^{J_k}) \left[\prod_{k=1}^{K-1} \tau_k^{N_k}\right]\left(1 - \sum_{k=1}^{K-1} \tau_k\right)^{N - \sum_{k=1}^{K-1} N_k}, \qquad (22)$$

where $\tau_1, \ldots, \tau_{K-1}$ are independent of Θ, N_k, $k = 1, \ldots, K$, are the observed totals for each category system, and $N = \sum_{k=1}^{K} N_k$.

One can notice that (21) is the true likelihood function obtained simply by multiplying the individual likelihood functions together, while (22) is obtained in such a way that it corresponds to the likelihood function of a joint GPT, with new parāmeters τ_k, $k = 1, \ldots, K - 1$. Clearly the MLE of τ_k will be the experimenter determined proportions, N_k/N, so the thrust of Observation 6 is that the preset N_k do not affect the variance-covariance matrix of the estimator of Θ.

An Example

Batchelder and Riefer (1990) studied a family of multinomial processing tree models for memory experiments involving source monitoring. In this section, we will apply some of the results from the previous sections to analyze a set of data with these models. Table 1 presents data obtained from Harvey (1985) and analyzed in Batchelder and Riefer (1990). Harvey studied manic and schizophrenia patients in a standard reality-monitoring task (Johnson & Ray, 1988). In this experiment, various groups of subjects first studied a list of written words and for each word were asked either to say the word aloud (say) or think the word (think). Then they were tested with old say and think words along with new distractors (new) and they were asked to correctly classify each word by responding "say", "think", or "new". There were five groups of subjects: Manic, nonthought disordered (NTD); Manic, thought disordered (TD); Schizophrenic NTD; Schizophrenic TD; and Normal subjects.

The model postulates three processing trees depicted in Figure 2 corresponding to say items (Source 1, S_1), think items (S_2), and new items (S_3). Each branch terminates in one of three responses: R_1 (say), R_2 (think), and R_3 (new). The model is a joint multinomial model because there are three category systems, each with (the same) three response categories. In our notation for (17), (18), and (19), $K = 3$, $J_1 = J_2 =$

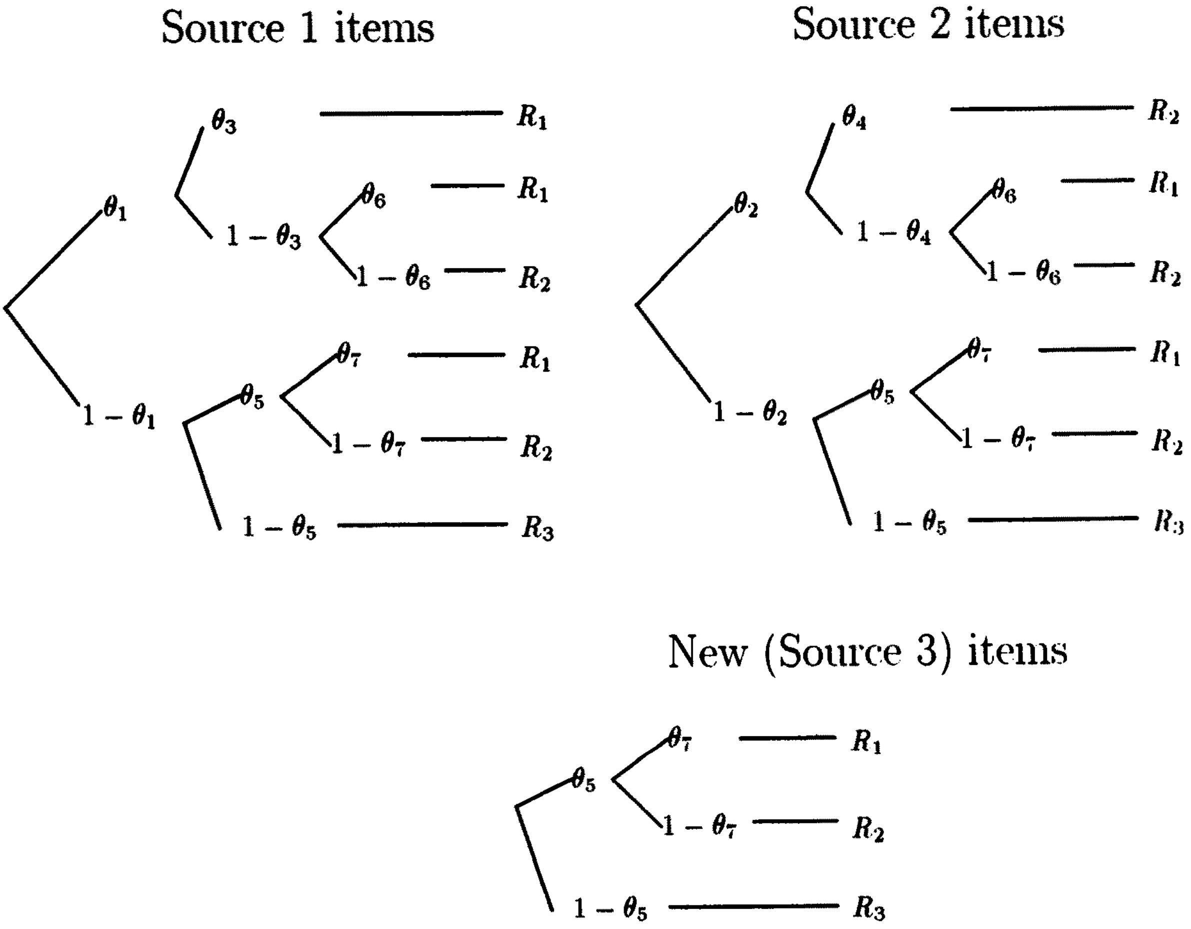

FIGURE 2.
Tree representation for the source monitoring model.

$J_3 = 3$, and $I_{11} = 3$, $I_{21} = 2$, $I_{31} = 1$, etcetera. In keeping with the notation in the paper, we denote the parameters by $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7)$; however, Batchelder and Riefer (1990) use different symbols, namely, D_1, D_2, d_1, d_2, b, a, g, respectively. To motivate the example, θ_1 and θ_2 are the probabilities of correctly detecting as old say and think items, respectively; θ_3 and θ_4 are conditional probabilities of correctly discriminating the source of detected say and think items, respectively; θ_5 is a false alarm probability for nondetected items; and θ_6 and θ_7 are guessing biases for say items that occur in different places in the tree (see Batchelder & Riefer, 1990, for more concerning the substantive details).

There are a total of fifteen branches in the three trees in Figure 2, and the branch probabilities are products of the link probabilities. The experimenter determines the proportions of items presented to the subjects in each category system, and here we denote them by the constants τ_1, τ_2, τ_3, with $\tau_1 + \tau_2 + \tau_3 = 1$. The three trees can be joined as described in (20), and it is easy to obtain the nine category probabilities, $\underline{p}_{jk}(\Theta)$, from Figure 2 as follows:

$$P(S_1, R_1) = \underline{p}_{11} = \tau_1[\theta_1\theta_3 + \theta_1(1 - \theta_3)\theta_6 + (1 - \theta_1)\theta_5\theta_7], \tag{23}$$

$$P(S_1, R_2) = \underline{p}_{21} = \tau_1[\theta_1(1 - \theta_3)(1 - \theta_6) + (1 - \theta_1)\theta_5(1 - \theta_7)],$$

$$P(S_1, R_3) = \underline{p}_{31} = \tau_1(1 - \theta_1)(1 - \theta_5),$$

$$P(S_2, R_1) = \underline{p}_{12} = \tau_2[\theta_2(1 - \theta_4)\theta_6 + (1 - \theta_2)\theta_5\theta_7],$$

$$P(S_2, R_2) = \underline{p}_{22} = \tau_2[\theta_2\theta_4 + \theta_2(1 - \theta_4)(1 - \theta_6) + (1 - \theta_2)\theta_5(1 - \theta_7)],$$

$$P(S_2, R_3) = \underline{p}_{32} = \tau_2(1 - \theta_2)(1 - \theta_5),$$

$$P(S_3, R_1) = \underline{p}_{13} = \tau_3\theta_5\theta_7,$$

$$P(S_3, R_2) = \underline{p}_{23} = \tau_3\theta_5(1 - \theta_7),$$

$$P(S_3, R_3) = \underline{p}_{33} = \tau_3(1 - \theta_5).$$

The corresponding joint tree for (23) has 15 branches leading to 9 categories that jointly describe each source-response pair. It is straightforward to obtain the structural constants for this tree, namely, the $\langle a_{ijks} \rangle$, $\langle b_{ijks} \rangle$, and $\langle \underline{c}_{ijk} \rangle$, for example, in the top equation in (23) there are three branches leading to category $j = 1$ in tree $k = 1$. Numbering these branches from left to right in (23) yields the structural constants

$$a_{1111} = a_{1113} = a_{2111} = a_{2116} = a_{3115} = a_{3117} = 1,$$

$$b_{2113} = b_{3111} = 1,$$

$$\underline{c}_{111} = \underline{c}_{211} = \underline{c}_{311} = \tau_1,$$

and the rest of the a_{i11s} and b_{i11s} equal zero, for $i = 1, 2, 3$ and $s = 1, \ldots, 7$.

Notice that (23) has 7 parameters and there are only 6 degree of freedom in the data (9 observed frequencies with 3 fixed marginals), so the GPT model in (23) is not identifiable. Batchelder and Riefer (1990) described a number of nested versions of the model that are identifiable. To illustrate, with the additional assumptions $\theta_3 = \theta_4 = \theta_3^*$ and $\theta_6 = \theta_7 = \theta_5^*$ along with $\theta_1 = \theta_1^*$, $\theta_2 = \theta_2^*$, $\theta_5 = \theta_4^*$ (see Figure 2), a testable and identifiable GPT model with structural constants $\langle a^*_{ijks} \rangle$, $\langle b^*_{ijks} \rangle$ can be obtained as follows:

$$a^*_{ijks} = a_{ijks}, \qquad s = 1, 2,$$

$$b^*_{ijks} = b_{ijks}, \qquad s = 1, 2,$$

$$a^*_{ijk3} = a_{ijk3} + a_{ijk4},$$

$$b^*_{ijk3} = b_{ijk3} + b_{ijk4},$$

$$a^*_{ijk4} = a_{ijk5},$$

$$b^*_{ijk4} = b_{ijk5},$$

$$a^*_{ijk5} = a_{ijk6} + a_{ijk7},$$

$$b^*_{ijk5} = b_{ijk6} + b_{ijk7},$$

$i = 1, \ldots, I_j, j = 1, 2, 3, k = 1, 2, 3$. Each $\underline{c}^*_{ijk}$ is given by τ_k where the ij-th branch comes from tree k. We denote the new model by

$$\mathcal{M}(\Theta^*; \langle \underline{c}^*_{ijk} \rangle; \langle a^*_{ijks} \rangle; \langle b^*_{ijks} \rangle), \tag{24}$$

TABLE 2
Parameter Estimates and Goodness-of-Fit Tests for Harvey's (1985) Experiment

	Group	Parameter estimate					Goodness-of-fit
		θ_1^*	θ_2^*	θ_3^*	θ_4^*	θ_5^*	$G^2(1)$
Group 1:	Manic NTD	.39	.62	.51	.37	.17	0.50
Group 2:	Manic TD	.53	.29	.43	.18	.69	9.94*
Group 3:	Schizophrenic NTD	.11	.36	.87	.34	.21	0.25
Group 4:	Schizophrenic TD	.47	.18	.03	.39	.80	0.18
Group 5:	Normal	.44	.59	.42	.21	.30	1.20
							$G^2(5)=12.07$

*$p < 0.01$.

where $\Theta^* = (\theta_1^*, \ldots, \theta_5^*)$. With the observed data in Table 1, and the structural constants $\langle \underline{c}^*_{ijk}\rangle$, $\langle a^*_{ijks}\rangle$, $\langle b^*_{ijks}\rangle$, it is straightforward to use (11) and (12) to carry out the EM algorithm.

We ran the version of the model in (24) on all five groups of Table 1 separately using a program for the source monitoring model designed by Hu (1990) for an IBM PC and obtained the parameter estimates in Table 2 along with the goodness-of-fit $G^2(1)$ values and confidence intervals (not in Table 2). The fit statistic reported is the log-likelihood ratio statistic G^2 defined later in (29). The statistic G^2 is asymptotically chi-square distributed when the model is true. These computations consumed (on an IBM PC 386-16Mhz) about 5 seconds.

Next we consider hypothesis testing between the five groups in Table 1. To illustrate, suppose we want to test the hypothesis that parameter θ_3^* is the same in all five groups. It is easy to join the five models in the form of (24) using the methods described leading to (20). Note that this is a second use of the joining principle justified in Observation 6. In fact, there are no important differences between a single GPT model and a joint GPT model for the EM algorithm. In practice, it is very easy to join GPT models together by recoding the structural constants and changing the notations for the parameters. For example, by joining 5 structurally identical GPT models like (24), each corresponding to an experimental group in Harvey's (1985) experiment, a GPT model with 25 parameters ($\Theta^{**} = (\Theta_1^{**}, \ldots, \Theta_5^{**})$, $\Theta_k^{**} = (\theta_{1k}^{**}, \ldots, \theta_{5k}^{**})$, $k = 1, \ldots, 5$), 45 observed frequencies (subject to 15 marginal constraints), and 30 degrees of freedom are obtained. Let ξ_l be the proportion of observations in group l. Then using the notation of Observation 6, the model becomes

$$\mathcal{M}(\Theta^{**}; \langle \underline{c}^{**}_{ijkl}\rangle; \langle a^{**}_{ijkls}\rangle; \langle b^{**}_{ijkls}\rangle), \tag{25}$$

where $l = 1, \ldots, 5$ and $\underline{c}^{**}_{ijkl} = \tau_k \xi_l$. Then one can test hypotheses for the model in (25) such as:

$$H_0 : \theta_{31}^{**} = \theta_{32}^{**} = \theta_{33}^{**} = \theta_{34}^{**} = \theta_{35}^{**} \tag{26}$$

by estimating parameters of the new model in (25) with the restriction in (26) denoted by

$$\mathcal{M}(\Theta^{***}; \underline{\langle \underline{c}^{***}_{ijkl}\rangle}; \langle a^{***}_{ijkls}\rangle; \langle b^{***}_{ijkls}\rangle). \tag{27}$$

TABLE 3
Parameter Estimates and Goodness-of-Fit Test for Harvey's (1985) (joint tree model) with $\theta_{31}^{***} = \theta_{32}^{***} = \theta_{33}^{***} = \theta_{34}^{***} = \theta_{35}^{***}$

Group		Parameter estimate				
		θ_{1k}^{***}	θ_{2k}^{***}	θ_{3k}^{***}	θ_{4k}^{***}	θ_{5k}^{***}
Group 1 (k=1):	Manic NTD	.40	.62	.46	.37	.17
Group 2 (k=2):	Manic TD	.53	.29	.46	.19	.69
Group 3 (k=3):	Schizophrenic NTD	.19	.40	.46	.31	.19
Group 4 (k=4):	Schizophrenic TD	.41	.03	.46	.43	.78
Group 5 (k=5):	Normal	.43	.59	.46	.22	.30
$G^2(9)$=15.20						

The hypothesis H_0 in (26) is tested by comparing the fits of the unrestricted model in (25) with 25 parameters reported in Table 2 with the restricted model in (27) with 21 parameters. Table 3 presents the results for the restricted model, and subtracting the G^2 values yields $G^2(4) = 3.13$, $p > .50$.

It is not practical to illustrate all of our results in this example; however, the preceding analyses give some insight in how to apply the methods to a practical problem. The Hu (1990) program handles all of the above results, including restricting parameters for the model for one group and joining a particular restricted model over separate experimental groups, and the reader may write to us for it.

Extensions to the Power Divergence Family

So far we have explored how to apply the EM algorithm to GPT models to yield maximum likelihood estimators. Due to the special form of processing tree models, the idea behind the EM algorithm can also be used to obtain a broader family of estimators. Cressie and Read (1984; Read & Cressie, 1988) have proposed a family of power divergence statistics for categorical data that compare the fit of observed frequencies to expected frequencies based on estimated category probabilities. In our notation, the family of power divergence statistics is defined, for every real λ, by

$$2nI^{(\lambda)}\left(\left\langle\frac{n_j}{n}\right\rangle:\tilde{p}\right) = \frac{2}{\lambda(\lambda+1)}\sum_{j=1}^{J} n_j\left[\left(\frac{n_j}{n\tilde{p}_j}\right)^{\lambda} - 1\right], \tag{28}$$

where $\tilde{p}$ is a probability distribution over the categories. The cases of $\lambda = -1$ and $\lambda = 0$ are defined by taking limits in (28).

With different λ values, different test statistics can be obtained. For example, when $\lambda = 0$,

$$\lim_{\lambda\to 0} 2nI^{(\lambda)}\left(\left\langle\frac{n_j}{n}\right\rangle:\tilde{p}\right) = 2\sum_{j=1}^{J} n_j \log\left(\frac{n_j}{n\tilde{p}_j}\right) = G^2, \tag{29}$$

which is the test statistic based on the likelihood function. For $\lambda = 1$,

$$2nI^{(1)}\left(\left\langle\frac{n_j}{n}\right\rangle:\tilde{p}\right) = 2\sum_{j=1}^{J}\frac{(n_j - n\tilde{p}_j)^2}{n\tilde{p}_j} = X^2,$$

which corresponds to the Pearson statistic based on the chi-square function. The extension of our methods to the power divergence family is useful because different researchers have used fit criteria, such as X^2 instead of G^2, for parametric multinomial models (see Batchelder, 1991).

If we replace each $\tilde{p}_j$ by $p_j(\Theta)$ in (28), then (28) is a function of $\Theta \in \Omega = (0, 1)^S$ for fixed frequency counts. In fact, (29) becomes the criterion for parameter estimation and hypothesis testing based on the likelihood function, and minimizing it over Θ is the same as maximizing the likelihood function. Read and Cressie (1988, Appendix A5) show that all the asymptotic properties of estimation based on the likelihood function hold for any member of the power divergence family. Since these inferential methods are based on minimizing the criterion function, the rest of this section is devoted to extending the EM algorithm to minimize (28) as a function of Θ in the context of a GPT model.

Observation 7. Assume that the $\langle c_{ij}\rangle$, $\langle a_{ijs}\rangle$, $\langle b_{ijs}\rangle$ are chosen so that (3) defines a GPT model over J categories with parameter $\Theta = (\theta_1, \ldots, \theta_S)$. Also assume that $2nI^{(\lambda)}(\langle n_j/n\rangle:p(\Theta))$ has minima in the interior of $[0, 1]^S$. Then any $\tilde{\Theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_S)$ that minimizes $2nI^{(\lambda)}(\langle n_j/n\rangle:p(\Theta))$ will satisfy

$$\tilde{\theta}_s = \frac{\displaystyle\sum_{j=1}^{J} n_j\left(\frac{n_j}{np_j(\tilde{\Theta})}\right)^{\lambda}\sum_{i=1}^{I_j}\frac{p_{ij}(\tilde{\Theta})}{p_j(\tilde{\Theta})}a_{ijs}}{\displaystyle\sum_{j=1}^{J} n_j\left(\frac{n_j}{np_j(\tilde{\Theta})}\right)^{\lambda}\sum_{i=1}^{I_j}\frac{p_{ij}(\tilde{\Theta})}{p_j(\tilde{\Theta})}(a_{ijs} + b_{ijs})}, \quad s = 1, \ldots, S. \tag{30}$$

Having (11) as the E part, an EM$_\lambda$ algorithm can be constructed. Here, $\tilde{\Theta}$ is not the MLE of the model and EM indexed by λ means that the algorithm depends on λ. So (31), similar to (12), can be obtained in the same manner:

$$\phi_s^{(\lambda)} = \frac{\displaystyle\sum_{j=1}^{J}\left(\frac{n_j}{np_j(\Theta)}\right)^{\lambda}\sum_{i=1}^{I_j} m_{ij}(\Theta)a_{ijs}}{\displaystyle\sum_{j=1}^{J}\left(\frac{n_j}{np_j(\Theta)}\right)^{\lambda}\sum_{i=1}^{I_j} m_{ij}(\Theta)(a_{ijs} + b_{ijs})}, \tag{31}$$

$s = 1, \ldots, S$. Note that when $\lambda \neq 0$, (31) is not identical to (12), so theoretical results from the previous sections about the EM algorithm can not be directly used here. In fact, with large λ, EM$_\lambda$ does not always converge. However, an alternative algorithm can be obtained. The following observation gives the algorithm and the convergence result.

Observation 8. For any real number ε, such that $\varepsilon(1 + \lambda) > 0$, the limit of a convergent sequence $\{\Theta^{(n)}\}$ such that:

$$\Theta^{(n+1)} = \Theta^{(n)} - \varepsilon(\Theta^{(n)} - M_\lambda(\Theta^{(n)})), \tag{32}$$

is a minima of $2nI^{(\lambda)}(\langle n_j/n\rangle : p(\Theta))$, where $M_\lambda(\Theta)$ is the vector $(\phi_1^{(\lambda)}, \ldots, \phi_S^{(\lambda)})$.

We call the above algorithm $\mathrm{EM}_\lambda^{(\varepsilon)}$ algorithm, since it uses the original idea of EM applied to the power divergence family, with an improved iteration formula in (32) involving ε. It is easy to notice that

$$\mathrm{EM} = (\mathrm{EM}_\lambda)_{\lambda=0} = (\mathrm{EM}_\lambda^{(\varepsilon)})_{\lambda=0, \varepsilon=1}.$$

Observation 9. If $\bar{\Theta}$ is a minima of $2nI^{(\lambda)}(\langle n_j/n\rangle : p(\Theta))$, then there is a real number $\varepsilon \neq 0$ and a neighborhood of $\bar{\Theta}$ such that the $\mathrm{EM}_\lambda^{(\varepsilon)}$ sequence $\{\Theta^{(n)}\}$, with any initial value in the neighborhood, converges to $\bar{\Theta}$.

The $\mathrm{EM}_\lambda^{(\varepsilon)}$ algorithm can also be used to improve the speed of convergence for the regular EM algorithm developed earlier. In the literature, instances occur where EM sequences reach the neighborhood quite quickly but move slowly in the neighborhood itself (see Ruud, 1991). By using the above technique, the speed of convergence can be changed dramatically. From (32), we can see that ε controls the step size of the iteration, so that when the EM sequence is moving very slowly, increasing ε will speed up the search process. As a matter of fact, the only possible case where the EM sequence is not convergent is one in which the likelihood function has two or more local maxima with exactly the same value and the EM sequence jumps from one to the other. Even in this unlikely case, adjusting (in fact, decreasing) $|\varepsilon|$ can prevent the EM sequence from jumping and force it to converge to one of the maxima. It is easy to implement this idea in designing a computer program. We have done this in our cqmputer package by Hu (1991). (See the note after the proof of Observation 9.) Therefore, by adjusting the values of ε, we can obtain a convergent EM sequence regardless of the starting point.

Conclusion

Most of the methods we have described have been programmed for a PC, and they have been used to analyze many data sets by ourselves and others, for example, Batchelder, Hu, and Riefer (in press), Batchelder and Riefer (1990), Erdfelder and Bayen (1991). A program based on the EM algorithm for source monitoring (Hu, 1990) is in wide circulation, and a program for handling an arbitrary GPT model is available (for an early version, see Hu, 1991). Both programs generate point and region estimators and conduct goodness-of-fit tests for both within-group and between-group hypotheses. In their current form, they can handle up to 50 parameters and still get estimates in under a minute on a 80386 CPU based PC.

There are several salient advantages of the EM algorithm over other iterative methods based on gradient search for analyzing GPT models. First, no special symbolic or numerical differentiation is required for a specific model. Second, one does not have to worry about setting the step size (unless wanting to speed the algorithm—see Observation 8). Third, in our applications the algorithm has proved as fast or faster than other conventional iterative methods when applied to specific GPT models, such as the source monitoring models of Batchelder and Riefer (1990). Fourth, the entire class of GPT models can be analyzed by specifying structural constants (see Definition 1) in a single program (Hu, 1991). Finally, the form of (3) is quite general, and given suitable parameterizations, captures many models for categorical data that can not be formulated conveniently as log-linear models.

Appendix A

Proof of Observation 1. The result will follow recursively if we can show that the submodel obtained by setting a single parameter to a constant is a GPT model. Without loss of generality, set θ_1 to a constant, $0 \leq k_1 \leq 1$, and let $\Theta' = (\theta_2, \ldots, \theta_S)$. Then (2) can be rewritten

$$p_{ij}(\Theta) = p_{ij}(\Theta') = c^*_{ij} \prod_{s=2}^{S} \theta_s^{a_{ijs}}(1 - \theta_s)^{b_{ijs}}, \tag{33}$$

where $c^*_{ij} = c_{ij}k_1^{a_{ij1}}(1 - k_1)^{b_{ij1}}$. Clearly the system of equations in (33) satisfies (2), (3), and (4) for all $\Theta' \in [0, 1]^{S-1}$. If $0 < k_1 < 1$, it is obvious that the submodel is a GPT model on the old categories and branches with new parameter Θ'. If $k_1 = 0$ or $k_1 = 1$, then some of the branches and perhaps some of the other parameters or categories are eliminated. For example, if $k_1 = 0$, all branches with $a_{ij1} > 0$ have zero probability and all parameters that enter the $p_{ij}(\Theta')$ only on those branches disappear, that is, parameters where $a_{ijs} + b_{ijs} > 0$ implies $a_{ij1} > 0$. Nevertheless, once the unnecessary branches, categories, and parameters are eliminated, it is obvious that the remaining structure satisfies Definition 1. □

Proof of Observation 2. We apply mathematical induction on the number of categories J. If $J = 1$, the model has $S = 0$ parameters, so $S < J$. Let $J = 2$ and assume that model has parameters $\Theta = (\theta_1, \ldots, \theta_S)$, for $S \geq 1$. From (2) and (4), we have

$$c_1 \prod_{s=1}^{S} \theta_s^{a_{1s}}(1 - \theta_s)^{b_{1s}} + c_2 \prod_{s=1}^{S} \theta_s^{a_{2s}}(1 - \theta_s)^{b_{2s}} = 1, \tag{34}$$

for all $\Theta \in [0, 1]^S$. Without loss of generality, assume $a_{11} > 0$, and set $\theta_s = 1/2$, for $s = 2, \ldots, S$. Then (34) becomes

$$c_1^* \theta_1^{a_{11}}(1 - \theta_1)^{b_{11}} + c_2^* \theta_1^{a_{21}}(1 - \theta_1)^{b_{21}} = 1, \tag{35}$$

for all $0 \leq \theta_1 \leq 1$, where c_1^* and c_2^* are positive constants. It is easy to show that (35) and $a_{11} > 0$ require $c_1^* = c_2^* = a_{11} = b_{21} = 1$ and $b_{11} = a_{21} = 0$. But this is possible only if $S = 1$, since otherwise either $c_1^* = 1/2$ or $c_2^* = 1/2$. Thus, if $J = 2$, we have $S < J$ as required.

Now assume that any GPT model with $J \leq K$ categories has $S < J$ parameters, and set $J = K + 1$. For any $r = 1, \ldots, S$, denote

$$A_r = \{j | a_{jr} > 0\},$$

and

$$B_r = \{j | b_{jr} > 0\}.$$

We will use the fact that $A_r \neq B_r$, for $r = 1, \ldots, S$. To see this, suppose there is an r with $A_r = B_r$, and set all other parameters equal to 1/2 and retain θ_r. The resulting submodel is a GPT model by Observation 1, so by (2) and (4) it would have to satisfy

$$\sum_{j \in A_r} c_j^* \theta_r^{a_{jr}}(1 - \theta_r)^{b_{jr}} = 1,$$

for all $0 \leq \theta_r \leq 1$ and positive c_j^*, a_{jr}, b_{jr}. However, if $\theta_r = 0$ or $\theta_r = 1$, this is not possible; thus, $A_r \neq B_r$. The family of sets $\{A_r | r = 1, \ldots, S\}$ will contain at least one "minimal" non empty set A_{r_0} such that no other A_r properly contains A_{r_0}. Consider a submodel that is obtained by setting $\theta_{r_0} = 0$ and retaining the other $S - 1$ parameters. From Observation 1, it is a GPT model, and it is obvious that it has K or less categories. Furthermore, all other parameters are still in the model, because if any θ_{r_1} is missing,

$$A_{r_1} \cup B_{r_1} \subset A_{r_0},$$

and since $A_{r_1} \neq B_{r_1}$, A_{r_1} is properly contained in A_{r_0}, which contradicts that A_{r_0} is a minimal set. Thus, the new GPT model with $\theta_{r_0} = 0$ has at most K categories and it has $S - 1$ parameters. From the induction hypothesis, we can conclude $S - 1 < K$. Thus $S < K + 1$, and the observation follows by induction. □

Proof of Observation 3. First, compute $\ln L(\Theta; D)$ and then substitute for $p_{ij}(\Theta)$ from (2). It is straightforward to see that

$$\frac{\partial \ln L(\Theta; D)}{\partial \theta_s} = \sum_{j=1}^{J} \sum_{i=1}^{I_j} \left[\frac{a_{ijs} m_{ij}}{\theta_s} - \frac{b_{ijs} m_{ij}}{1 - \theta_s} \right]. \tag{36}$$

Setting each of the above equations to zero yields (10). It is straightforward to show that (10) maximizes $L(\Theta; D)$, since the matrix

$$\frac{\partial^2 \ln L(\Theta; D)}{\partial \theta_s \partial \theta_r} = \begin{cases} -\sum_{j=1}^{J} \sum_{i=1}^{I_j} \left[\frac{a_{ijs} m_{ij}}{\theta_s^2} + \frac{b_{ijs} m_{ij}}{(1 - \theta_s)^2} \right] & \text{for} \quad s = r \\ 0 & \text{for} \quad s \neq r, \end{cases} \tag{37}$$

is negative definite for all Θ in the domain. Therefore $L(\Theta; D)$ has only one global maxima given by (10). □

Proof of Observation 4. By using $\alpha_s(\Theta)$ and $\beta_s(\Theta)$, (12) becomes

$$\phi_s = \frac{\alpha_s(\Theta)}{\alpha_s(\Theta) + \beta_s(\Theta)},$$

so

$$\frac{\partial \phi_s}{\partial \theta_r} = \frac{\frac{\partial \alpha_s(\Theta)}{\partial \theta_r} \beta_s(\Theta) - \frac{\partial \beta_s(\Theta)}{\partial \theta_r} \alpha_s(\Theta)}{[\alpha_s(\Theta) + \beta_s(\Theta)]^2}.$$

Next consider any point $\hat{\Theta}$ in $(0, 1)^S$, such that $\hat{\Theta} = M(\hat{\Theta})$, i.e.,

$$\hat{\theta}_s = \frac{\alpha_s(\hat{\Theta})}{\alpha_s(\hat{\Theta}) + \beta_s(\hat{\Theta})}. \tag{38}$$

Then

$$\frac{\partial \phi_s}{\partial \theta_r} = \left[\frac{\frac{\partial \alpha_s(\hat{\Theta})}{\partial \theta_r}}{\hat{\theta}_s} - \frac{\frac{\partial \beta_s(\hat{\Theta})}{\partial \theta_r}}{1 - \hat{\theta}_s} \right] \left[\frac{\hat{\theta}_s(1 - \hat{\theta}_s)}{\alpha_s(\hat{\Theta}) + \beta_s(\hat{\Theta})} \right]. \tag{39}$$

Denote

$$I(\hat{\Theta}) = -\left(\frac{\partial^2 \ln L(\Theta;\, \langle n_j \rangle_{j=1}^{J})}{\partial \theta_s \partial \theta_r}\right)_{S \times S}\Bigg|_{\Theta = \hat{\Theta}},$$

$$V(\hat{\Theta}) = \left(\left[\frac{\partial \phi_s}{\partial \theta_r}\right]_{s,r}\right)_{S \times S}\Bigg|_{\Theta = \hat{\Theta}},$$

and a diagonal matrix

$$W(\hat{\Theta}) = \left(\left[\frac{\hat{\theta}_s(1 - \hat{\theta}_s)}{\alpha_s(\hat{\Theta}) + \beta_s(\hat{\Theta})}\right]_{s,s}\right)_{S \times S}.$$

Then by (16), (39), and using (38) again, and from

$$-\left[\frac{\alpha_s(\hat{\Theta})}{\hat{\theta}_s^2} + \frac{\beta_s(\hat{\Theta})}{(1 - \hat{\theta}_s)^2}\right] = -\left[\frac{\alpha_s(\hat{\Theta}) + \beta_s(\hat{\Theta})}{\hat{\theta}_s(1 - \hat{\theta}_s)}\right],$$

we have

$$V(\hat{\Theta}) = -I(\hat{\Theta})W(\hat{\Theta}) + E, \tag{40}$$

where E is the unit matrix, $E_{S \times S}$.

Consider any EM sequence $\langle \Theta^{(n)} \rangle$, where $\Theta^{(n+1)} = M(\Theta^{(n)})$. Then, we have

$$\Theta^{(n+1)} - \Theta^{(n)} = M(\Theta^{(n)}) - M(\Theta^{(n-1)}). \tag{41}$$

From (11) and (12), it is easy to see that $M(\Theta)$ is differentiable with respect to Θ, when $\Theta \in (0,\, 1)^S$, and by the mean value theorem in calculus, there is a Θ_{ε_n} such that

$$M(\Theta^{(n)}) - M(\Theta^{(n-1)}) = \left[\left(\left[\frac{\partial \phi_s(\Theta)}{\partial \theta_r}\right]_{s,r}\right)_{S \times S}\right]_{\Theta = \Theta_{\varepsilon_n}} [\Theta^{(n)} - \Theta^{(n-1)}]; \tag{42}$$

and if

$$\lim_{n \to \infty} \left[\left(\left[\frac{\partial \phi_s(\Theta)}{\partial \theta_r}\right]_{s,r}\right)_{S \times S}\right]_{\Theta = \Theta_{\varepsilon_n}} = V(\hat{\Theta})$$

exists, from (41) and (42), when n is very large, we have

$$\Theta^{(n+1)} - \Theta^{(n)} \approx V(\hat{\Theta})[\Theta^{(n)} - \Theta^{(n-1)}].$$

Thus $\langle \Theta^{(n)} \rangle$ is a convergent sequence if and only if the eigenvalues of V are bounded by 1, and it is obvious that when $\langle \Theta^{(n)} \rangle$ converges to any point $\Theta^* \in (0,\, 1)^S$, $V = V(\Theta^*)$.

When $\Theta^* = \lim_{n \to \infty} \Theta^{(n)}$ exists, then Θ^* is a stationary point of (7). On the other hand, the eigenvalues of $V(\Theta^*)$ are bounded by 1, so all eigenvalues of $-I(\Theta^*)W(\Theta^*)$

are negative. The eigenvalues of $W(\Theta^*)$ are already positive, so that the eigenvalues of $-I(\Theta^*)$, the matrix of second derivatives of the log-likelihood function, must be negative. Therefore Θ^* (at least locally) maximizes (7). □

Note: As Wu (1983) points out, the reason that a convergent EM sequence is not necessarily convergent to a maxima is because the log-likelihood function is written in such a form that $I(\hat{\Theta})$ is the difference of two negative definite matrices (Wu, 1983, p. 97). This prevents one from establishing the definiteness status of $I(\hat{\Theta})$ from the expression. In our case, $I(\hat{\Theta})$ can be expressed by (40), so it is easy to see that it is negative definite.

Proof of Observation 5. It is clear that the restricted model defined above is a multinomial model over the same J categories with $S - z$ functionally independent parameters, each defined in [0, 1]. Further, each $p_j(\Theta'')$ can be written as

$$p_j(\Theta'') = \sum_{i=1}^{I_j} c_{ij} X_{ij} \prod_{k=z+1}^{S} \theta_k^{a_{ijk}}(1 - \theta_k)^{b_{ijk}}, \tag{43}$$

where

$$X_{ij} = \prod_{s=1}^{z} f_s(\Theta'')^{a_{ijs}}(1 - f_s(\Theta''))^{b_{ijs}}.$$

It is obvious from Items i, ii, and iii that $f_k(\Theta'')^{a_{ijs}}(1 - f_k(\Theta''))^{b_{ijs}}$ has one of the following forms, for each $s = 1, \ldots, z$:

$$\alpha^{a_{ijs}}(1 - \alpha)^{b_{ijs}}, \tag{44}$$

$$\left[\prod_{k=z+1}^{S} \theta_k^{\mu_k}\right]^{a_{ijs}} \left[1 - \prod_{k=z+1}^{S} \theta_k^{\mu_k}\right]^{b_{ijs}}, \tag{45}$$

or

$$\left[\sum_{k=z+1}^{S} \nu_k \theta_k^{\rho_k}\right]^{a_{ijs}} \left[1 - \sum_{k=z+1}^{S} \nu_k \theta_k^{\rho_k}\right]^{b_{ijs}}. \tag{46}$$

Next, we show that (43) can be rewritten in GPT model form, namely,

$$p_j(\Theta'') = \sum_{i=1}^{I_j^*} c_{ij}^* \prod_{k=z+1}^{S} \theta_k^{a_{ijs}^*}(1 - \theta_k)^{b_{ijs}^*}, \tag{47}$$

for new constants $\langle c_{ij}^* \rangle$, $\langle a_{ijs}^* \rangle$ and $\langle b_{ijs}^* \rangle$.

Equation (44) provides no problem since it is a constant independent of Θ''. We show that the other two cases can be handled by adding unobserved categories but preserving the observed categories and the GPT model form. First consider (45). Without loss of generality, let $\mu_{z+1} = 1$. Then

$$\left[1 - \prod_{k=z+1}^{S} \theta_k^{\mu_k}\right] = (1 - \theta_{z+1}) + \theta_{z+1}\left[1 - \prod_{k=z+2}^{S} \theta_k^{\mu_k}\right].$$

Using mathematical induction, it is not hard to show that

$$\left[1-\prod_{k=z+1}^{S}\theta_k^{\mu_k}\right]=\sum_{k=1}^{S'}(1-\theta_{K_k})\prod_{l=1}^{k-1}\theta_{K_l},$$

where $z+1\le K_k\le S$ and $\mu_{K_k}\neq 0$ and $S'_s=\sum_{k=z+1}^{S}\mu_k$. So expand (45) to the form

$$\sum_u d_u\prod_{k=z+1}^{S}\theta_k^{\gamma_{uk}}(1-\theta_k)^{\eta_{uk}}. \tag{48}$$

Finally, consider (46). It is easy to see that

$$\left[1-\sum_{k=z+1}^{S}\nu_k\theta_k^{\rho_k}\right]=\sum_{k=z+1}^{S}\nu_k[1-\theta_k^{\rho_k}]=\sum_{k=z+1}^{S}\nu_k\left[(1-\theta_k)\sum_{l=1}^{\rho_k-1}\theta_k^{l}\right].$$

Therefore (46) yields the same form as (48). Since X_{ij} is just a product of forms like (48), it has the same form as (48). Hence (43) can be written as (47), and the restricted model is a GPT model. □

Proof of Observation 6. Denote (22) by

$$U(\Theta;\,T;\,\Delta)=L(\Theta;\,\Delta)V(T;\,\Delta),$$

where

$$L(\Theta;\,\Delta)=\prod_{k=1}^{K}L_k(\Theta;\,\langle n_{jk}\rangle_{j=1}^{J_k})$$

and

$$V(T;\,\Delta)=\binom{N}{N_1\cdots N_K}\left[\prod_{k=1}^{K-1}\tau_k^{N_k}\right]\left(1-\sum_{k=1}^{K-1}\tau_k\right)^{N-\sum_{k=1}^{K-1}N_k},$$

with $T=(\tau_1,\ldots,\tau_{K-1})$, $\Delta=(\langle n_{j1}\rangle_{j=1}^{J_1},\ldots,\langle n_{jK}\rangle_{j=1}^{J_K})$.

The Fisher information matrix for $(\hat{\Theta};\,\hat{T})$, from (22) is

$$\begin{pmatrix}A & 0\\ 0 & B\end{pmatrix}, \tag{49}$$

where

$$A=A_{S\times S}=\left[\mathbf{E}\left(\frac{\partial^2\log(L(\Theta;\,\Delta))}{\partial\theta_i\partial\theta_j}\right)\right]_{\Theta=\hat{\Theta}},$$

$$B=B_{(K-1)\times(K-1)}=\left[\mathbf{E}\left(\frac{\partial^2\log(V(T;\,\Delta))}{\partial\tau_u\partial\tau_v}\right)\right]_{T=\hat{T}}.$$

Since A is the Fisher information matrix that can be obtained from (21), and it is also the Fisher information matrix for $\hat{\Theta}$ in (49), we conclude the proof. □

Proof of Observation 7. Differentiate $2nI^{(\lambda)}(\langle n_j/n\rangle:p(\Theta))$ with respect to θ_s, $s = 1, \ldots, S$, assign each of them zero, and then rearrange the derivatives. Then (30) can be obtained by the assumption that $\tilde{\Theta}$ minimizes $2nI^{(\lambda)}(\langle n_j/n\rangle:p(\Theta))$ in the interior of $[0, 1]^S$. □

Proof of Observation 8. In order to prove Observation 8, we need to introduce some notations and use two lemmas. Denote

$$\alpha_s^{(\lambda)}(\Theta) = \sum_{j=1}^{J} n_j\left(\frac{n_j}{np_j(\Theta)}\right)^{\lambda} \sum_{i=1}^{I_j} \frac{p_{ij}(\Theta)}{p_j(\Theta)}\, a_{ijs},$$

$$\beta_s^{(\lambda)}(\Theta) = \sum_{j=1}^{J} n_j\left(\frac{n_j}{np_j(\Theta)}\right)^{\lambda} \sum_{i=1}^{I_j} \frac{p_{ij}(\Theta)}{p_j(\Theta)}\, b_{ijs},$$

$$I_\lambda(\tilde{\Theta}) = \left(\left[\frac{\partial^2 nI^{(\lambda)}\left(\left\langle\frac{n_j}{n}\right\rangle:p(\Theta)\right)}{\partial\theta_s\partial\theta_r}\right]_{s,r}\right)_{S\times S}\Bigg|_{\Theta=\tilde{\Theta}},$$

$$V_\lambda(\tilde{\Theta}) = \left(\left[\frac{\partial\phi_s^{\lambda}}{\partial\theta_r}\right]_{s,r}\right)_{S\times S}\Bigg|_{\Theta=\tilde{\Theta}},$$

$$W_\lambda(\tilde{\Theta}) = \left(\left[\frac{\tilde{\theta}_s(1-\tilde{\theta}_s)}{\alpha_s^{\lambda}(\tilde{\Theta}) + \beta_s^{\lambda}(\tilde{\Theta})}\right]_{s,s}\right)_{S\times S}.$$

Lemma 1. $I_\lambda(\Theta)$, $W_\lambda(\Theta)$, $V_\lambda(\Theta)$ are continuous with respect to $\Theta \in (0, 1)^S$ and $\lambda \in (-\infty, +\infty)$.

Lemma 2. If $\tilde{\Theta}$ satisfies (30) then

$$V_\lambda(\tilde{\Theta}) = E - (1+\lambda)I_\lambda(\tilde{\Theta})W_\lambda(\tilde{\Theta}), \tag{50}$$

where E is the unit matrix $E = E_{S\times S}$.

Proof of Lemma 1 is straightforward. Proof of (50) in Lemma 2 is almost the same as that for (40), except one has to consider $\lambda \neq 0$.

From (32),

$$(\Theta^{(n+1)} - \Theta^{(n)}) = (\Theta^{(n)} - \Theta^{(n-1)}) - \varepsilon((\Theta^{(n)} - \Theta^{(n-1)}) - (M_\lambda(\Theta^{(n)}) - M_\lambda(\Theta^{(n-1)})).$$

Using the same argument for (42) and denoting

$$V_\lambda^{(\varepsilon)}(\Theta) = (1-\varepsilon)E + \varepsilon V_\lambda(\Theta),$$

we have

$$(\Theta^{(n+1)} - \Theta^{(n)}) = (V_{\lambda}^{(\varepsilon)}(\Theta))_{\Theta = \Theta_{\varepsilon_n}} [\Theta^{(n)} - \Theta^{(n-1)}], \tag{51}$$

for some $\Theta_{\varepsilon_n} \in (0, 1)^S$. From Lemma 1, and the convergence of $\{\Theta^{(n)}\}$,

$$V_{\lambda}(\tilde{\Theta}) = \lim_{n \to \infty} \left[\left(\left[\frac{\partial \phi_s^{\lambda}(\Theta)}{\partial \theta_r} \right]_{s,r} \right)_{S \times S} \right]_{\Theta = \Theta_{\varepsilon_n}}$$

By (50) and the continuity of $M_{\lambda}(\Theta)$, and Lemma 1, matrix

$$V_{\lambda}^{(\varepsilon)}(\tilde{\Theta}) = E - \varepsilon(1 + \lambda) I_{\lambda}(\tilde{\Theta}) W_{\lambda}(\tilde{\Theta}),$$

has eigenvalues in $(-1, 1)$. Since $\varepsilon(1 + \lambda) > 0$, $I_{\lambda}(\tilde{\Theta})$ is positive definite, and $\tilde{\Theta}$ is a minima of $2nI^{(\lambda)}(\langle n_j/n \rangle{:}p(\Theta))$. □

Proof of Observation 9. The key to this proof is to show that there is an ε and a neighborhood of $\tilde{\Theta}$, such that whenever Θ is in that neighborhood, eigenvalues of $V_{\lambda}^{(\varepsilon)}(\Theta)$ are in $(-1, +1)$.

By assumption, $I_{\lambda}(\tilde{\Theta})$ is positive definite. By the continuity of $I_{\lambda}(\Theta)$, $W_{\lambda}(\Theta)$ and (50) of Lemma 2, it is easy to show that there is a δ_1 neighborhood $\mathcal{N}(\tilde{\Theta}; \delta_1)$ of $\tilde{\Theta}$, such that

1. $2nI^{(\lambda)}(\langle n_j/n \rangle{:}p(\Theta))$ has only one minimum in $\mathcal{N}(\tilde{\Theta}; \delta_1)$,
2. for any $\Theta \in \mathcal{N}(\tilde{\Theta}; \delta_1)$, $I_{\lambda}(\Theta) W_{\lambda}(\Theta)$ is positive definite. So, there is $\varepsilon \neq 0$, such that $\varepsilon(1 + \lambda) > 0$ and

   $$E - \varepsilon(1 + \lambda) I_{\lambda}(\Theta) W_{\lambda}(\Theta)$$

   has eigenvalues in $(-1, +1)$, for all $\Theta \in \mathcal{N}(\tilde{\Theta}; \delta_1)$.
3. $V_{\lambda}(\Theta) = E - (1 + \lambda) I_{\lambda}(\Theta) W_{\lambda}(\Theta) + \Gamma(\Theta)$, where $\lim_{\Theta \to \tilde{\Theta}} \Gamma(\Theta) = 0$.

In other words, when $\Theta \in \mathcal{N}(\tilde{\Theta}; \delta_1)$,

$$\begin{aligned} V_{\lambda}^{(\varepsilon)}(\Theta) &= (1 - \varepsilon)E + \varepsilon V_{\lambda}(\Theta) \qquad (52) \\ &= E - \varepsilon(1 + \lambda) I_{\lambda}(\Theta) W_{\lambda}(\Theta) + \varepsilon \Gamma(\Theta). \end{aligned}$$

Since

$$\lim_{\Theta \to \tilde{\Theta}} V_{\lambda}^{(\varepsilon)}(\Theta) = E - \varepsilon(1 + \lambda) I_{\lambda}(\tilde{\Theta}) W_{\lambda}(\tilde{\Theta}),$$

and using Lemma 1, there is a δ_2 neighborhood $\mathcal{N}(\tilde{\Theta}; \delta_2)$, such that the eigenvalues of $V_{\lambda}^{(\varepsilon)}(\Theta)$ are in $(-1, +1)$, for any $\Theta \in \mathcal{N}(\tilde{\Theta}; \delta_2)$. Let $\delta = \min\{\delta_1, \delta_2\}$, then whenever $\Theta \in \mathcal{N}(\tilde{\Theta}; \delta)$, eigenvalues of $V_{\lambda}^{(\varepsilon)}(\Theta)$ are in $(-1, +1)$. So, the $\mathrm{EM}_{\lambda}^{(\varepsilon)}$ sequence $\{\Theta^{(n)}\}$, with initial value in $\mathcal{N}(\tilde{\Theta}; \delta)$, is convergent.

Since there is only one minimum in $\mathcal{N}(\tilde{\Theta}; \delta)$, by Observation 8, the $\mathrm{EM}_{\lambda}^{(\varepsilon)}$ sequence has $\tilde{\Theta}$ as its limit. □

Note: From (52), ε is controlling the eigenvalues of $V_{\lambda}^{(\varepsilon)}(\Theta)$. Decreasing $|\varepsilon|$ and keeping $\varepsilon(1 + \lambda) > 0$, one can assure that eigenvalues of $V_{\lambda}^{(\varepsilon)}(\Theta)$ remain in $(-1, 1)$, which guarantees the convergence. We have designed a PC program for the $\mathrm{EM}_{\lambda}^{(\varepsilon)}$ algorithm using the above results. For many data sets and models that we have tried, when λ is very small (for example, $|\lambda| < 0.5$), convergence has always been obtained regardless of the starting value. When λ is large, we have been able to adjust the values of ε, so that the $\mathrm{EM}_\lambda^{(\varepsilon)}$ sequence converges.

References

Batchelder, W. H. (1991). Getting wise about minimum distance measures [Review of *Goodness-of-fit statistics for discrete multivariate data* by T. R. C. Read & N. A. C. Cressie]. *Journal of Mathematical Psychology, 35*, 267–273.

Batchelder, W. H., Hu, X., & Riefer, D. M. (in press) Analysis of a model for source monitoring. In G. H. Fischer & D. Laming (Eds.), *Mathematical psychology: New developments*. Berlin: Springer-Verlag. (Available as Technical Report No. 92-07, Institute for Mathematical Behavior Sciences, School of Social Sciences, UC, Irvine)

Batchelder, W. H., & Riefer, D. M. (1986). Statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology, 39*, 129–149.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548–564.

Bäuml, K.-H. (1991). Experimental analysis of storage and retrieval processes involved in retroactive inhibition: The effect of presentation mode. *Acta Psychologica, 77*(2), 103–119.

Bernstein, F. (1925). Zusammenfassende Betrachtungen über die erblichen Blutenstructuren des Menschen [Summarizing considerations on the inheritable blood structures of mankind]. *Z. Abstamm. Vererbgsl., 37*, 237–270.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of Royal Statistical Society, Series B, 45*, 47–50.

Ceppellini, R., Siniscalco, M., & Smith, C. A. B. (1955). The estimation of gene frequencies in random mating populations. *Annals of Human Genetics, 20*, 97–115.

Chechile, R., & Meyer, D. L. (1976). A Bayesian procedure for separately estimating storage and retrieval components of forgetting. *Journal of Mathematical Psychology, 13*, 269–295.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B, 39*, 1–38.

Efron, B., & Hinkley, D. V. (1978). The observed versus expected information. *Biometrika, 65*, 457–487.

Elandt-Johnson, R. C. (1971). *Probability models and statistical methods in genetics*. New York, Wiley & Sons.

Erdfelder, E., & Bayen, U. J. (1991). Episodisches Gedächtnis im Alter: Methodologische und empirische Arguments für einen Zugang über mathematische Modelle [Episodic memory in old age: Methodological and empirical arguments for an access through mathematical models]. In D. Frey (Eds.), *Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990, Band 2* (pp. 172–180). Göttingen: Hogrefe.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics, 14*, 174–194.

Harvey, P. D. (1985). Reality monitoring in mania and schizophrenia. *The Journal of Nervous and Mental Disease, 173*, 67–72.

Hu, X. (1990). *Source monitoring program* (Version 1.0). Irvine: University of California. (available upon request)

Hu, X. (1991). *General program for processing tree models* (Version 1.0). Irvine: University of California. (available upon request)

Humphreys, M. S., & Bain, J. D. (1983). Recognition memory: A cue and information analysis. *Memory and Cognition, 11*, 583–600.

Johnson, M. K., & Raye, C. L. (1980). Reality monitoring. *Psychological Review. 88*, 67–85.

Landsteiner, K. (1901). Über Agglutinationserscheinungen normalen menschlichen Blutes. [On agglutination appearances of normal human blood] *Wien. Klin. Wschr, 14*, 1132–1134.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton Mifflin.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society, Series B, 44*, 226–233.

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899–909.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318–339.

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In

J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313–335). New York: Springer-Verlag.

Riefer, D. M., & Rouder, J. M. (1992). A multinomial modeling analysis of the mnemonic benefits of bizarre imagery. *Memory and Cognition, 20,* 601–611.

Rosenbloom, P. S., Laird, J. E., Newell, A., & McCarl, R. (1991). A preliminary analysis of the SOAR architecture as a basis for general intelligence. *Artificial Intelligence, 47,* 289–325.

Ross, B. H., & Bower, G. H. (1981). Comparisons of models of associative recall. *Memory & Cognition, 9,* 1–16.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56,* 241–254.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing* (Vol. 1). Cambridge: MIT Press.

Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics, 49,* 305–341.

Smith, C. A. B. (1957). Counting methods in genetical statistics. *Annals of Human Genetics, 21,* 97–115.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology.* San Francisco: Freeman.

Weir, B. S. (1990). *Genetic data analysis.* Sunderland, MA: Sinaver Associates.

Wu, J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics, 11,* 95–103.

Manuscript received 3/3/92
Final version received 12/1/92