Journal of Classification 12:21-55 (1995)

A Mixture Likelihood Approach for Generalized Linear Models

Michel Wedel
University of Groningen

Wayne S. DeSarbo
University of Michigan

Abstract: A mixture model approach is developed that simultaneously estimates the posterior membership probabilities of observations to a number of unobservable groups or latent classes, and the parameters of a generalized linear model which relates the observations, distributed according to some member of the exponential family, to a set of specified covariates within each Class. We demonstrate how this approach handles many of the existing latent class regression procedures as special cases, as well as a host of other parametric specifications in the exponential family heretofore not mentioned in the latent class literature. As such we generalize the McCullagh and Nelder approach to a latent class framework. The parameters are estimated using maximum likelihood, and an EM algorithm for estimation is provided. A Monte Carlo study of the performance of the algorithm for several distributions is provided, and the model is illustrated in two empirical applications.

Keywords: Mixture models; Generalized linear models; EM algorithm; Maximum likelihood estimation.

1. Introduction

Finite mixture models have been extensively developed in the statistics and classification literature. The development of these models dates back to the work of Newcomb (1886) and Pearson (1894). In such finite mixture models, it is assumed that a sample of observations arises from a specified

---

The authors thank Lars Muus for helpful suggestions on the Monte Carlo Study.
Author's Address for Correspondence: Michel Wedel, Department of Business Administration and Management Science, Faculty of Economics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands. (Bitnet: M.WEDEL@ECO.RUG.NL)

number of the underlying populations of unknown proportions. A specific form of the density of the observations in each of the underlying populations is specified, and the purpose of the finite mixture approach is to decompose the sample into its mixture components. Specific forms of densities which have been extensively used include the normal (e.g., Hasselblad 1966; Day 1969; Wolfe 1970), exponential (Thomas 1966; Teicher 1961) and bernouili densities (the latter models are typically known as latent structure models, e.g., Lazarsfeld and Henry 1968; Goodman 1974).

Initially, the parameters of finite mixtures have been estimated using the method of moments (cf. Pearson 1894; Charlier and Wicksell 1924; Quandt and Ramsey 1978), but attention later focused on graphical techniques for the detection of (univariate) mixtures (e.g., Harding 1948; Cassie 1954; Bhattacharya 1967; Fowlkes 1979). Hasselblad (1966, 1969) was among the first to use maximum likelihood estimation for mixtures of two or more distributions from the exponential family. As maximum likelihood has been shown to be superior to the method of moments for the estimation of finite mixtures (cf. Fryer and Robertson 1972), the likelihood approach for finite normal mixtures has recently become increasingly popular (e.g., Wolfe 1970; Day 1969; Symons 1981; McLachlan 1982; Basford and McLachlan 1985).

The likelihood of finite mixtures can be maximized basically in two ways: by using standard optimization routines such as the Newton-Raphson method (McHugh 1956, 1958), or by using the Expectation-Maximization (EM) algorithm (Dempster, Laird and Lubin 1977). The Newton-Raphson method requires relatively few iterations to converge, and provides the asymptotic variances of the parameter estimates as a by-product, but convergence is not ensured (Atkinson 1989; McLachlan and Basford 1988). In the Em algorithm, iterations are computationally attractive, the algorithm can usually be programmed easily, convergence is ensured, but the algorithm requires many iterations, and may converge to local optima (Titterington, Smith and Makov 1985; McLachlan and Basford 1988). It is as yet unclear which of the two methods is to be preferred in general (cf. Everitt 1984; McLachlan and Basford 1988; Mooijaart and van der Heijden 1992), but the EM algorithm has apparently been the most popular (Titterington 1990).

While most of the mixture likelihood approaches (beginning with Newcomb 1886, and later Hasselblad 1966, 1969, and Wolfe 1970) have used iterative schemes corresponding to particular instances of the EM algorithm, the formal applicability of this EM algorithm with its attractive convergence properties of the likelihood solution to finite mixture problems was recognized only after the developments of Dempster, Laird and Rubin (1977), which were later supplemented by Boyles (1983) and Wu (1983). (See also Redner and Walker 1984; Titterington 1990.) The literature on mixture

models is quite extensive, and for a complete review, we refer to the books on finite mixtures by Everitt and Hand (1981), Titterington, Smith and Makov (1985), McLachlan and Basford (1988), and Langeheine and Rost (1988).

The likelihood approach to fitting normal mixtures has been used by a large number of authors (see Titterington, Smith and Makov 1985, for a review). Whereas in these "unconditional" approaches for finite mixtures of normal distributions the mean and variance of the underlying densities are estimated, "conditional" mixture models allow for the simultaneous probabilistic classification of observations and the estimation of regression models relating covariates to the expectations of the dependent variable within latent classes. DeSarbo and Cron (1988) propose a "conditional" mixture model which enables the estimation of separate regression functions (and corresponding object memberships) in a number of classes using maximum likelihood. The model specifies a finite mixture of univariate normal densities in which the expectations of these densities are specified as linear functions of a set of explanatory variables. As such, the model generalizes the Quandt (1972), Hosmer (1974), and Quandt and Ramsey (1978) stochastic switching regression models to more than two classes, and uses maximum likelihood rather than method of moments estimators. DeSarbo and Cron (1988) use an EM algorithm to estimate their conditional mixture model. A large number of mixture regression models has now been developed (see Wedel and DeSarbo 1994, for a review). Lwin and Martin (1989), De Soete and DeSarbo (1991) and Wedel and DeSarbo (1993) developed conditional mixture binomial probit and logit regression models. Conditional mixture multinomial logit and probit regression models were developed by respectively Kamakura and Russell (1989) and Kamakura (1991). Wedel, DeSarbo, Bult and Ramaswamy (1993) proposed a univariate poisson mixture regression model. DeSarbo, Ramaswamy, Reibstein and Robinson (1993), DeSarbo, Wedel, Vriens and Ramaswamy (1992) and Jones and McLachlan (1992) developed conditional multivariate normal regression mixtures.

Whereas the application of the above regression or "conditional" mixture models has been predominantly in business research, their potential for substantive applications exist in virtually all the physical and social sciences. A large number of distributions from the exponential family have been used to describe the random variation of observations in these sciences, and the applications of generalized linear models, which include as special cases linear regression, logit and probit models, loglinear and multinomial models, inverse polynomial models, and some models used for survival data, have been enormous (cf. McCullagh and Nelder 1989). The historical development of generalized linear models can be traced to the pioneering work of Gauss, Legendre and Fisher (cf. Stigler 1986), but the term "generalized linear model" was coined by Nelder and Wedderburn (1972) who

demonstrated that the technique of iterative weighted least-squares can be used to obtain maximum likelihood estimates of the parameters of linear models with observations distributed according to some member of the exponential family. A special case of this procedure, known as the scoring method, was first introduced by Fisher (1935) in the context of Probit analysis. Developments in the estimation of linear exponential family models have been published by Dempster (1971), Berk (1972), Haberman (1977), Green (1984), and Jorgensen (1984). A comprehensive review of theory and application of generalized linear models is provided by McCullagh and Nelder (1989).

As mentioned, there are many applications of generalized linear models that may arise in the physical and social sciences. A number of such applications are listed in the book by McCullagh and Nelder (1989). However the estimation of a single set of regression coefficients across all observations may be inadequate and potentially misleading if the observations arise from a number of unknown groups in which the coefficients differ. It is in these applications that the application of the conditional finite mixture approaches may prove to be of great use. (Note that if group membership is known the group membership variable can be included in the model and there is no need to employ a mixture approach.) In this paper, we will propose a generalized linear regression latent class or mixture model, which contains the previously proposed conditional mixtures as special cases, as well as a host of other parametric specifications heretofore not dealt with in the literature. Thus, we propose a methodology which generalizes the McCullagh and Nelder (1989) work to a latent class framework. In the following sections, we will describe this mixture likelihood approach for generalized linear models, as well as a general method for its estimation based on an EM algorithm. The approach allows for the simultaneous estimation of a probabilistic classification of observations and the generalized linear model to explain the observations from a set of covariates in each class. The model is described in Section 2, and the EM algorithm in Section 3. Section 4 provides a Monte Carlo study investigating the performance of the algorithm for various distributions. In Sections 5 and 6 illustrative applications are provided, and Section 7 contains conclusions and suggestions for future research.

2. The Model

Assume that the multivariate random variables $\mathbf{y}_j = (y_{jk})$, $j = 1, \ldots, n$, and $k = 1,...,K$, arise from a superpopulation which is a mixture of a finite number (I) of populations in proportions $\pi_1, \ldots, \pi_I$, where it is not known in advance from which class a particular vector of observations arises. The probabilities π_i obey the following constraints:

$$\sum_{i=1}^{I} \pi_i = 1 \ , \ \pi_i \geq 0 \ , \ i = 1, \ldots, I \, . \tag{1}$$

We assume that the conditional probability density function of y_{jk} given that y_{jk} comes from Class i, is one taking the general form:

$$f_{jk|i}(y_{jk} \mid \theta_{ijk}, \lambda_i) = \exp \{(y_{jk}\theta_{ijk} - b(\theta_{ijk})) / a(\lambda_i) + c(y_{jk}, \lambda_i)\} \, , \tag{2}$$

for specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$, where conditional upon Class i, the y_{jk} are independently distributed with canonical parameters θ_{ijk} and means μ_{ijk}. The parameter λ_i is called the dispersion parameter, and is assumed to be constant over observations in Class i, while $a(\lambda_i) > 0$. If λ_i is known, then the distribution is a member of the exponential family with canonical parameter θ_{ijk}. The distribution may or may not be a member of the exponential family if λ_i is unknown (cf. McCullagh and Nelder 1989). Table 1, adapted from McCullagh and Nelder (1989), presents several characteristics of some common univariate distributions in the exponential family that we shall use in developing this class of models. We specify a linear predictor η_{ijk}, and a link function $g(\cdot)$ such that in Class i:

$$\eta_{ijk} = g(\mu_{ijk}) \, , \tag{3}$$

where the linear predictor is produced by P covariates $\mathbf{X}_1, \ldots, \mathbf{X}_P$ ($\mathbf{X}_p = (\mathbf{X}_{jkp})$; $p = 1, \ldots, P$, and the parameter vectors $\beta_i = (\beta_{ip})$ in Class i:

$$\eta_{ij} = \sum_{p=1}^{P} X_{jp} \, \beta_{ip} \, . \tag{4}$$

Thus, conditional upon Class i, a generalized linear model is formulated consisting of a specification of the distribution of the random variable, y_{jk}, a linear predictor, η_{ijk}, and a function $g(\cdot)$, which links the random and systematic components (so called canonical links occur when $\theta_{ijk} = \eta_{ijk}$, which are respectively the identity, log, logit, inverse and squared inverse functions for the Normal, Poisson, Binomial, Gamma and Inverse Gaussian distributions; see Table 1 from McCullagh and Nelder 1989).

The unconditional probability density function of an observation vector $\mathbf{y}_j$ can therefore be expressed in the finite mixture form (McLachlan and Basford 1988):

Table 1:
Characteristics of Some Common Univariate Distributions in the Exponential Family†

	Normal	Poisson	Binomial*	Gamma@	Inverse Gaussian
Notation	$N(\mu,\sigma^2)$	$P(\mu)$	B(m,p)/m	$G(\mu,\nu)$	$IG(\mu,\sigma^2)$
Range of y	$(-\infty,\infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0,\infty)$	$(0,\infty)$
Dispersion parameter: λ	$\lambda=\sigma^2$	1	1/m	$\lambda=\nu^{-1}$	$\lambda=\sigma^2$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y;\lambda)$	$-\frac{1}{2}\left(\frac{y^2}{\lambda}+\log(2\pi\lambda)\right)$	-log y!	$\log\binom{m}{my}$	$\nu\log(\nu y)-\log y$ $-\log\Gamma(\nu)$	$-\frac{1}{2}\{\log(2\pi\lambda y^3)+\frac{1}{\lambda y}\}$
$\mu(\theta)$ = E(Y;θ)	θ	$\exp(\theta)$	$e^\theta/(1+e^\theta)$	$-1/\theta$	$(-2\theta)^{1/2}$
Canonical Link: $\theta(\mu)$	identity	log	logit	reciprocal	$1/\mu^2$
Variance function: $V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

\+ Table 1 is taken from McCullagh and Nelder (1989).
* The mean-value parameter is denoted by μ, or by π for the binomial distribution.
@ The parameterization of the gamma distribution is such that its variance is μ^2/ν.

$$f_j(\mathbf{y}_j \mid \Phi) = \sum_{i=1}^{I} \pi_i \prod_{k=1}^{K} f_{jk|i}\,(y_{jk} \mid \beta_i, \lambda_i)\,, \tag{5}$$

where

$$\Phi' = (\pi', \beta', \lambda'); \pi = (\pi_1, \ldots, \pi_I)'; \beta = (\beta'_1, \ldots, \beta'_I)'; \lambda = (\lambda_1, \ldots, \lambda_I)'.$$

The purpose of the analysis is to estimate the parameter vector Φ. To accomplish this, we formulate the likelihood for Φ:

$$L(\Phi; y) = \prod_{j=1}^{n} f_j(y_j \mid \Phi)\,. \tag{6}$$

An estimate of Φ can be obtained by maximizing the likelihood equation (6) with respect to Φ subject to the restrictions in (1). It will be shown below that this problem can be solved using an EM-algorithm (Dempster, Laird and Rubin 1977). Once an estimate of Φ has been obtained, estimates of the posterior probability, α_{ij}, that observation j comes from latent Class i can be calculated for each observation vector $\mathbf{y}_j$, by means of Bayes' Theorem where this posterior probability is given by:

$$\alpha_{ij}(y_j, \Phi) = \frac{\pi_i \prod_{k=1}^{K} f_{jk|i}\,(y_{jk} \mid \beta_i, \lambda_i)}{\sum_{i=1}^{I} \pi_i \prod_{k=1}^{K} f_{jk|i}\,(y_{jk} \mid \beta_i, \lambda_i)}\,. \tag{7}$$

The proposed approach is similar to ordinary mixture models, except for the specification of the within class generalized linear model. Note that these methods as well as several other published methods, including univariate normal regression mixtures (DeSarbo and Cron 1988), binomial probit and logit regression mixtures (De Soete and DeSarbo 1991, Wedel and DeSarbo 1993), univariate poisson regression mixtures (Wedel et al. 1993), and latent class analysis (Goodman 1974) can be obtained as special cases of the proposed approach.

3. Estimation

3.1 The EM Algorithm

To derive the EM algorithm, we introduce non-observed data, z_{ij}, indicating if observation j belongs to latent class i: $z_{ij} = 1$ if j comes from Class i, and $z_{ij} = 0$ otherwise. It is assumed that z_{ij} are i.i.d. multinomial:

$$f(\mathbf{z}_j \mid \pi) = \prod_{i=1}^{I} \pi_i^{z_{ij}} , \tag{8}$$

where the vector $\mathbf{z}_j = (z_{ij}, \ldots, z_{Ij})'$, and we will denote the matrix $(\mathbf{z}_1, \ldots, \mathbf{z}_n)'$ by $\mathbf{Z}$ and the matrix $(\mathbf{X}_1, \ldots, \mathbf{X}_p)$ by $\mathbf{X}$. Further, it is assumed that the y_{jk} given $\mathbf{z}_j$ are conditionally independent, and that y_{jk} given $\mathbf{z}_j$ has the density:

$$f(y_{jk} \mid \mathbf{z}_j) = \prod_{i=1}^{I} f_{jk \mid i} (y_{jk} \mid \beta_i, \lambda_i)^{z_{ij}} . \tag{9}$$

With z_{ij} considered as missing data, the log-likelihood function for the complete data $\mathbf{X}$ and $\mathbf{Z}$ can be formed from equations (8) and (9) (Dempster, Laird and Rubin 1977):

$$\ln L_c(\Phi;\mathbf{y},\mathbf{Z}) = \sum_{j=1}^{n} \sum_{k=1}^{K} \sum_{i=1}^{I} z_{ij} \ln f_{jk \mid i}(y_{jk} \mid \beta_i, \lambda_i) + \sum_{j=1}^{n} \sum_{k=1}^{K} \sum_{i=1}^{I} z_{ij} \ln \pi_i . \tag{10}$$

This complete log-likelihood is maximized using an iterative EM-algorithm. In the E-step the log-likelihood is replaced by its expectation, calculated on the basis of provisional estimates of Φ. In the M-step, the expectation of $\ln L_c$ is maximized with respect to Φ to obtain new provisional estimates. The E- and M-steps are alternated until no further improvement in the likelihood function is possible. Dempster, Laird and Rubin (1977) prove that the EM-algorithm provides monotone increasing values of $\ln L_c$. Under mild conditions, $\ln L_c$ is bounded from above, in this case by 0, and convergence to at least a local optimum can be established using a limiting sums argument and/or Jensen's inequality (cf. Titterington, Smith and Makov 1985). Boyles (1983) and Wu (1983) provide a discussion of the convergence properties of the EM algorithm. The estimation of our model using the EM algorithm is conceptually similar to the estimation of ordinary mixture models (e.g., Titterington, Smith and Makov 1985) except for the solution of the within-class likelihood functions in the M-step as will be detailed below.

3.2 The E-step

In the E-step the expectation of $\ln L_c$ is calculated with respect to the conditional distribution of the non-observed data $\mathbf{Z}$, given the observed data $\mathbf{y}$ and provisional estimates of Φ. It can easily be seen that $E(\ln L_c(\Phi;\mathbf{y},\mathbf{Z}))$ is obtained by replacing z_{ij} in equation (10) by their current expected values, $E(z_{ij} \mid \mathbf{y},\Phi)$. To obtain this expectation, we first calculate the conditional

distribution of $\mathbf{y}_j$, given $\mathbf{Z}$, which is:

$$f(\mathbf{y}_j \mid \mathbf{Z},\Phi) = \prod_{i=1}^{I} \Big(\prod_{k=1}^{K} f_{jk\mid i}(y_{jk} \mid \beta_i,\lambda_i)\Big)^{z_{ij}} . \tag{11}$$

Using Bayes' rule we can now derive the conditional distribution of z_{ij} given $\mathbf{y}_j$ from equations (11) and (8), which is in turn used to calculate the required conditional expectation:

$$E(z_{ij} \mid \mathbf{y}_j,\Phi) = \frac{\pi_i \prod_{k=1}^{K} f_{jk\mid i}(y_{jk} \mid \beta_i,\lambda_i)}{\sum_{i=1}^{I} \pi_i \prod_{k=1}^{K} f_{jk\mid i}(y_{jk} \mid \beta_i,\lambda_i)} , \tag{12}$$

which is easily seen to be identical to the posterior probability α_{ij} defined in equation (7). Estimates of the posterior probabilities, $\hat{\alpha}_{ij}$ are obtained by evaluating equation (12) at the current estimates of β and λ.

3.3 The M-step

In order to maximize the expectation of $\ln L_c$ with respect to ϕ in this M-step, the non-observed data $\mathbf{Z}$ in (10) are replaced by their current expectations $\hat{\alpha}_{ij}$:

$$E(\ln L_c(\Phi;\mathbf{y},\mathbf{Z})) = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{n} \hat{\alpha}_{ij} \ln f_{jk\mid i}(y_{jk} \mid \beta_i,\lambda_i) + \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{n} \hat{\alpha}_{ij} \ln \pi_i . \tag{13}$$

As the cross-derivatives of the two terms on the right are zero, they may be maximized separately. The maximum of (13) with respect to π, subject to the constraints in equation (1), is obtained by maximizing the augmented function:

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{n} \hat{\alpha}_{ij} \ln \pi_i - \mu\Big(\sum_{i=1}^{I} \pi_i - 1\Big) , \tag{14}$$

where μ is a Lagrangian multiplier. Setting the derivative of (14) with respect to π_i equal to zero and solving for π_i yields:

$$\hat{\pi}_i = \sum_{j=1}^{n} \hat{\alpha}_{ij}/n . \tag{15}$$

Maximizing (13) with respect to β and λ is equivalent to independently maximizing each of the I expressions:

$$L_i^* = \sum_{j=1}^{n} \sum_{k=1}^{K} \hat{\alpha}_{ij} \ln f_{jk|i}(y_{jk} \mid \beta_i, \lambda_i) . \tag{16}$$

The maximization of L_i^* is equivalent to the maximization problem of the generalized linear model for the complete data, except that each observation (y_{jk}) contributes to the log-likelihood for each Class i with a known weight $\hat{\alpha}_{ij}$, which is obtained in the preceding E-step (cf. Dempster, Laird and Rubin 1977). The stationary equations are obtained by equating the first order partial derivatives of equation (16) to zero:

$$\frac{\partial L_i^*}{\partial \beta_{ip}} = \sum_{j=1}^{n} \sum_{k=1}^{K} \hat{\alpha}_{ij} \frac{\partial \ln f_{jk|i}(y_{jk} \mid \beta_i, \lambda_i)}{\partial \beta_{ip}} = 0 . \tag{17}$$

Using equations (3) and (4) and the chain rule yields:

$$\frac{\partial L_i^*}{\partial \beta_{ip}} = \sum_{j=1}^{n} \sum_{k=1}^{K} \hat{\alpha}_{ij} \frac{\partial \ln f_{jk|i}(y_{jk} \mid \theta_{ijk}, \lambda_i)}{\partial \theta_{ijk}} \frac{d\theta_{ijk}}{d\mu_{ijk}} \frac{d\mu_{ijk}}{d\eta_{ijk}} \frac{d\eta_{ijk}}{d\beta_{ip}} . \tag{18}$$

McCullagh and Nelder (1989) show that:

$$\frac{db(\theta_{ijk})}{d\theta_{ijk}} = \mu_{ijk} , \quad \frac{d\mu_{ijk}}{d\theta_{ijk}} = \frac{d^2 b(\theta_{ijk})}{d\theta_{ijk}^2} = V_{ijk} , \quad \frac{\partial \eta_{ijk}}{\partial \beta_{ip}} = X_{jkp} ,$$

where V_{ijk} is the variance function of $f_{jk|i}(y_{jk} \mid \theta_{ijk}, \lambda_i)$. Therefore:

$$\frac{\partial \ln f_{jk|i}(y_{ik} \mid \beta_i, \lambda_i)}{\partial \beta_{ip}} = \frac{r_{ijk}}{a(\lambda_i)} \frac{d\eta_{ijk}}{d\mu_{ijk}} (y_{ik} - \mu_{ijk}) X_{jkp} , \tag{19}$$

where

$$r_{ijk} = \left[\frac{d\mu_{ijk}}{d\eta_{ijk}} \right]^2 V_{ijk} , \tag{20}$$

Substituting (19) in (18), and assuming $a(\lambda_i) = \lambda_1$ (McCullagh and Nelder 1989) yields:

$$\frac{\partial L_i^*}{\partial \beta_{ip}} = \sum_{j=1}^{n} \sum_{k=1}^{K} \hat{\alpha}_{ij} r_{ijk} (y_{jk} - \mu_{ijk}) X_{jkp} \frac{d\eta_{ijk}}{d\mu_{ijk}} = 0 . \tag{21}$$

It can be easily seen that equation (21) is the ordinary stationary equation of the generalized linear model fitted across all observations, where observation j contributes to the estimating equations with fixed weight $\hat{\alpha}_{ij}$.

Therefore, for each Class i, L_i^* can be maximized by the iterative reweighted least-squares procedure proposed by Nelder and Wedderburn (1972) for ML estimation of generalized linear models, with each observation y_{jk} weighted additionally with $\hat{\alpha}_{ij}$. Given provisional estimates of β_{ip}, this procedure involves forming an adjusted dependent variable for each Class i:

$$\hat{\mu}_{ijk} = \hat{\eta}_{ijk} + (y_{jk} - \hat{\mu}_{ijk}) \frac{d\eta_{ijk}}{d\mu_{ijk}}, \tag{22}$$

where η_{ijk} is the current estimate of the linear predictor (4) and $\hat{\mu}_{ijk}$ is the corresponding fitted value derived from the link function (3). The derivative of the link function in (22) is evaluated at the current estimate of μ_{ijk}. The adjusted dependent variable $\hat{\mu}_{ijk}$ is regressed on $X_{ikp}(p = 1, \ldots, P)$, with weight $\hat{w}_{ijk}$ for each Class i, to obtain new estimates of β_{ip}, where:

$$\hat{w}_{ijk} = \left[\frac{d\mu_{ijk}}{d\eta_{ijk}}\right]^2 V_{ijk}^{-1} \hat{\alpha}_{ij}. \tag{23}$$

The derivative of μ_{ijk} in equation (23) is evaluated at the current estimate of η_{ijk}. On the basis of these revised estimates of β_{ip}, new estimates of η_{ijk} and μ_{ijk} are calculated. These values are input for a new weighted regression with the dependent variable calculated according to equation (22), and weights calculated according to equation (23). This procedure is repeated until changes in the log-likelihood (16) are sufficiently small.

Estimates of λ_i are obtained by setting the derivative of (16) with respect to λ_i equal to zero and solving for λ_i:

$$\sum_{j=1}^{n} \sum_{k=1}^{K} \hat{\alpha}_{ij} \left[\frac{y_{ik}\theta_{ijk} - b(\theta_{ijk})}{a(\lambda_i)^2} \cdot \frac{da(\lambda_i)}{d\lambda_i} + \frac{dc(y_{ik}, \lambda_i)}{d\lambda_i}\right] = 0. \tag{24}$$

The iterative weighted least-squares procedure maximizes the likelihood equation (16) according to Fishers scoring method, and is equivalent to a Newton-Raphson procedure for canonical link functions (McCullagh and Nelder 1989).

3.4 Standard Errors of the Estimates

Under typical regularity conditions, the estimators of β_i, being ML estimators, are asymptotically normal (cf. Cramér 1946; Redner and Walker 1984; DeSarbo and Cron 1988). The asymptotic covariance matrix of the estimates of β_i, conditional upon Class i can be calculated from the inverse of the observed Fisher information matrix (e.g., McLachlan and Basford 1988; Louis 1982):

$$\mathbf{I}(\hat{\Phi}) = \mathbf{X}\hat{\mathbf{W}}_i\mathbf{X}'\mathbf{a}(\hat{\lambda}_i)^{-1} , \tag{25}$$

where $\hat{\mathbf{W}}_i$ is a $(n \times n)$ diagonal matrix containing the weights $\hat{w}_{ijk}$, defined in equation (23).

3.5 Identifiability

Throughout the development of the algorithm above it was assumed that Φ is identifiable. Note that the above approach for mixture models is conceptually similar to unconditional mixture models (cf. McLachlan and Basford 1988; Titterington, Smith and Makov 1985), but now the population mean μ_{ijk} is modeled as a function of a set of covariates. Titterington, Smith and Makov (1985) provide an extensive overview of the identifiability of unconditional mixtures with specific component densities, including a survey of the literature. Here, many mixtures involving members of the exponential family including the univariate Normal, Poisson, Exponential and Gamma distributions are identifiable (see Titterington, Smith and Makov 1985). The lack of identifiability due to invariance of the likelihood under interchanging of the labels of the latent classes (Aitkin and Rubin 1985) is of no concern here, and we follow the solution of McLachlan and Basford (1988) in reporting results for only one of the possible arrangements of the classes.

3.6 The Algorithm

Summarizing the proposed EM algorithm for fitting latent class generalized linear models consists of the following steps:

1. At the first step of the iteration, $s = 0$, initialize the procedure by fixing the number of Classes, I, and generating a starting partition $\hat{\alpha}_{ij}^{(0)}$. A random starting partition can be obtained, or a rational start can be used (e.g., using K-means cluster analysis).
2. Given $\hat{\alpha}_{ij}^{(s)}$, M.L. estimates of β_i are obtained from the iterative reweighted least-squares regression procedure, using the adjusted dependent variable and weight function defined in equations (22) and (23) respectively. Starting values for this procedure are discussed by Nelder and Wedderburn (1972), where the link function is applied to the data themselves as an initial estimate of the linear predictor. Estimates of the dispersion parameters λ_i are obtained from (24), estimates of the mixing proportions π_i from (15). This step constitutes the M-step of the algorithm.

3. Convergence test: stop if $| \ln L(\Phi^{(s+1)} | \mathbf{y}) - \ln L(\phi^{(s)} | \mathbf{y}) |$ is sufficiently small.
4. Calculate new estimates of the posterior membership, $\hat{\alpha}_{ij}^{(s+1)}$, according to equation (7). This step constitutes the E-step of the algorithm.
5. Repeat steps 2 to 4.

3.7 Limitations of the Algorithm

A potential problem associated with the application of the EM algorithm to mixture problems is its convergence to local maxima. This problem is caused by the sensitivity of the algorithm to the (random) starting values used, and is well documented (McLachlan and Basford 1988, p. 16; Titterington, Smith and Makov 1985, p. 84). The convergence to local optima seems to be exacerbated when the component densities are not well separated, when the number of parameters estimated is large, and when the information embedded in each observation is limited, leading to a relatively weak posterior update in the E-step. Solutions that have been suggested to overcome this problem are having the algorithm started from a wide range of (random) starting values, by having the algorithm started from a larger number of classes working down to a smaller number, or using some clustering procedure such as K-means, applied to the dependent variable to obtain an initial partition of the data (e.g., McLachlan and Basford 1988; Banfield and Bassil 1977). Another problem associated with the EM algorithm is its slow rate of convergence (Titterington, Smith and Makov 1985, p. 88). Several procedures have been proposed to improve the rate of convergence (e.g., Peters and Walker 1978; Louis 1982; Meilijson 1989; Jones and McLachlan 1992).

3.8 Identification of the Number of Classes

When applying the above models to real data, the actual number of classes, $\mathbf{I}$, is unknown and has to be inferred from the data. The problem of identifying the number of clusters is, as yet, an inference problem in mixture models with the least satisfactory statistical treatment (Titterington 1990). Suppose we wish to test the null-hypothesis ($\mathbf{H}_o$) of $\mathbf{I}$ classes against the alternative hypothesis ($\mathbf{H}_1$) of $\mathbf{I} + 1$ classes. Unfortunately, the standard generalized likelihood ratio statistic for this test is not asymptotically distributed as chi-square since $\mathbf{H}_0$ corresponds to a boundary of the parameter space for $\mathbf{H}_1$, so that under $\mathbf{H}_0$, the generalized likelihood ration test statistic is not (asymptotically) a full rank quadratic form (Aitkin and Rubin 1985; Gosh and Sen 1985; Li and Sedransk 1988; Titterington 1990). Other procedures for determining the number of classes that have been proposed include the tests proposed by Davies (1977) which are applicable to some mixtures and based

on score statistics, and the Monte Carlo test procedure (Hope 1968) applied to mixture problems by Aitkin, Anderson and Hinde (1981), McLachlan (1987), and De Soete and DeSarbo (1991) which involves comparing the likelihood ratio statistic from the real data with a distribution of that statistic obtained from a number of datasets generated conform $\mathbf{H}_0$. Both of these procedures, however, are computationally cumbersome (cf. Titterington 1990). See McLachlan and Basford (1988, p. 21-29) for a recent overview of the literature on the determination of the number of components in mixtures.

Sclove (1987) and Bozdogan and Sclove (1984) proposed the use of Akaike's Information criterion (AIC; Akaike 1974) to determine the number of such Classes. AIC is defined in the context of our model as:

$$AIC = -2 \ln L + 2(P \cdot I + I - 1). \tag{26}$$

The major problem with the use of this criterion is that it relies on the same asymptotic properties as the likelihood ratio test (Sclove 1987; Titterington, Smith and Makov 1985). Bozdogan's (1987) consistent AIC (CAIC) imposes an additional sample size penalty on the log-likelihood. This statistic is more conservative than the AIC statistic and therefore tends to favor more parsimonious models.[1] The CAIC statistic is defined as:

$$CAIC = -2 \ln L + (P \cdot I + I - 1)(\ln(n) + 1). \tag{27}$$

Our approach to determine the appropriate number of Classes involves the use of the CAIC criterion as a heuristic, where that value of **I** is chosen that minimizes that statistic. As the CAIC statistic is burdened with the same problems as the likelihood ratio test and the AIC statistic, CAIC will be used only as a guide to the possible number of underlying groups.

4. Monte Carlo Study of Algorithm Performance

4.1 Design of the Study

In order to assess the performance of the proposed algorithm a Monte Carlo study was performed. In this study, synthetic datasets were generated according to the following six factors:

---

1. A related statistic is the Schwartz (1978) Bayesian Information Criterion (BIC): $BIC = \ln L - (P \cdot I + I - 1) \ln(n)$.

1. Distribution of the dependent variable: Normal, Poisson, Binomial or Gamma
2. Number of segments: $\mathbf{I} = 2$ or $\mathbf{I} = 4$
3. Number of x-variables: $P = 2$ or $P = 5$
4. Number of replicated measures: $K = 1$ or $K = 5$
5. Cluster separation: low or high
6. Starting values: random or K-means

The above six factors are expected to affect the performance of the proposed algorithm. The factors and their levels were chosen to reflect a variation in conditions representative of practical applications. In the construction of the synthetic data, the linear predictor was calculated for 500 hypothetical subjects according to equation (4). The x-variates were generated form a uniform distribution in the interval (0,10), and the coefficients were generated in the interval (-2,2). The (absolute) difference of the coefficients between successive segments was 0.2 in the low cluster separation condition, and 0.4 in the high cluster separation condition. The expected values for each of the distributions were generated by applying the canonical link function to the linear predictor, i.e., using the identity link for the normal distribution, the log-link for the Poisson distribution, the logit-link for the Binomial distribution, and the inverse-link for the Gamma distribution. Random variates were then generated for each distribution using the procedures described by Rubinstein (1981); algorithm IT-1 was used for the Normal distribution (p. 41) G-1 for the Gamma distribution (p. 71), algorithm IT-2 for the Binomial distribution (p. 96), and the algorithm provided for the Poisson distribution was used (p. 103). The number of trials was set to 25 for the Binomial distribution. For the Normal distribution the parameter σ_i^2 was set to 0.4 and 0.8 for $i = 1,2$ in the two-cluster condition and to 0.2 (0.2) 0.8 for $1 = 1, \ldots, 4$ in the four-cluster condition. For the Gamma distribution the parameter ν_i was set to 10 and 14 for $i = 1,2$ in the two-cluster condition and to 10 (2) 16 for $i = 1, \ldots, 4$ in the four-cluster condition. Each dataset was analyzed using two different sets of random starting values, generated form a uniform distribution in the interval (0,1) and scaled to satisfy the sum constraint in (1), and one from a partition obtained from a K-means clustering of the dependent variable, yielding 0/1 initial values.

The design used in the study was a $4 \times 3 \times 2^4$ full factorial design, which resulted in 192 observations. This design enables the investigation of all main effects and first and higher order interactions. The design has 99% power to detect effects that account for about 14% of the total variance at $p < 0.01$, and 70% power to detect effects that account for about 6% of the total variance (Cohen 1988, pp. 290, 294). The following four measures of algorithm performance were calculated, assessing computational effort,

goodness of fit, and parameter recovery:

1. ***ITER:*** the number of (major) iterations required for convergence,
2. $R^2 = 1 - (L(\hat{\Phi}_0 \mid \mathbf{y})/L(\hat{\Phi} \mid \mathbf{y}))^{2/n}$,
3. $RMS(\hat{\beta}) = \left[\sum_{i=1}^{I} \sum_{p=1}^{P} (\beta_{ip} - \hat{\beta}_{ip})^2 / \mathbf{P} \cdot \mathbf{I} \right]^{1/2}$,
4. $RMS(\hat{\pi}) = \left[\sum_{i=1}^{I} (\pi_i - \hat{\pi}_i)^2 / \mathbf{I} \right]^{1/2}$.

These statistics 3 and 4 are calculated after appropriate permutation of the recovered classes ($L(\hat{\Phi}_0 \mid \mathbf{y})$ denotes the likelihood of the model including the intercept only).

4.2 Results

The 192 observations for each dependent measure were analyzed using Analysis of Variance. In the ANOVA, the F-tests for all main effects and two-factor interactions were significant at $p < 0.01$ for the four dependent measures, while the F-tests for all three-factor interactions were not significant. The major part of the significant two-factor interactions pertained to interactions of the type of distribution with one of the other factors. Therefore, Table 2 shows the means of the four dependent measures according to type of distribution and each of the five other factors. Other relevant and significant ($p < 0.01$) interactions will be mentioned in the text.

From Table 2 it appears that the number of iterations required was highest for the Gamma and for the Normal distributions. Increasing the number of replications per subject decreases computational effort. Increasing the number of segments increases the number of iterations required. These effects differ somewhat by type of distribution, as shown in Table 2. Less well separated clusters and a smaller number of x-variables increases the number of iterations required, but only for the Normal and the Gamma distributions. (The effect of the number of x-variates may be explained by the fact that increasing the number of x-variates increases cluster separation as well). The type of start used had no effect on computational effort.

The R^2 measure showed some differences between the types of distribution, but was not affected by number of replications or the type of start used. R^2 increased for increasing numbers of x-variates and increasing numbers of segments for all distributions. A higher cluster separation likewise resulted in a higher R^2 for the Normal, Binomial and Poisson distributions.

Table 2:
Results of the Monte Carlo Study

		Iter				R^2				RMS $(\hat{\beta})$				RMS $(\hat{\pi})$			
		N**	B	P	G	N	B	P	G	N	B	P	G	N	B	P	G
REPLICATION	1	58.7^{a*}	44.2^{a}	42.6^{a}	115.4^{b}	0.713^{a}	0.943^{b}	0.800^{c}	0.814^{c}	0.018^{a}	0.077^{ab}	0.103^{b}	0.300^{c}	0.041	0.048	0.066	0.063
	5	25.8^{c}	22.4^{c}	23.3^{c}	81.7^{d}	0.822^{c}	0.953^{b}	0.839^{c}	0.817^{c}	0.032^{a}	0.037^{a}	0.093^{b}	0.143^{d}	0.024	0.027	0.044	0.061
X-VARIATES	2	53.8^{a}	36.6^{a}	38.0^{a}	76.7^{b}	0.671^{a}	0.896^{b}	0.639^{a}	0.889^{b}	0.013^{a}	0.046^{ab}	0.078^{b}	0.057^{ab}	0.034^{a}	0.027^{a}	0.075^{b}	0.056^{a}
	5	30.6^{b}	30.0^{ab}	27.9^{ab}	120.4^{c}	0.864^{b}	0.999^{c}	0.999^{c}	0.741^{a}	0.036^{a}	0.068^{ab}	0.118^{b}	0.386^{c}	0.031^{a}	0.047^{a}	0.036^{a}	0.068^{a}
SEGMENTS	2	29.8^{a}	11.0^{a}	21.2^{a}	38.1^{b}	0.634^{a}	0.900^{b}	0.659^{a}	0.772^{c}	0.008^{a}	0.010^{a}	0.027^{a}	0.174^{c}	0.031^{a}	0.006^{a}	0.045^{b}	0.027^{a}
	4	54.7^{b}	55.7^{b}	44.7^{b}	159.0^{c}	0.901^{b}	0.996^{d}	0.980^{d}	0.859^{b}	0.041^{a}	0.104^{b}	0.169^{c}	0.269^{d}	0.034^{ab}	0.068^{b}	0.066^{b}	0.097^{bc}
SEPARATION	low	56.8^{a}	30.3^{b}	37.0^{b}	111.6^{c}	0.713^{a}	0.906^{bc}	0.768^{ac}	0.826^{c}	0.019^{a}	0.022^{a}	0.050^{a}	0.212^{c}	0.021^{a}	0.016^{a}	0.068^{b}	0.066^{b}
	high	27.7^{b}	36.3^{b}	28.9^{b}	85.5^{d}	0.822^{cd}	0.989^{b}	0.871^{d}	0.804^{c}	0.030^{a}	0.092^{b}	0.146^{b}	0.231^{c}	0.044^{ab}	0.058^{b}	0.043^{ab}	0.057^{b}
START	random	40.4^{a}	33.2^{a}	34.2^{a}	98.6^{b}	0.789^{a}	0.948^{b}	0.819^{c}	0.823^{c}	0.010^{a}	0.042^{ab}	0.100^{b}	0.225^{c}	0.025	0.028	0.054	0.060
	K-means	45.9^{a}	33.5^{a}	30.4^{a}	98.6^{b}	0.723^{a}	0.948^{b}	0.819^{c}	0.800^{c}	0.034^{a}	0.086^{a}	0.094^{ab}	0.214^{c}	0.047	0.056	0.057	0.064

* Means sharing a superscript are not significantly different at $p < 0.01$.

** N = Normal, B = Binomial, P = Poisson, G = Gamma distributions.

The RMS ($\hat{\beta}$) was low for the Normal, Binomial and Poisson, but higher for the Gamma distribution. (The effect that the Gamma model performs less well may be explained from the fact that the variance of the observations increases quadratically with mean, due to which some of the classes have high variance.) For the Gamma distribution, recovery of the coefficients significantly improves for a larger number of replications, for the Binomial and Poisson distributions this effect is also present but not significant. A larger number of x-variates or a larger number of segments results in a decrease in the accuracy of parameter recovery. These effects are most pronounced for the Gamma distribution. For the Binomial and Poisson distributions increasing cluster separation decreases parameter recovery. (This may be explained from the fact that increasing cluster separation increases the variance in some of the segments.) The effects of cluster separation and number of x-variables are larger at higher numbers of segments. Type of start has no effect on parameter recovery.

The RMS ($\hat{\pi}$) was lowest for the Normal and Binomial distributions. Whereas the effects of the number of x-variates and of cluster separation are relatively small, increasing the number of segments decreases recovery of the prior probabilities, especially for the Binomial and Gamma distributions. Type of start and number of replications have no significant effects.

In general, computational performance of our procedure decreases with the number of parameters estimated (number of x-variables, number of segments) decreases with decreasing numbers of replications per subject and decreases when clusters are less well separated. Parameter recovery is negatively affected by the number of parameters estimated and by the separation of clusters relative to the within cluster variance (note that for the Binomial, Poisson and Gamma distributions within cluster variance increases for a number of clusters when cluster separation increases). A lower number of repeated measures per subject also negatively affects the parameter recovery, possibly due to a weak posterior update in the E-step. Deterioration of the performance of the algorithm under conditions in which many parameters are to be estimated and clusters are not well separated may largely be due to convergence to local optima. The type of local optimum observed most in our study was a solution in which classes occurred more than once. Such solutions are of course easily recognizable in practical applications. Local optima of this type occurred in 6% of the analyses for the Normal distribution, 10% for the Binomial distribution, 25% for the Poisson distribution, and 31% for the Gamma distribution. They also occurred more frequently for 5 than for 2 x-variables (27% versus 9%) and for 4 as compared to 2 segments (33% versus 3%). This underlies the necessity of using several random starts in applications to identify locally optimum solutions under such conditions. The use of a rational start using the K-means procedure did not improve

algorithm performance.

5. Application I: Coupon Usage

5.1 Study Design

Coupons provide consumers a reduced price on the next purchase of a specific brand of such things as grocery products, food products, laundry and cleaning services, and restaurant meals. Coupons are distributed in newspapers, magazines, by mail, in stores, or on packages. Price discounts in coupon form are reported to produce significantly larger increases in sales than equivalent promotional reductions in price (Cotton and Babb 1978). In the last decade, coupons have become an important promotional tool in marketing, and their use has grown by more than 500% (Narasimhan 1984); in 1986, a total of 190 billion coupons were distributed in the US (Vilcassim and Wittink 1987). Accordingly, managerial interest in the identification of the characteristics of consumers that respond to coupon offers has increased. Shimp and Kavas (1984) noted that although research that demonstrated the effects of coupons on sales had been abundant (cf. Nielsen 1965; Ward and Davis 1978; Dodson, Tybout, and Sternthal 1978; Bawa and Shoemaker 1989), little research had addressed the issue of consumers' response to coupons in attempts to understand couponing behavior itself.

Much of the research on couponing behavior to date has focused on the description of coupon users or coupon-prone consumers with demographic characteristics. (Lichtenstein, Netemeyer, and Burton 1990.) Teel, Williams, and Bearden (1980) demonstrated that households who would try a new product with a coupon tended to be larger, younger, and to have higher incomes. A number of authors have developed models of consumer couponing behavior to predict how coupon usage should vary by demographic characteristics of households (Blattberg, Buesing, Peacock, and Sen 1978; Narasimhan 1984; Bawa and Shoemaker 1987).

Recently, Lichtenstein, Netemeyer, and Burton (1990) argued that coupon responsive behavior is a manifestation of a number of latent psychological determinants. In relation to these underlying psychological variables, segments of consumers may exist, such as coupon prone shoppers, value conscious shoppers (Lichtenstein, Netemeyer, and Burton 1990), involved/activist shoppers, or routinized/brand loyal shoppers (Bawa and Shoemaker 1987) which exhibit differences in the extent to which coupons are used. Theories of coupon redemption behavior predict these segments to exhibit different patterns of associations of coupon usage and household demographic variables.

In research on coupon usage to date, typically a single set of regression coefficients is estimated for the entire sample (Bawa and Shoemaker 1987, 1989; Narasimhan 1984; Lichtenstein, Netemeyer, and Burton 1990). A problem of this approach is that multiple regression neglects the integer properties of the dependent variable (the number of coupons redeemed). Further, this approach may be potentially misleading if the sample consists of a number of unknown segments, as hypothesized above, in which the association of household characteristics with the coupon redemption rate differs. This circumstance requires disaggregation of the sample into groups. We will use the generalized linear model mixture likelihood approach developed in this paper to simultaneously identify such unobserved segments and estimate the associations of household demographic variables with coupon redemption in each of these segments. The data of our study pertain to household purchases of yogurt which were derived from a static sample of 2484 households who took part in a scanner panel over a 104 weeks period in Sioux Falls, South Dakota, USA, and who purchased yogurt in this period. These data were collected as follows. Panel members were given a card that was to be presented at the checkout counter of supermarkets. Code numbers on the card allowed information on the entire set of products purchased to be recorded for each individual on each purchase occasion by the store's electronic bar code reader. Information on the characteristics of the brands purchased, shelf-prices, amounts bought, coupons redeemed, purchase date, etc. were registered. Further, information on demographic characteristics, such as household size, income, and education, was collected for each of the households.

The dataset used for the analyses contains (amongst others) the following variables (average values are given in parenthesis): the dependent variable was the total number of coupons redeemed in the 104 weeks period: CPN (1.168); the total volume (in ounces $\times$ 100) of yogurt bought: TOTVOL (2.876); the average price (in dollars) paid per ounce: PRICEOZ (6.478); home ownership (1 = owned, 0 = rented): HOMEOWN (0.835); income (in 1000 dollars) per month: INCOME (2.444); the average number of hours the female head of the household works each week: FHAVHRS (24.50); the average number of hours the male head of the household works each week: MHAVHRS (30.18); the size of the household: HHSIZE (2.994); education level (1 = high school not completed, 0 = high school completed) of the female head of the household: FHEDUCN (0.322). These variables were included in our analysis given the literature on coupon redemption reviewed above, where they had been hypothesized to affect coupon usage behavior. The total volume purchased (TOTVOL) is hypothesized to have a positive association with coupon usage, as heavy users of yogurt are more often in the market and take more advantage of coupons (Narasimhan 1984; Bawa and

Shoemaker 1989). Price (PRICEOZ) should also have a positive association with couponing, as higher priced brands offer higher savings per unit through coupons (Narasimhan 1984). Since homeowners have lower storage costs, home ownership (HOMEOWN) is hypothesized to result in higher coupon redemption (Blattberg et al. 1978; Bawa and Shoemaker 1989). INCOME should have a negative association with couponing, as higher incomes tends to decrease leisure time (Narasimhan 1984; Bawa and Shoemaker 1987), while economic theory suggests that lower income households may be more price sensitive (cf. Blattberg et al. 1978; Bawa and Shoemaker 1989). Average working hours of female and male household heads (FHAVHRS respectively MHAVHRS) are hypothesized to be negatively associated with coupon usage, as time spent working negatively affects the time available for handling coupons (Blattberg et al. 1978; Narasimhan 1984; Bawa and Shoemaker 1987). The effect of household size (HHSIZE) is not equivocal. Larger families are likely to have a larger set of acceptable brands, and therefore lower substitution costs, which may result in higher coupon usage (Bawa and Shoemaker 1987, 1989). However, larger households may also have less time available for handling coupons, which would result in a lower coupon usage rate (Bawa and Shoemaker 1987). Education of the female head (FHEDUCN) is likely to positively affect couponing, as higher educated more often display variety seeking behavior which result in lower substitution costs and therefore higher coupon usage (Bawa and Shoemaker 1987), while a higher educational level will also be associated with a higher efficiency of housewives in organizing their time, and higher coupon usage consequently (Narasimhan 1984).

5.2 Aggregate I = 1 Poisson regression Results

As the variable to explained consists of counts of the number of coupons used per household in the 104 weeks period, we use the Poisson mixture regression model. This model arises as a special case of the general mixture model described above. For the Poisson mixture, equation (2) amounts to:

$$f_{jk|i}(y_{ik} \mid \mu_{ijk}) = \exp\{y_{jk}\mu_{ijk} - \exp(\mu_{ijk}) - \log(y_{jk}!)\}, \qquad (28)$$

while the link-function $g(\cdot) = \log(\cdot)$ (cf. Table 1), and $K = 1$. The parameters of the model are estimated using the EM algorithm described above, with the adjusted dependent variable and the weight function to be evaluated in the M-step of the algorithm equal to:

$$\hat{\mu}_{ijk} = \eta_{ijk} + (y_{jk} - \hat{\mu}_{ijk}) / \hat{\mu}_{ijk}, \quad \hat{w}_{ijk} = \hat{\alpha}_{ij} / \hat{\mu}_{ijk}. \qquad (29)$$

Table 3:
Coupon Usage I = 1 Aggregate Poisson Mixture Regression Parameter Estimates

INTERCEPT	-1.1620**
TOTVOL	0.0473**
PRICEOZ	0.1335**
HOMEOWN	0.4619**
INCOME	-0.0696**
FHAVHRS	-0.0021
MHAVHRS	-0.0024*
HHSIZE	0.0500**
FHEDUCN	-0.1069

* denotes $p \leq 0.05$

** denotes $p \leq 0.01$

Table 4:
Coupon Usage I = 4 Poisson Mixture Regression Parameter Estimates

	i = 1	i = 2	i = 3	i = 4
INTERCEPT	0.2500	-1.1280**	-4.2930**	-2.5230**
TOTVOL	0.0853**	0.1967**	0.1587**	0.1418**
PRICEOZ	0.1611**	0.2440**	0.1778**	0.2241**
HOMEOWN	-0.3410	0.5925**	2.5900**	-2.0660**
INCOME	-0.1081	-0.0837**	-0.0703*	-0.2434**
FHAVHRS	-0.0766**	0.0010	0.0046	0.0079
MHAVHRS	-0.0776**	-0.0024	0.0030	0.0016
HHSIZE	0.6325**	-0.0443	-0.0194	0.3493**
FHEDUCN	-3.3020**	0.0610	-0.0480	-0.2550
π_i	0.187	0.148	0.421	0.243

* denotes $p \leq 0.05$

** denotes $p \leq 0.01$

The effects of the household characteristics on coupon redemption were estimated treating all 2484 household as one group (i.e., **I** = 1). The results of the analysis are provided in Table 3. The model has a log-likelihood value of -3079.7, and an R^2 of 0.162. Table 2 shows significant effects of TOTVOL, PRICEOZ, HOMEOWN, MHAVHRS, HHSIZE, which are in the hypothesized directions. FHAVHRS has a negative effect which, although not significant, is in the hypothesized direction and of the same order of magnitude as the effect of MHAVHRS.

5.3 Poisson mixture regression Results

The issue remains whether all subjects exhibit the same association of household characteristics with coupon usage or whether segments of subjects exist that exhibit different associations. The purpose of the subsequent analysis is potentially to identify such segments of households that exhibit different associations of coupon usage with demographic variables, indicating a potentially different psychological bases for coupon usage. To address this research issue, the poisson mixture regression model was applied to the data for **I** = 2 to 5 classes. As the CAIC statistic reaches a minimum at **I** = 4, we select **I** = 4 as the appropriate number of classes. This solution has a log-likelihood of -1916.8 and a R^2 of 0.478. Table 4 presents the estimated coefficients for this 4-segment solution.

In the first segment, consisting of 18.7% of the sample, TOTVOL, PRICEOZ, FHAVHRS, MHAVHRS, HHSIZE and FHEDUCN show a significant association with the number of coupons redeemed. Compared to the other three segments, the coefficients of both working hour variables, as well as of education are much larger (moreover, in the other three segments the the effects are not significant). Apparently, the opportunity costs of the households time is a major determinant of coupon usage in this segment, which may be called *time conscious*. Perceived handling costs are higher for households with higher average working hours (Blattberg et al. 1978; Narasimhan 1984; Bawa and Shoemaker 1987), while more educated housewives are more efficient in organizing their time, leading to higher coupon usage (Narasimhan 1984). The large positive effect of household size as compared to the other segments supports the hypothesis that larger households have a larger set of preferred brands to choose from (Bawa and Shoemaker 1987, 1989), and thereby less costs of time in searching for the couponed brands, leading to higher coupon usage. Alternatively, in larger households, more household members may be active in acquiring and handling coupons which decreases the opportunity costs of time of the household, with an associated increase in coupon redemption. The effects of TOTVOL and PRICEOZ are, although significant and in the hypothesized direction, considerably smaller than in the other three segments.

The second segment consists of 14.8% of the sample, and shows significant effects of TOTVOL, PRICEOZ, HOMEOWN, and INCOME. In this segment, as compared to the other three, TOTVOL and PRICEOZ have high coefficients, while the signs of the coefficients are as hypothesized (Narasimhan 1984; Bawa and Shoemaker 1987, 1989). As the constant term indicates, average coupon usage is high in this segment. Therefore, we designated consumers in segment 2 as *coupon-prone*. This segment is hypothesized to consist of consumers that have a predominant commitment to coupon usage, and appear to rely on coupons as extrinsic cues of deals without much attention to factors such as price, value, or time costs (cf. Lichtenstein, Netemeyer, and Burton 1990). Consumers in this segment that buy larger amounts of yogurt have more opportunity to take advantage of coupons (Narasimhan 1984), while higher priced brands offer higher savings with a result that consumers resort to these brands when coupons are available (Narasimhan 1984).

Segment 3 is the largest segment and contains 42.1% of the sample. In this segment TOTVOL, PRICEOZ, and HOMEOWN are significant at the $p < 0.01$ level. The effects of TOTVOL and PRICEOZ (and INCOME, which is significant at $p < 0.05$) are as hypothesized. The coefficient for home-ownership in this segment is larger than in the other three segments, and this variable has a major effect on coupon usage. This segment is concerned with *storage facilities*. Apparently, holding costs strongly determine coupon usage in this segment. Home owners generally have more storage space than apartment dwellers, resulting in lower storage costs and greater (deal- and) coupon proneness (Blattberg et al. 1978).

Finally, Segment 4 consists of 24.3% of the sample. TOTVOL, PRICEOZ, HOMEOWN, INCOME and HHSIZE have significant effects. This segment was designated as *value conscious*, as the usage of coupons in this segment appears to be predominantly value-based: price has a significant positive effect as higher priced brands offer larger savings (Narasimhan 1984), while the coefficient of income is over two times that of its closest rival (Segment 1). A given face value of a coupon is more important to a low income household than to a high income household because of the decreasing utility of increasing income (Bawa and Shoemaker 1989). Interestingly, the coefficient of home-ownership (HOMEOWN) in this segment is opposite to the hypothesized direction: consumers with rented houses use more coupons. This finding could be related to the fact that the homeownership variable might, in relation to the cost of housing, capture additional effects of low income levels.

The results of our study demonstrate the existence of multiple market segments with different associations of household characteristics and coupon usage behavior. The study has provided important insights for managers who

wish to understand and predict the responsiveness of consumers to coupons, and allow coupon promotions to be directed at the most responsive segments through newspapers, magazines, direct mail, etc. Our analysis revealed four segments that can be interpreted on the basis of existing theories of coupon behavior (Blattberg et al. 1978; Narasimhan 1984; Bawa and Shoemaker 1987, 1989) as groups of households in which opportunity time costs, deal-proneness, storage facility, and value consciousness, appear to be the main factors related to coupon usage. Moreover, the heterogeneity in effects apparent at the segment level appear to have masked significance of effects at the aggregate level (notably the effects of FHAVHRS and FHEDUCN). It must be noted however, that while our analysis has revealed insights into coupon usage behavior, it has not addressed the issue of the effect of coupon redemption on sales which is an important issue from a management point of view. Profitability of coupon promotions cannot be measured by redemption rates, as a coupon promotion could result in a high redemption rate, but yet be unprofitable if coupons are used by consumers that would have purchased the product anyway. Further, coupons may cause an increase in sales even if they are not redeemed by acting as an advertisement and increasing awareness of the brand in question (Bawa and Shoemaker 1989). Future research should address these issues, as well as replicate our study for different brands and different product categories.

5.4 Validation

In order to investigate the sensitivity of the solution to different starting values, five I = 4 Class solutions were obtained with different (random) starting values. The log-likelihoods of the solutions fall within a dispersion range of 0.5% of each other, ranging from -1916.8 to -1926.3. We calculated the correlation between the coefficients for each solution (after appropriate permutation) and averaged the resulting correlations across the 4 classes. These 10($= 5 \times 4/2$) correlations ranged from 0.752 to 0.975. Three of the five solutions were close with correlations of 0.912, 0.945 and 0.975. Even though due to the structure of the data (i.e., one replication, large number of x-variates and segments) the model is more sensitive to local optima, there is a reasonable congruence between the solutions.

Application II: Customer Satisfaction

6.1 Study Design

With increased competitive pressures faced by firms in a global economy, more and more emphasis has been placed on listening to the "voice of

the customer'' and customer satisfaction. Howard and Sheth (1969) defined customer satisfaction as ''the buyer's cognitive state of being adequately or inadequately rewarded for the sacrifices he has undergone'' (p. 145). Oliver and DeSarbo (1988), DeSarbo, Oliver and De Soete (1986), and DeSarbo, Oliver and Rangaswamy (1989) have conducted a variety of different consumer psychology experiments to examine the impact of some five hypothesized determinants of customer satisfaction:

a. *Expectations* — prepurchase beliefs the consumer has about how the product/service will perform;
b. *Performance* — how the purchased product/service is perceived to perform for the consumer;
c. *Disconfirmation* — whether the purchased product/service performed better than expected (positive disconfirmation), the same as expected (zero disconfirmation), or worse than expected (negative disconfirmation) as experienced by the consumer;
d. *Attribution* — whether the outcome of the purchase, construed as either a success or failure, is attributed to the consumer himself or to some external agent;
e. *Inequity* — how a consumer's outcomes in an exchange compare to those received by the other party.

In these various studies, the authors show that higher levels of consumer satisfaction are typically associated with high expectations, high performance, positive disconfirmation (product performs better than expected), internal (self) attribution, and favorable inequity (lower outcome ratio for the other party). In order to quantify these effects, Oliver and DeSarbo (1988) devised simulated stock market trading scenarios and embodied these interrelated constructs vis a vis experimental designs. A stock market transaction was selected since it contained naturally all of the five factors listed above. These were expectations for the stock's performance, the ability to make the investment decision personally (internal attribution) or to rely on one's broker (external attribution), a performance outcome easily compared to expectation (disconfirmation), and a comparison of the investor's outcome (gain) to that of the broker (the commission).

In the Oliver and DeSarbo (1988) study, and as described in DeSarbo, Oliver and Rangaswamy (1989), attribution was manipulated by suggesting to the subject that the decision to buy the stock was either his/her decision (internal attribution) or was that of a broker (external attribution). For the expectation treatment, the stock was predicted either to exceed the Standard and Poor's 500 index by 5% in six months (''high'' expectations) or would just match the overall market in this time period (''low'' expectations). Performance was manipulated by describing stock as having risen 12%

(regardless of the market's performance) in six months ("high" performance) or as having risen only half that amount ("low" performance). Disconfirmation was manipulated relative to the expectation treatment. For positive disconfirmation, the stock was described as exceeding the expectation treatment level by 5%; for negative disconfirmation, the stock was described as falling short of the expectations by 5%, whether they were "high" or "low." Finally, for favorable inequity, the investor's gain structure was described so that the actual monetary outcome net of commissions was 20% above the broker's two-way (purchase, sale) commission while, for unfavorable inequity, the broker's commission exceeded the investor's gain by 20%.

Business school students at a large northeastern United States university with stock market experience (screened) were recruited to participate in a study of responses to market transaction outcomes. Oliver and DeSarbo (1988) employed a 2^5 full factorial design in a conjoint analysis task (cf. Green and Rao 1971) involving 32 profiles, but found no significant disordinal interactions. Forty-one subjects, 20 male and 11 females, were presented with the 32 hypothetical profiles in random order and asked to rate each profile on a seven point satisfaction scale (see Oliver and DeSarbo 1988, for further details of the experiment).

6.2 Aggregate I = 1 Regression Results

Given the lack of significant disordinal interactions in Oliver and DeSarbo (1988), we use a main-effect dummy variable coding in the resulting analyses representing the five treatments where a "1" was used to indicate the presence of the highest or most favorable level, and a "0" for the lowest or least favorable level. The satisfaction dependent measure was standardized by subject to eliminate differential scale and variance effects. A generalization of the DeSarbo and Cron (1988) conditional normal mixture regression methodology was employed, which arises as a special case of the general mixture regression model described above. Here, the complete likelihood function accounts for replications by subject (K = 32 profiles). As such, this approach is equivalent to a generalization of the DeSarbo, Wedel, Vriens and Ramaswamy (1992) latent class metric conjoint analysis involving finite mixtures of multivariate conditional normal distributions, where the covariance matrix, Σ_i, is restricted to equal $\sigma_i \mathbf{I}$. Table 5 presents the $\mathbf{I} = 1$ aggregate solution. All five factors are significant and the highest or most favorable level of each of the factors display positive impact on overall satisfaction as witnessed by the positive coefficients. Positive disconfirmation appears to have the greatest impact on overall satisfaction for the entire sample, followed by performance, expectations, equity, and attribution.

Table 5:
Customer Satifaction I = 1 Aggregate Normal Mixture Regression Parameter Estimates

INTERCEPT	-1.27**
ATTRIBUTION (INTERNAL)	0.13*
EXPECTATIONS (HIGH)	0.38**
PERFORMANCE (HIGH)	0.52**
INEQUITY (FAVORABLE)	0.24**
DISCONFIRMATION (POSITIVE)	1.26**

* $p \leq 0.05$

** $p \leq 0.01$

Table 6:
Customer Satisfaction I = 2 Normal Mixture Regression Parameter Estimates

	i = 1	i = 2
INTERCEPT	-1.29**	-1.24**
ATTRIBUTION (INTERNAL)	0.18**	0.08
EXPECTATIONS (HIGH)	0.44**	0.32**
PERFORMANCE (HIGH)	0.45**	0.58**
INEQUITY (FAVORABLE)	0.01	0.50**
DISCONFIRMATION (POSITIVE)	1.53**	0.99**
π_i	0.51	0.49

* $p \leq 0.05$

** $p \leq 0.01$

6.3 Normal Mixture Regression Results

In order to investigate the presence of consumer segments, the Normal mixture regression model was run for **I** = 1 to 5 classes. For this Normal mixture, equation (2) amounts to:

$$f_{jk|i}(y_{jk} \mid \mu_{ik}, \sigma_{ik}) = \exp\{(y_{jk}\mu_{ik} - \mu_{ik}^2)/\sigma_{jk}^2 - \frac{1}{2} y_{jk}^2/\sigma_{ik}^2 - \frac{1}{2} \ln(2\pi)\}, \quad (30)$$

while the identity-link is used, and $\hat{w}_{ijk} = \hat{\alpha}_{ij}$. The minimum CAIC statistic occurs at **I** = 2 latent classes. This solution has a log-likelihood of -1320.94. Table 7 presents the estimated coefficients for the **I** = 2 latent class solution. For the first Class, consisting of 51% of the sample, the impact of disconfirmation dominates in terms of influencing satisfaction judgments; the magnitude of this coefficient is over three times as large as its next competitor (performance or expectations). Note that inequity is not significant to the members of this latent Class. Here, the delight of actual stock performance exceeding high expectations (positive disconfirmation) due to one's own judgment appears to provide the greatest amount of satisfaction to the members of this latent class.

While positive disconfirmation also has the greatest impact on satisfaction for the members of latent class two, (49% of the sample) the relative magnitude of the coefficient in relation to performance is much less than in latent class one. Here, inequity is significant, while attribution is not. Members of this latent Class appear to be more concerned with their gains in relationship to what their exchange partners, the broker, receives. In addition, the members of this group do not appear to be concerned with who made the decisions.

Thus, we see evidence of factors common to both latent Classes which similarly drive customer satisfaction (high expectations, high performance, positive disconfirmation). This result is quite consistent with Oliver (1980) who documented the strong impact pf these three components of customer satisfaction working in tandem. The **I** = 2 latent class solution also provides some interesting information concerning the nature of respondent heterogeneity and the differential impact inequity and attribution have in each of the two latent classes — an aspect that is obviously masked in the aggregate solution displayed in Table 5.

6.4 Validation

To examine potential problems of local optimum solutions, five **I** = 2 Class solutions were obtained with different starting values. All values of the

log-likelihood fall within a dispersion range of .015% of each other, ranging from -1320.791 to -1320.969. The resulting average correlations between the coefficients estimated for the five solutions (after permutation) were calculated. Four correlations equaled 0.999, six equaled 1.000. Thus, there is substantial congruence between the five solutions derived from these alternative starting values.

7. Conclusions

The generalized linear mixture regression model developed in this paper finds potential applications in many physical and psychological sciences, especially if the application of single classical generalized linear models is suspected to be inadequate because of observations arising from a number of classes (i.e., sample heterogeneity) which differ in the parameters of the linear model. The mixture likelihood approach proposed simultaneously estimates the membership of the observations in an (initially specified) number of classes and the parameters of the generalized linear model that relates the dependent to the independent variables within each class, accommodating a large number of possible distributions for the dependent variable. This paper generalizes the McCullagh and Nelder (1989) work to a latent class framework, as well as previously published models (e.g., DeSarbo and Cron 1988; De Soete and DeSarbo 1991; Kamakura and Russell 1989; Jones and McLachlan 1992; Wedel and DeSarbo 1993; Wedel et al. 1993) to accommodate any distribution from the exponential family (as well as others) and provides a simple and unifying estimation approach. The EM-algorithm proposed for the estimation of the model has the advantages of being computationally attractive, of being easy to program, and of convergence being ensured. Disadvantages of the algorithm are that its convergence rate may be slow and that it is burdened with convergence to local optima. Acceleration procedures for the EM algorithm proposed by e.g. Peters and Walker (1978), Louis (1982), Meilijson (1989), and Jones and McLachlan (1992) should be investigated in future research. The problem of local optima was investigated in a Monte Carlo study and the conditions under which they may pose a problem for estimation were identified. Other issues to be addressed in future research are tests for identifying the number of classes present, finite sample properties of the significance tests, and the development of models that allow for overdispersion within classes.

References

AITKIN, M., and RUBIN, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Distributions," *Journal of the Royal Statistical Society, Series B, 47*, 67-75.

AITKIN, M., ANDERSON, D., and HINDE, J. (1981), "Statistical Modelling of Data on Teaching Styles (with discussion)," *Journal of the Royal Statistical Society, A 144*, 419-461.

AKAIKE, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control, AC-19*, 716-723.

ATKINSON, K. E. (1989), *An Introduction to Numerical Analysis*, New York: Wiley.

BANFIELD, C. F., and BASSIL, L. C. (1977), "A Transfer Algorithm for Non-Hierarchical Classification," *Applied Statistics, 26*, 206-210.

BASFORD, K. E., and MCLACHLAN, G. J. (1985), "The Mixture Method of Clustering Applied to Three-Way Data," *Journal of Classification, 2*, 109-125.

BAWA, K., and SHOEMAKER, R. W. (1987), "The Coupon-Prone Consumer: Some Findings Based on Purchase Behavior Across Product Classes," *Journal of Marketing, 51*, 99-110.

BAWA, K., and SHOEMAKER, R. W. (1989), "Analyzing Incremental Sales from a Direct Mail Coupon Promotion," *Journal of Marketing, 53*, 66-78.

BERK, R. H. (1972), "Consistency and Asymptotic Normality of MLS's for Exponential Models," *Annals of Mathematical Statistics, 43*, 193-204.

BHATTACHARYA, C. G. (1967), "A Simple Method for Resolution of a Distribution into its Gaussian Components," *Biometrics, 23*, 115-135.

BLATTBERG, R. C., BUESING, T., PEACOCK, P., and SEN, S. K. (1978), "Identifying the Deal Prone Segment," *Journal of Marketing Research, 15*, 369-377.

BOYLES, R. A. (1983), "On Convergence of the EM Algorithm," *Journal of the Royal Statistical Society, Series B, 45*, 47-50.

BOZDOGAN, H. (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions," *Psychometrika, 52*, 345-370.

BOZDOGAN, H., and SCLOVE, S. L. (1984), "Multi-sample Cluster Analysis Using Akaike's Information Criterion," *Annals of the Institute of Statistics and Mathematics, 36*, 163-180.

CASSIE, R. M. (1954), "Some Uses of Probability Paper for the Graphical Analysis of Polymodel Frequency Distributions," *Australian Journal of Marine and Freshwater Research, 5*, 513-522.

CHARLIER, C. V. L., and WICKSELL, S. D. (1924), "On the Dissection of Frequency Functions," *Arkiv für Mathematik, Astronomi och Fysik, BD 18*, 6.

COHEN, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale: Lawrence Erlbaum.

COTTON, B. C., and BABB, E. M. (1978), "Consumer Response to Promotional Deals," *Journal of Marketing, 42*, 109-113.

CRAMÉR, H. (1946), *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

DAVIES, R. B. (1977), "Hypothesis Testing When a Nuissance Parameter is Present Only Under the Alternative," *Biometrika, 64*, 247-254.

DAY, N. E. (1969), "Estimating the Components of a Mixture of two Normal Distributions," *Biometrika, 56*, 463-474.

DEMPSTER, A. P. (1971), "An Overview of Multivariate Data Analysis," *Journal of Multivariate Analysis, 1*, 316-346.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, R. B. (1977), "Maximum Likelihood from Incomplete Data via the EM-Algorithm," *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

DESARBO, W. S., OLIVER, R. L., and DE SOETE, G. (1986), "A Probabilistic Multi-Dimensional Scaling Vector Model," *Applied Psychological Measurement, 10*, 79-98.

DESARBO., W. S., and CRON, W. L. (1988), "A Maximum Likelihood Methodology for Clusterwise Linear Regression," *Journal of Classification, 5*, 249-282.

DESARBO, W. S., OLIVER, R. L., and RANGASWAMY, A. (1989), "A Simulated Annealing Methodology for Clusterwise Regression," *Psychometrika, 54*, 707-736.

DESARBO, W. S., WEDEL, M., VRIENS, M., and RAMASWAMY, V. (1992), "Latent Class Metric Conjoint Analysis," *Marketing Letters, 3*, 273-288.

DESARBO, W. S., RAMASWAMY, V., REIBSTEIN, D. J., and ROBINSON, W. T. (1993), "A Latent Pooling Methodology for Regression Analysis with Limited Time Series of Cross Sections: a PIMS Data Application," *Marketing Science, 12*, 103-124.

DE SOETE, G., and DESARBO, W. S. (1991), "A Latent Class Probit Model for Analyzing Pick Any/N Data," *Journal of Classification, 8*, 45-63.

DODSON, J. A., TYBOUT, A. M., and STERNTHAL, B. (1978), "Impact of Deals and Deal Retraction on Brand Switching," *Journal of Marketing Research, 15*, 72-81.

EVERITT, B. S. (1984), "Maximum Likelihood Estimation of the Parameters in a Mixture of two Univariate Normal Distributions: A Comparison of Different Algorithms," *Statistican, 33*, 205-215.

EVERITT, B. S., and HAND, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

FISHER, R. A. (1935), "The Case of Zero Survivors," (Appendix to Bliss, C.I. (1935)), *Annals of Applied Biology, 22*, 164-165.

FOWLKES, E. B. (1979), "Some Methods for Studying Mixtures of two Normal (Lognormal) Distributions," *Journal of the American Statistical Association, 74*, 561-575.

FRYER, I. G., and ROBERTSON, C. A. (1972), "A Comparison of Some Methods for Estimating Mixed Normal Distributions," *Biometrika, 59*, 639-648.

GHOSH, J. M., and SEN, P. K. (1985), "On the Asymptotic Performance of the Log-likelihood Ratio Statistic for the Mixture Model and Related Results," *Proceedings of the Berkeley Conference, Neyman and Kiefer, II*, Monterey: Wadsworth, 789-806.

GOODMAN, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika, 61*, 215-231.

GREEN, P. E., and RAO, V. R. (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research, 8*, 355-363.

GREEN, P. J. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives," *Journal of the Royal Statistical Society, Series B, 46*, 149-192.

HABERMAN, S. J. (1977), "Maximum Likelihood Estimates in Exponential Response Models," *Annals of Statistics, 5*, 815-841.

HARDING, I. P. (1948), "The Use of Probability Paper for the Graphical Analysis of Polymodel Frequency Distributions," *Journal of the Marine Biological Association (UK), 28*, 141-153.

HASSELBLAD, V. (1966), "Estimation of Parameters for a Mixture of Normal Distributions," *Technometrics, 8*, 431-444.

HASSELBLAD, V. (1969), "Estimation of Finite Mixtures of Distributions from the Exponential Family," *Journal of the American Statistical Association, 64*, 1459-1471.

HOPE, A. C. A. (1968), "A Simplified Monte Carlo Significance Test Procedure," *Journal of the Royal Statistical Society, Series B, 30*, 582-598.

HOSMER, D. W. (1974), "Maximum Likelihood Estimates of the Parameters of a Mixture of two Regression Lines," *Communications in Statistics, 3*, 995-1006.

JONES, P. N., and MCLACHLAN, G. J. (1992), "Fitting Finite Mixture Models in a Regression Context," *Australian Journal of Statistics, 43*, 233-240.

JONES, P. N., and MCLACHLAN, G. J. (1992), "Improving the Convergence Rate of the EM Algorithm for a Mixture Model Fitted to Grouped and Truncated Data," *Journal of Statistical Computation and Simulation, 43*, 31-44.

JORGENSEN, B. (1984), "The Delta Algorithm and GLIM," *International Statistical Review, 52*, 283-300.

KAMAKURA, W. A., and RUSSELL, G. J. (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research, 26*, 379-390.

KAMAKURA, W. A. (1991), "Estimating Flexible Distributions of Ideal-points with External Analysis of Preference," *Psychometrika, 56*, 419-448.

LANGEHEINE, R., and ROST, J. (1988), *Latent Trait and Latent Class Models*, New York: Plenum.

LAZARSFELD, P. F., and HENRY, N. W. (1968), *Latent Structure Analysis*, Boston: Houghton-Mifflin.

LI, L. A., and SEDRANSK, N. (1988), "Mixtures of Distributions: A Topological Approach," *Annals of Statistics, 16*, 1623-1634.

LICHTENSTEIN, D. R., NETEMEYER, R. G., and BURTON, S. (1990), "Distinguishing Coupon Proneness from Value Consciousness: An Acquisition-Transaction Utility Theory Perspective," *Journal of Marketing, 54*, 54-67.

LOUIS, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society, Series B, 44*, 226-233.

LWIN, T., and MARTIN, P. J. (1989), "Probits of Mixtures," *Biometrics, 45*, 721-732.

MCCULLAGH, P., and NELDER, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

MCHUGH, R. B. (1956), "Efficient Estimation and Local Identification in Latent Class Analysis," *Psychometrika, 21*, 331-347.

MCHUGH, R. B. (1958), "Note on Efficient Estimation and Local Identification in Latent Class Analysis," *Psychometrika, 23*, 273-274.

MCLACHLAN, G. J. (1982), "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis," in *Handbook of Statistics* (vol 2), Eds., P. R. Krishnaiah and L. N. Kanal, Amsterdam: North-Holland, 199-208.

MCLACHLAN, G. J. (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Applied Statistics, 36*, 318-324.

MCLACHLAN, G. J., and BASFORD, K. E. (1988), *Mixture Models: Inference and Application to Clustering*, New York: Marcel Dekker.

MEILIJSON, J. (1989), "A Fast Improvement of the RM Algorithm on Its Own Terms," *Journal of the Royal Statistical Society, B 51*, 127-138.

MOOIJAART, A., and VAN DER HEIJDEN, P. G. M. (1992), "The EM Algorithm for Latent Class Analysis with Constraints," *Psychometrika, 57*, 261-271.

NARASIMHAN, C. (1984), "A Price Discrimination Theory of Coupons," *Marketing Science, 3*, 125-145.

NELDER, J. A., and WEDDERBURN, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A, 135*, 370-384.

NEWCOMB, S. (1886), "A Generalized Theory of the Combination of Observations So As To Obtain the Best Result," *American Journal of Mathematics, 8*, 343-366.

NIELSEN, A. C. (1965), "The Impact of Retail Coupons," *Journal of Marketing* (October), 11-15.

OLIVER, R. L. (1980), "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions," *Journal of Marketing Research, 17*, 460-469.

OLIVER, R. L., and DESARBO, W. S. (1988), "Response Determinants in Satisfaction Judgments," *Journal of Consumer Research, 14*, 495-507.

PEARSON, K. (1894), "Contributions to the Mathematical Theory of Evolution," *Philosophical Transactions, A, 185*, 71-110.

PETERS, B. C., and WALKER, H. F. (1978), "An Iterative Procedure for Obtaining Maximum Likelihood Estimates of the Parameters of a Mixture of Normal Distributions," *Journal of Applied Mathematics, 35*, 362-378.

QUANDT, R. E., and RAMSEY, J. B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association, 73*, 730-738.

QUANDT, R. E. (1972), "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association, 67*, 306-310.

REDNER, R. A., and WALKER, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review, 26*, 195-239.

RUBINSTEIN, R. Y. (1981), *Simulation and the Monte Carlo Method*, New York: Wiley.

SCHWARTZ, G. (1978), "Estimating the Dimensions of a Model," *Annals of Statistics, 6*, 461-464.

SCLOVE, S. L. (1987), "Applications of Model-Selection Criteria to some Problems in Multivariate Analysis," *Psychometrika, 52*, 333-343.

SHIMP, T. A., and KAVAS, A. (1984), "The Theory of Reasoned Action Applied to Coupon Usage," *Journal of Consumer Research, 11*, 795-809.

STIGLER, S. M. (1986), *The History of Statistics*, Cambridge, Mass: Harvard University Press.

SYMONS, M. J. (1981), "Clustering Criteria and Multivariate Normal Mixtures," *Biometrics, 37*, 35-43.

TEEL, J. E., WILLIAMS, R. H., and BEARDEN, W. O. (1980), "Correlates of Consumer Susceptibility to Coupons in New Grocery Product Introductions," *Journal of Advertising, 3*, 31-35.

TEICHER, H. (1961), "Identifiability of Mixtures," *Annals of Mathematical Statistics, 31*, 55-73.

TITTERINGTON, D. M. (1990), "Some Recent Research in the Analysis of Mixture Distributions," *Statistics, 4*, 619-641.

TITTERINGTON, D. M., SMITH, A. F. M., and MAKOV, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

THOMAS, E. A. C. (1966), "Mathematical Models for the Clustered Firing of Single Cortical Neurons," *British Journal of Mathematical and Statistical Psychology, 19*, 151-162.

VILCASSIM, N. J., and WITTINK, D. R. (1987), "Supporting a Higher Shelf Price Through Coupon Distributions," *Journal of Consumer Marketing, 4*, 29-39.

WARD, R. W., and DAVIS, J. E. (1978), "A Pooled Cross-Section Time Series Model of Coupon Promotions," *American Journal of Agricultural Economics*, (August), 193-401.

WEDEL, M., and DESARBO, W. S. (1994), "A Review of Recent Developments in Latent Class Regression Models," in *Advanced Methods of Marketing Research*, Ed., R. Bagozzi, 352-388.

WEDEL, M., DESARBO, W. S., BULT, J. R., and RAMASWAMY, V. (1993), "A Latent Class Poisson Regression Model for Heterogeneous Count Data With an Application to Direct Mail," *Journal of Applied Econometrics, 8*, 397-411.

WEDEL, M., and DESARBO, W. S. (1993), "A Latent Class Binomial Logit Methodology for the Analysis of Paired Comparison Data: An Application Reinvestigating the Determinants of Perceived Risk," *Decision Sciences, 24*, 1157-1170.

WOLFE, J. H. (1970), "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research, 5*, 329-350.

WU, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *Annals of Statistics, 11*, 95-103.