19

LATENT STRUCTURE REGRESSION

WAYNE S. DESARBO
Pennsylvania State University

WAGNER A. KAMAKURA
Duke University

MICHEL WEDEL
University of Michigan

Ordinary least squares (OLS), or multiple regression as it is also designated, is the most popular method in applied marketing research to summarize the relationship between a predesignated set of independent variables and a single dependent variable. To set the scene and establish notation for the subsequent exposition, we start to briefly summarize this well-known method. Let

$i = 1, \ldots, I$ consumers,

$j = 1, \ldots, J$ independent variables,

Y_i = the value of the dependent variable for consumer i,

X_{ij} = the value of the jth independent variable for consumer i,

b_j = the value of the jth OLS regression coefficient,

e_i = error for consumer i.

Then, the standard linear multiple regression model can be expressed as

$$y_i = \sum_{j=1}^{J} X_{ij} b_j + e_i \tag{1}$$

or, in matrix form,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \tag{2}$$

where $\mathbf{y} = ((y_i))$, $\mathbf{X} = ((X_{ij}))$, $\mathbf{b} = ((b_j))$, and $\mathbf{e} = ((e_i))$. Given an independent sample of consumers/observations for $\mathbf{y}$ and $\mathbf{X}$, one is typically interested in estimating b_j by minimizing the following error sums of squares:

$$\begin{aligned} \underset{b_j}{Min} Z &= \sum_{i=1}^{I} \left[y_i - \sum_{j=1}^{J} X_{ij} b_j \right]^2 \\ &= (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\ &= \sum_{i=1}^{I} e_i^2 = \mathbf{e}'\mathbf{e}. \end{aligned} \tag{3}$$

Johnston (1984) and others derive (by taking partial derivatives of (3), setting them equal to zero, and then solving) the well-known analytical expression for estimating **b** and σ^2 that minimize (3):

$$\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y},$$

$$\sigma^2 = \mathbf{e'e}/(I - J). \quad (4)$$

Maddala (1976) and others show that if the assumption is made that the random vector **e** is multivariate normally distributed, then the likelihood function can be written (assuming $E(\mathbf{ee'}) = \sigma^2\mathbf{I}$, where **I** is an identity matrix), as

$$L(y|\mathbf{b}, \sigma^2) = (2\pi\sigma^2)\exp\left[-\frac{(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{2\sigma^2}\right], \quad (5)$$

and the corresponding maximum likelihood estimates for **b** and σ^2 that maximize the likelihood function in (5) are identical to those obtained from least squares estimation (i.e., in Expression (4)).

There are many applications that arise in marketing, however, where the estimation of a single set of regression coefficients may prove to be "misleading." Consider, for example, the small illustrative synthetic data set provided in Table 19.1 with $J = 5$ independent variables $(X_1, \ldots, X_5)$ and $I = 50$ consumers. Suppose this sample of 50 consumers was taken and, unknown to the marketing researcher, the first 25 consumers were extracted from a drastically different market segment than the last 25 consumers. That is, suppose this sample represents two radically different market segments that were *unknown* a priori. Table 19.2 represents the aggregate regression analysis computed over all $I = 50$ customers. As shown, all five independent variables are *not* significantly different from zero, as indicated by the various t statistics. The calculated F statistic is also not significant, which indicates that there would be no significant linear relationship between the five independent variables and the dependent variable. Indeed, $R^2 = 0.06$, and the adjusted R^2 is actually negative. The marketing researcher concludes that nothing in the regression equation is significant.

Now, suppose this marketing researcher understands that there are two disparate sets of observations, has access to a variable that identifies these two sets, and a priori analyzes each market segment separately. Here, in addition to Equation (1), we now have further notation:

$k = 1, \ldots, K$ market segments (known, with $K = 2$ in the example),

b_{jk} = the value of the jth regression coefficient for the kth segment,

σ_k^2 = the variance term for the kth segment,

z_{ik} = the 0/1 membership of consumer i in segment k.

We now assume y_j has a univariate normal distribution, conditional upon known segment membership:

$$y_{i/k} \sim \sum_{k=1}^{K} z_{ik}(2\pi\sigma_k^2)^{-1/2}\exp\left[\frac{-(y_i - \mathbf{X_i b_k})^2}{2\sigma_k^2}\right]. \quad (6)$$

The purpose here is to estimate the b_{jk} and σ_k^2, which is done in the example by conducting one separate regression analysis for Market Segment 1 (the first 25 observations) and a different regression analysis for Market Segment 2 (the last 25 observations). Tables 19.3 and 19.4 display the corresponding results for each market segment. As seen from the t tests, almost every regression coefficient is significant in both estimated equations, as is the total regression function (F test). The corresponding R^2 are 0.993 and 0.994, indicating near-perfect fits. As shown by the estimated regression coefficients, Segment 1's regression function is near the opposite (in sign) that of Segment 2. Thus, we see an illustration of how an aggregate regression function can mask the underlying structure of consumer heterogeneity in the presence of unknown and discrete market segments, as well as how that problem is easily resolved if the actual market segments are known.

Where such market segments are precisely known in advance, the marketing researcher merely has to divide/classify the total sample of observations and conduct separate regression analyses per market segment, as has been

Table 19.1 Synthetic Regression Data

X_1	X_2	X_3	X_4	X_5	Y
0.995	1.893	−1.291	−0.560	0.508	−3.287
−1.268	−0.988	0.244	0.716	−2.261	−8.254
0.135	0.725	−0.511	0.712	−1.551	−11.867
0.687	0.223	1.441	0.098	−0.352	10.859
1.020	−0.355	−0.660	0.090	−1.544	−4.733
0.313	0.999	0.365	−0.991	1.000	12.191
0.584	−1.775	−1.702	1.843	−1.267	−25.508
−1.342	−0.976	−1.381	0.290	1.525	−11.996
−0.159	−0.132	0.217	−1.643	−0.278	9.533
1.089	0.629	1.148	−1.138	−1.299	15.716
0.322	−0.983	−1.673	0.101	−0.287	−14.693
0.082	−0.009	1.068	−1.023	0.565	16.379
−0.923	2.327	−0.989	1.637	1.023	−14.245
−0.138	0.122	−0.967	−0.365	−0.535	−6.939
−0.650	1.494	0.040	−0.086	−0.806	−1.569
1.227	0.475	2.090	−0.294	0.643	21.104
0.987	0.389	−1.791	0.758	1.574	−12.890
−1.786	0.484	0.495	0.545	−0.448	−1.547
−1.330	−1.583	0.606	0.948	−0.137	−3.635
−0.213	−1.400	−0.942	−0.659	0.042	−1.644
−0.305	0.531	0.627	0.068	0.214	3.152
−0.907	−0.296	−0.528	−0.423	−0.603	−3.425
−0.051	−1.204	0.553	−0.044	0.373	4.460
0.247	−0.288	−0.623	0.371	−1.285	−9.167
−0.261	1.024	−2.361	0.188	−1.183	−20.099
−1.072	−0.044	0.486	0.176	0.788	−2.941
−0.179	0.345	−0.655	2.860	2.211	18.596
−0.777	−0.390	−1.587	0.682	1.438	14.039
−1.311	−0.671	0.715	0.440	0.181	−0.625
0.145	0.321	−1.310	−0.085	0.843	7.317
−1.229	−1.133	−0.148	0.865	−0.007	8.799
0.421	0.147	−0.215	0.075	−0.492	2.029
1.088	−1.352	−0.531	−0.222	−0.048	0.930
−1.210	−0.100	−0.935	−0.142	−0.742	10.174
−0.202	1.088	−0.514	1.695	0.146	14.756
0.651	−0.118	−2.153	−1.061	0.036	9.358
−0.908	0.304	0.403	0.332	0.099	0.087
−0.242	−1.206	0.520	0.647	−1.529	3.671
2.010	−0.312	1.958	−0.275	0.289	−19.342
0.170	−0.957	0.580	−0.027	−1.039	−1.826
−1.594	0.188	−0.331	1.407	−2.152	16.582
1.041	−1.211	−0.610	0.036	−0.252	5.588
−0.391	0.143	−0.186	−0.574	0.227	−3.652
1.227	−0.476	0.908	−0.733	−0.146	−14.729
−0.774	−3.297	−0.893	1.470	−0.468	18.368
0.807	0.030	1.033	−0.533	−0.585	−12.020
1.597	−1.308	0.837	−0.702	0.455	−11.534
1.595	−0.281	−1.685	1.117	−0.632	18.330
−0.664	0.113	−0.098	−0.423	−0.432	−0.882
1.529	−2.106	0.592	0.126	1.466	−6.664

Table 19.2 SPSS Aggregate Regression Analysis for Entire Sample

Model Summary

Model	R	R-*Square*	*Adjusted* R-*Square*	*Standard Error of the Estimate*
1	.244[a]	.060	–.047	11.836924

a. Predictors: (Constant), X5, X3, X2, X1, X4.

ANOVA[b]

Model		*Sum of Squares*	df	*Mean Square*	F	*Sig.*
1	Regression	390.412	5	78.082	.557	.732[a]
	Residual	6164.962	44	140.113		
	Total	6555.374	49			

a. Predictors: (Constant), X5, X3, X2, X1, X4.
b. Dependent Variable: Y.

Coefficients[a]

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		B	*Std. Error*	*Beta*	t	*Sig.*
1	(Constant)	.864	1.776		.487	.629
	X1	–1.633	1.878	–.136	–.870	.389
	X2	.352	1.671	.031	.211	.834
	X3	1.932	1.693	.175	1.142	.260
	X4	.493	2.165	.037	.228	.821
	X5	1.639	1.767	.137	.928	.359

a. Dependent Variable: Y.

illustrated above. However, there are many circumstances where either the underlying market segments are unknown or the existing market segments, defined a priori, fail to display such heterogeneity in terms of the structure of the regression function (e.g., their market consumption behavior). For example, market segments defined solely on the basis of demographics tend not to display structurally different regression functions when modeling purchase behavior. Plotting the data may be beneficial when there are one or two independent variables but would be of little assistance for the data in Table 19.1, where there are five independent variables, as detailed plotting in six dimensions is impossible.

The purpose of this chapter is to review a class of models for problems of this nature where the underlying basis of customer heterogeneity (i.e., discrete market segments) is *unknown* a priori. Our objective is to *simultaneously* estimate the number of market segments, their size and composition, and the segment-specific regression coefficients. In essence, we

Table 19.3 SPSS Regression Analysis for Market Segment 1

Model Summary

Model	R	R-*Square*	*Adjusted* R-*Square*	*Standard Error of the Estimate*
1	.996[a]	.993	.991	1.120445

a. Predictors: (Constant), X5, X3, X1, X2, X4.

ANOVA[b]

Model		*Sum of Squares*	df	*Mean Square*	F	*Sig.*
1	Regression	3360.091	5	672.018	535.303	.000[a]
	Residual	23.853	19	1.255		
	Total	3383.943	24			

a. Predictors: (Constant), X5, X3, X1, X2, X4.
b. Dependent Variable: Y.

Coefficients[a]

Model		*Unstandardized Coefficients*		*Standardized Coefficients*		
		B	*Std. Error*	*Beta*	t	*Sig.*
1	(Constant)	.308	.240		1.284	.215
	X1	1.520	.281	.109	5.401	.000
	X2	.370	.222	.033	1.670	.111
	X3	7.533	.216	.721	34.800	.000
	X4	−5.868	.312	−.407	−18.798	.000
	X5	1.877	.232	.160	8.074	.000

a. Dependent Variable: Y.

use the same data as in ordinary multiple regression and attempt to learn much more about the underlying structure of the data. The problem is essentially the same as that in Equation (6), but now, the segment indicators z_{ik} are *unknown.* We will first continue to treat problems involving assumptions of normality of the dependent variable, which are very common in marketing, and later discuss alternative formulations.

This class of (latent structure regression) methods enables marketers to engage in response-based segmentation (i.e., identifying segments that are homogeneous in how they respond to market stimuli) and enables prediction of a dependent measure of interest such as attitude, preference, or choice. Since predicting measures of marketing effectiveness is both common and important, response-based segmentation is a particularly powerful approach to segmentation. Many applied and academic studies have used this approach (cf. Wedel & Kamakura, 2000). Previously, marketing researchers have

Table 19.4 SPSS Regression Analysis for Market Segment 2

Model Summary

Model	R	R-*Square*	*Adjusted* R-*Square*	*Standard Error of the Estimate*
1	.997[a]	.994	.992	.960690

a. Predictors: (Constant), X5, X2, X3, X4, X1.

ANOVA[b]

Model		*Sum of Squares*	df	*Mean Square*	F	*Sig.*
1	Regression	2781.179	5	556.236	602.688	.000[a]
	Residual	17.536	19	.923		
	Total	2798.714	24			

a. Predictors: (Constant), X5, X2, X3, X4, X1.
b. Dependent Variable: Y.

Coefficients[a]

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		B	*Std. Error*	*Beta*	t	*Sig.*
1	(Constant)	−.154	.231		−.666	.514
	X1	−1.258	.209	−.125	−6.014	.000
	X2	−.509	.217	−.044	−2.344	.030
	X3	−7.546	.211	−.679	−35.783	.000
	X4	6.257	.245	.522	25.542	.000
	X5	−1.762	.220	−.152	−8.022	.000

a. Dependent Variable: Y.

used combinations of approaches to do segmentation and prediction (e.g., cluster analysis followed by regression analysis), but these approaches have been shown to be much less effective or even not applicable at all in many situations (Wedel & Kamakura, 2000). Two important areas of application of these response-based segmentation models are conjoint analysis and scanner data modeling. Here, one can estimate the importance of product attributes for preference formation in segments of consumers that are not identified a priori but are formed based on their latent, unobservable preferences for product attributes (cf. DeSarbo & Ramaswamy, 1995; DeSarbo, Wedel, Vriens, & Ramaswamy, 1992). Or one estimates the effect of price and promotion on consumer choice, grouping consumers into segments that are maximally different in terms of these effects. Other applications of response-based segmentation include the analyses of transaction databases where segments of customers are derived for optimal targeting of offerings. In essence, one can apply such latent structure regression methods in the same scenarios as where ordinary multiple regression is typically applied.

Thus, latent structure regression models are a powerful and useful tool for marketing research, enabling the development of response models of consumer behavior and grouping of consumers based on differences in the response models across segments. Latent structure regression provides a powerful approach to model data that can be considered continuous, such as many rating scales in marketing (if the number of scale points is 7 or larger). Here, the regression framework allows one to build quite general representations of consumer behavior through the inclusion of main effects, interaction effects, nonlinear effects, dummy variables, and so on. In combination with the latent structure that facilitates segmentation based on such general response models, this gives rise to a broad class of response-based segmentation models. The great appeal of this approach derives from the flexibility of the linear modeling framework to represent consumer behavior. However, these latent structure regression models cannot be estimated with OLS but have to be estimated with algorithms such as the E-M algorithm. The E-M algorithm is a very robust algorithm that iterates between estimating the probability that each consumer belongs to each segment (E step) and fitting a regression weighted with these probabilities for each segment (M step) until convergence. The next section of this chapter describes the latent structure regression framework and the E-M estimation algorithm. Also, we discuss several technical issues that are relevant, including the simultaneous estimation of the posterior probabilities of membership of consumers into classes. These probabilities reflect the (un)certainty with which each respondent can be classified into every segment, and have important applications in targeting segments or individuals (e.g., in applications in direct marketing and customer relationship management [CRM]). Also, we will discuss how to decide which number of segments describes the data best, for which various (information) statistics have been proposed.

However, many phenomena in marketing are not well captured by the normal distribution that is assumed in the latent structure regression approach. For example, in many marketing studies, choice data are collected where the dependent variable indicates whether a product is chosen. Or, one wishes to investigate the frequency with which products or services are used, where the dependent variable is a count taking on integer values of zero and larger. Or, one collects data on times, such as response latencies or retention times. Such times typically are positive and skewed, features that are not very well captured by the normal distribution. Consequently, for each of these types of variables, a different distribution needs to be chosen. Mostly, one chooses the binomial or multinomial distribution for choices, the Poisson distribution for counts, and the gamma distribution for times. While such a regression of such outcome variables is seemingly more complex, it is in fact not much more so. These distributions (normal, Poisson, binomial, and gamma) are all members of the so-called exponential family of distributions, and since they are in that same family, all these distributions can be accommodated through the same principles, as a single class rather than as a collection of special cases. Regression models that involve an outcome variable in the exponential family are very well studied, often applied, and are called "generalized linear models." It takes not much more human or computer effort to estimate them than linear regression models. Models that are well-known and often applied in marketing, including logistic regression, probit regression, and count data regression, are special cases. Rather than using least squares, generalized linear models are estimated with the method of maximum likelihood.

The likelihood is a measure of fit. It is, loosely speaking, obtained as the probability that a specific set of model parameters has generated the data set at hand. The distribution of the dependent variable (normal, binomial, etc.) is instrumental in computing those probabilities for each consumer, and since consumers are assumed to come from a random sample of the population, the probability of the sample is the product of the probabilities for each consumer. In many cases, the likelihood function needs to be maximized numerically to obtain the parameter estimates. A variety of numerical search algorithms can be used for that purpose, and in most cases, these converge in a few iterations only (because the likelihood is concave in the parameters for generalized linear models). Given the popularity of generalized linear models for the analysis of consumer behavior, their extension to a latent structure framework opens up a broad array of response-based segmentation models. Because it allows the simultaneous segmentation and estimation of generalized linear models for each class, this class of models has received much interest in marketing theory and practice, particularly after Windows-based software (GLIMMIX, Latent Gold, Sawtooth CBC) had been developed to facilitate their estimation. Most of the packages apply E-M-type algorithms for the estimation of latent structure regression models. Now, marketers can do response-based segmentation in conjoint and scanner data applications and identify segments in such a way that consumer response to the marketing mix in these segments is optimal. These mixtures of generalized linear models constitute a very broad class of models, where one can choose from various types of dependent measures, select very flexible predictor functions as described above, and deal with consumer heterogeneity and the identification of segments all at the same time. The second part of this chapter describes the details of this even broader class of models, and technical details are given in the appendix. We provide applications to synthetic and empirical data. Finally, we discuss in more depth why these models are conceptually appealing to marketing managers and what their advantages and disadvantages are over other approaches that assume a continuous distribution of consumer heterogeneity.

Normal Latent Structure Regression

The Model Framework

In addition to the notation developed above, following DeSarbo and Cron (1988), let

$k = 1, \ldots, K$ market segments (now unknown),

b_{jk} = the value of the jth regression coefficient for the kth segment,

σ_k^2 = the variance term for the kth segment,

λ_k = the size (proportion) of segment k.

We now assume y_i is distributed as a finite sum or "mixture" of conditional univariate normal densities:

$$y_i \sim \sum_{k=1}^{K} \lambda_k f_{ik}(y_i | X_{ij}, \sigma_k^2, b_{jk}) = \sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{X_i b_k})^2}{2\sigma_k^2}\right], \quad (7)$$

where $\mathbf{X_i} = ((X_{ij}))_i$ and $\mathbf{b_k} = ((b_j))_k$. That is, we assume an independent sample of subjects' (or observations') dependent variable $y_1, y_2, \ldots, y_1$ drawn randomly from a finite mixture of conditional normal densities of underlying groups or segments in unknown proportions $\lambda_1, \lambda_2, \ldots, \lambda_K$. That is, (7) is equivalent to (6), but the unobserved segment proportions λ_K replace the observed 0/1 segment indicators z_{ik}. Thus, in (7), the contributions of the segment-specific regression functions are weighted with the segment sizes, as in (6), but here those segment proportions are unknown.

The parameters of the latent structure regression model can be estimated by maximizing a likelihood function. The likelihood is based on the assumption of a conditional normal distribution of the dependent variable within each segment. Once the likelihood is formed, maximizing it over the parameters provides those parameter values that are most likely to have produced the data. Given a sample of I independent subjects/observations, one can thus form a log-likelihood (maximizing the log of the likelihood is equivalent to maximizing the likelihood, but simpler) to estimate λ_k σ_k^2 and b_{jk} (as is done in ordinary multiple regression using maximum likelihood estimation [MLE], as developed earlier) and is described in detail in the appendix. As mentioned above, here one relaxes the assumption that the segments and their sizes are known a priori and estimates the segment sizes λ_k rather than being able to use the observed segment indicators z_{ik} to run separate regressions for each segment. However, after these model parameters are estimated, one may still want to assign each of the subjects i to one of the k segments. One does this by obtaining estimates of the posterior probability of membership for any consumer i in each derived segment k_i, $\hat{p}_{jk}$ once one has obtained estimates of the parameters $\hat{\lambda}_k$, $\hat{\sigma}_k^2$ and $\hat{b}_{jk}$ via the application of Bayes rule, as shown in the appendix. This result renders a "fuzzy" clustering of the I consumers/observations, which is analogous to the "hard" partition of the sample obtained if segment composition were known a priori in the form of the indicators z_{ik}. But one could form similar "hard" partitions by applying the rule (which is the optimal Bayes rule):

$$\text{Assign } i \text{ to } k \text{ if } \hat{p}_{ik} > \hat{p}_{il} \text{ for all } l \neq k = l \ldots K.$$

The appendix describes the technical aspects of the estimation procedure devised by DeSarbo and Cron (1988) to estimate λ_k, b_{jk}, and σ_k^2 that maximize the log likelihood in (9), given **y, X,** and a value of K. Finally, note that estimates of λ_k, σ_k^2, P_{ik}, and b_{jk} are obtained all *simultaneously* by latent structure regression estimation methods such as the procedure outlined in the appendix. Maximum likelihood estimates of the class sizes (λ_k) *and of the regression parameters within each class* (σ_k^2 and b_{jk}) are obtained via the E-M algorithm, which involves two steps. In the expectation step, cases are classified into the latent classes based on the current parameter estimates, using the Bayes rule (shown in the appendix). In the maximization step, a weighted least squares regression is fit to each latent class, using the current membership probabilities obtained from the previous expectation step as weights. The E-M steps are iterated until convergence, leading to maximum likelihood estimates of the model parameters.

Determining K^* = the Optimal Number of Segments

When applying the above models to data, the actual number of segments K^* is unknown and must be inferred from the data. Unfortunately, the standard likelihood ratio statistic (the difference of the maximized likelihood of the two models with K and $K + 1$ segments) for this test does not apply. Wedel and Kamakura (2000) discuss a number of alternative heuristics for selecting K^*. One class of criteria for investigating the number of segments that is frequently used is various information criteria. Such measures attempt to balance the increase in fit obtained against the larger number of parameters estimated for models with more segments. Basically, such criteria impose a penalty on the likelihood that is monotone with the number of free parameters estimated:

$$C = -21n\ L_{max} + Pd. \qquad (8)$$

Here, P is the number of free parameters estimated ($JK + 2K - 1$), and d is some constant. The constant imposes a penalty on the likelihood, which balances the increase in fit (more parameters yield a higher likelihood) against the additional number of parameters estimated. The classical Akaike (1974) information criterion, AIC, arises when $d = 2$. Two criteria that penalize the likelihood more heavily than AIC are the Bayesian information criterion (BIC) and the consistent Akaike information criterion (CAIC). For those criteria, $d = \ln(I)$ and $d = \ln(I + 1)$, respectively. Note that the CAIC penalizes the likelihood somewhat more than BIC, although the two criteria render similar decisions as to K^* in most applications. Both statistics impose a larger penalty on the likelihood than AIC and are more conservative than the AIC statistic in that they tend to favor more parsimonious models (i.e., models with fewer segments). Studies by Bozdogan (1987, 1994) indicate that CAIC is preferable in general for mixture models. Bozdogan also proposed the modified AIC3 criterion, for which $d = 3$.

The preceding heuristics account for overparameterization as large numbers of segments are derived, but one must also ensure that the segments are sufficiently separated for the solution that is selected. To assess the separation of the segments, an entropy statistic can be formulated to investigate the degree of separation in the estimated posterior probabilities. E_k is a relative and nonlinear measure that is bounded between 0 and 1. Values close to 1 indicate that the derived segments are well separated. A value close to 0, indicating that all the posteriors are equal for each observation, is of concern as it implies that the centroids of the segments are not sufficiently well separated. Celeux and

Soromenho (1996) proposed a normalized entropy criterion (NEC) for the selection of the number of segments in mixture models. A potential problem with the measure is that it is not defined for $K = 1$. Hence, NEC should preferably be applied in conjunction with one of the other informational criteria in determining the number of segments for $K > 2$. See Wedel and Kamakura (2000) for a discussion of other heuristics. (Note that one can also compute an overall R^2 for such latent structure regression solutions as in ordinary multiple regression.)

The major difficulty one faces with the use of these heuristics is that there are occasions when they may point to different optimal values of K^*. This is especially pronounced with the information-based heuristics, where AIC and AIC3 tend to select overparameterized (K^* too large) solutions as opposed to the more conservative BIC and CAIC measures. Another problem is that in general, these measures will select a larger number of segments when the sample size increases. In a recent extensive Monte Carlo study, Dias (2004) shows that if one compares AIC, BIC, CAIC, AIC3, ICOMP, and NEC statistics for the selection of the number of segments in a (discrete) latent class model, AIC3 appears to outperform the other measures, with an overall hit rate of more than 70%. BIC and CAIC performed relatively well, with hit rates around 60%. We therefore recommend these three measures.

The Synthetic Data Example

Recall the synthetic data illustration presented in Table 19.1 and how the overall aggregate regression function (see Table 19.2) masks the true data structure. The true structure was revealed in Tables 19.3 and 19.4 once the previously unknown market segments were revealed. Now, let's assume we know nothing in advance about the underlying market segments and perform the latent structure regression analysis on the Table 19.1 data. Table 19.5a presents the various heuristics for $k = 1, \ldots, 4$. As aptly noted, the procedure accurately reveals the structure underlying the data as *all* heuristics select the $K^* = 2$ segment solution. Table 19.5b presents the $K^* = 2$ segment solution parameter values, which are indeed similar to those found in Tables 19.3 and 19.4 when we assumed the market segments were known. Thus, for this synthetic example, the latent structure regression performs very well in *simultaneously* estimating the number of market segments, their size and composition, and the segment-specific regression parameters.

Alternative Latent Structure Regression Specifications

The number of applications of generalized linear models (which include as special cases linear regression, logit and probit models, log-linear, and multinomial models) in marketing has been enormous. Generalized linear models (Nelder & Wedderburn, 1972) are regression models in which the dependent variable is specified to be distributed according to one of the members of the exponential family (see Table 19.6). Generalized linear models deal with continuous variables, which can be specified to follow a normal, gamma, or exponential distribution; for discrete variables, the binomial, multinomial, Poisson, or negative binomial distributions can be used, among others. Each of those distributions has extensive applications in marketing. For example, the normal is used for continuous measures, often rating scales, although these are strictly speaking categorical. Nevertheless, in applications, one often treats them as continuous. Consumer choices of products are discrete and often described with the binomial distribution; purchase counts are mostly modeled with a Poisson distribution, which describes them to be positive and accommodates their nature as integer (0, 1, 2, . . .) variables. Response and retention times are positive and skewed and therefore best described with a gamma (or exponential as a special case) distribution. The expectation of these dependent variables is modeled as a function of a set of explanatory variables as in standard multiple regression models discussed earlier. Thus, generalized linear models provide a very rich framework for modeling marketing response, enabling one to deal independently with a wide variety of stochastic forms of the dependent variable, reflecting uncertainty in consumer behavior or the marketer's knowledge of it, as

Table 19.5 Latent Structure Regression Results for the Synthetic Data in Table 19.1

(a) Results

K*	*ln* L	*# Parameters*	*AIC*	*AIC3*	*BIC*	*CAIC*	R^2	*Entropy*	*# Iterations*
1	–191.31	7	396.62	403.62	410.01	417.01	0.06	0.00	2
2	–99.09	15	228.18	243.18	256.86	271.86	0.99	0.93	9
3	–92.17	23	230.33	253.33	274.31	297.31	1.00	0.90	16
4	–85.89	31	233.78	264.78	293.06	324.06	1.00	0.90	18

(b) The* K* *=2 Solution

	k = *1*	k = *2*
*b*0	0.289	−0.181
*b*1	1.584	−1.238
*b*2	0.358	−0.533
*b*3	7.513	−7.567
*b*4	−5.866	6.256
*b*5	1.875	−1.764
$\hat{\sigma}^2$	0.99	0.85
λ	0.52	0.48

Note: AIC = Akaike information criterion; AIC3 = modified Akaike information criterion, where *d* = 3; BIC = Bayesian information criterion; CAIC = consistent Akaike information criterion.

well as the richness of the structure of the marketing problem in terms of how explanatory variables affect the expectation of the dependent variable through linear, nonlinear (e.g., quadratic), and interactive functions.

However, as we have seen, the estimation of a single aggregate (generalized linear) regression equation across all consumers in a sample may be inadequate if the consumers belong to a number of unknown segments in which the regression coefficients differ. An alternative motivation for such mixture regression models comes from random coefficient specifications. In random coefficient models, the coefficients of a generalized linear model are assumed to follow some distribution across the population to account for heterogeneity. Often, the normal distribution is assumed, but a discrete distribution can be assumed instead (e.g., the multinomial in the case of a finite mixture model). If one assumes a normal distribution of the coefficients, each consumer has his or her own coefficient, but these are constrained to come from a continuous normal distribution. The random coefficient models are not as easy to estimate and necessitate the application of, for example, simulated likelihood or hierarchical Bayes methods (Wedel et al., 1999). However, if one does not have a good prior about the shape of the population distribution of the parameters and assumes that it is discrete (binomial, multinomial), then a latent structure model arises. In that case, a finite number of support points with accompanying probability mass are used to approximate the distribution of the coefficients over the population of consumers (i.e., instead of estimating a parameter for each individual, a limited set of parameters is estimated for relatively homogeneous groups of individuals). This is the finite mixture (or alternatively called latent class/latent structure) formulation, which is often more convenient than the continuous heterogeneity approximation because it is easy to interpret the coefficients for each class, interpret

Table 19.6 Some Distributions in the Exponential Family

Distribution	*Notation*	*Distribution Function*	*Link*
Binomial	$B(K,\mu)$	$\binom{K}{y}\left(\frac{m}{K}\right)^{y}\left(1-\frac{\mu}{K}\right)^{(K-y)}$	$\text{In}(\mu/(K-\mu))$
Poisson	$p(\mu)$	$\frac{e^{-\mu}\mu^{y}}{y!}$	$\text{In}(\mu)$
Negative binomial	$NB(\mu,\upsilon)$	$\left(\frac{v}{v+\mu}\right)^{v}\frac{\Gamma(v+y)}{y!\Gamma(v)}\left(\frac{\mu}{v+\mu}\right)^{y}$	$\text{In}(\mu/(\upsilon+\mu))$
Multinomial	$M(\mu)$	$\prod_{k=1}^{K}\mu_k^{y_k}$	$\text{In}(\mu/(\mu_k+\mu_K))$
Normal	$N(\mu,\sigma)$	$\frac{1}{\sqrt{2\pi\sigma}}\exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right]$	μ
Multivariate normal	$MVN(\mu,\Sigma)$	$\frac{1}{(2\pi)^{K/2}\lvert\Sigma\rvert^{1/2}}\exp[-1/2(y_n-\mu)'\Sigma]$	μ
Exponential	$E(\mu)$	$\frac{1}{\mu}\exp\left[-\frac{y}{\mu}\right]$	$\frac{1}{\mu}$
Gamma	$G(\mu,\upsilon)$	$\frac{1}{\Gamma(v)}\left(\frac{yv}{\mu}\right)^{v-1}\exp\left[-\frac{vy}{\mu}\right]$	$\frac{1}{\mu}$
Dirichlet	$D(\mu)$	$\frac{\Gamma\left(\sum_{k=1}^{K}\mu_k\right)\prod_{k=1}^{K}y_k^{\mu_k-1}}{\prod_{k=1}^{K}\Gamma(\mu_k)}$	$\text{In}(\mu/(\mu_k+\mu_K))$

Source: Adapted from McCullagh and Nelder (1989).

the classes as segments connecting well to strategic marketing, and estimate the coefficients. The maximization of the likelihood is relatively simple using such algorithms as the E-M algorithm, where the application of continuous distributions of the coefficients necessitates the evaluation of high dimensional integration. In addition, the finite mixture formulation has much conceptual appeal to marketers since it connects elegantly to the theory of market segmentation, which uses discrete market segments to explain differences in consumer behavior and product differentiation.

Following Wedel and DeSarbo (1995), we assume that the multivariate random variables $y_i = (y_{ir})$, $i = 1, \ldots, I$ and $r = 1, \ldots, R$ (replications) arise from a superpopulation that is a mixture of a finite number (K) of populations in proportions $\lambda_1, \ldots, \lambda_k$, where it is not known in advance from which class a particular vector of observations arises. Here, $r = 1, \ldots, R$ denotes repeated observations of the dependent variable for each consumer i. Although such repeated measures can also be accommodated in the latent class linear regression model described above, we have omitted them for ease of exposition. In fact, repeated measures provide additional information for the classification of cases into the latent class in the E step of the E-M algorithm, leading to better defined latent classes or segments. It is relatively easy to deal with such repeated measures in particular when they are considered to be independent. In that case, the modeling exercise involves stringing out the measurements on the I consumers and the R replications in one long I^*R vector $y_i = (y_{ir})$. Dependencies of the repeated measures, if assumed to be present, can be modeled through the structural component of the model (i.e., by including lagged effects in the independent variables, assuming the repeated measures are made in time).

Thus far, the setup is similar to that for the mixture of normal regressions described above, except that we now have multiple replications (r) for each consumer i. However, we now assume that the conditional probability density function of y_{ir}, given that y_{ir} comes from segment k, takes the general exponential family form as described in the appendix. Conditional upon segment k, the y_{ir} are independently distributed according to a particular distribution in this class with means μ_{kir}. These are the means of the normal or Poisson distributions or the expected probability of choice for the binomial. Table 19.6, adapted from McCullagh and Nelder (1989), presents several characteristics of some common univariate distributions in the exponential family that are used in developing this class of models. We specify a linear predictor, η_{kir}, that captures the structural part of consumer response that one wishes to describe, such that in each segment, the linear predictor is produced by J covariates $X_1, \ldots, X_J$ ($\mathbf{X}_J = (X_{irj})$), $j = 1, \ldots, J$, and the parameter vectors $\boldsymbol{\beta}_k = (\beta_{ij})$ in segment k:

$$\eta_{kir} = \sum_{j=1}^{J} X_{irj}\beta_{kj}. \tag{9}$$

Thus, conditional upon segment k, a generalized linear model is formulated with the specification of the distribution of the random variable, y_{ir}, a linear predictor η_{kir}. A link-function $g(\cdot)$ links the mean of the random and the systematic components. For the normal distribution, the link function is linear ($\mu_{kir} = \eta_{kir}$), the Poisson uses a convenient log-link function ($\log(\mu_{kir}) = \eta_{kir}$), and the binomial uses the logit link function ($\log(\mu_{kir}/1 - \mu_{kir}) = \eta_{kir}$); see McCullagh and Nelder (1989). These link functions have the property that the predicted values of the model are consistent with the assumptions on the values that the dependent variable can take on (i.e., any value for the normal, positive values for the Poisson, and between 0 and 1 for the binomial). For example, the binomial distribution and the logit link leads to the popular logistic regression model of choice. Note that the linear predictor can contain main effects, interactions, nonlinear effects, dummy-coded discrete variables, and more, which, in combination with the stochastic nature of the dependent variable and the mixture formulation accommodating segments and heterogeneity, affords a greatly flexible and general modeling framework.

The purpose of the latent structure analysis is to estimate the parameters characterizing the segments and the within-segment regression models. To accomplish this, we formulate the likelihood function and maximize it, as described in the appendix. The likelihood can be maximized using an E-M algorithm (Dempster, Laird, & Rubin, 1977) or through direct numerical maximization using such algorithms as Newton-Raphson. Once estimates of the parameters have been obtained in any iteration of estimation, estimates of the posterior probability, p_{ki}, that observation i comes from latent segment k can be calculated for each observation vector, y_i, by means of Bayes theorem as before.

The proposed approach is similar to finite 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 & Cron, 1988), binomial probit and logit regression mixtures (DeSoete & DeSarbo, 1991; Wedel & DeSarbo, 1993), univariate Poisson regression mixtures (Wedel, DeSarbo, Bult, & Ramaswamy, 1993), latent class multinomial logit model (Kamakura & Russell, 1989), and latent class analysis (Goodman, 1974)—can be obtained as special cases of this framework.

Application: The Dodge Viper

In this illustrative application, we extend the analysis originally presented by Malhotra (1996) to demonstrate how the latent structure regression approach can provide valuable insights otherwise overlooked by traditional regression analysis. Malhotra reports an interesting psychographic study conducted in 1989 among visitors of auto shows as part of a concept test for the Dodge Viper. Dodge had produced a concept version of a new automobile (to be later named the Dodge Viper) that was featured in a number of auto shows throughout the United States; it wanted to know what type of market existed for such a car priced around $60,000. Initially, Dodge hypothesized that the car would be attractive to a Yuppie crowd—highly educated, affluent baby boomers who tended to prefer imported vehicles. After a series of in-depth personal interviews with consumers, a list of some 30 psychographic questions was constructed to measure important constructs such as patriotism, styling, prestige, personality, and so on. These appear in Table 19.7 and are measured on 9-point Likert scales (1 = *definitely disagree* to 9 = *definitely agree*). In addition, an intention/likelihood to buy question was asked in an effort to understand how these psychological constructs affected buying intentions for the new concept car. Knowledge of these key constructs could drive advertising decisions and the positioning of the automobile. Targeted respondents were interviewed at such auto shows, and some 400 completed surveys were collected.

Traditional Analysis Results

Given the large number of intercorrelated psychographic variables present in the study, Malhotra (1996) first conducted a principal components analysis (PCA) among the 30 psychographic items to attempt to reduce the total number of dimensions of the data. Using the "eigenvalue-greater-than1" rule, Malhotra extracted nine principal components that account for 78.53% of the total variance in these data. Table 19.8 presents the varimax rotated PCA loadings for the nine-component solution. Malhotra labels these varimax-rotated factors as follows:

Component	*Label*
1	Financial concerns
2	Style consciousness
3	Societal concern
4	Patriotism
5	Optimism
6	Adventurous
7	Opinion leadership
8	Traditionalism
9	Intensity

As mentioned earlier, the questionnaire also includes an intention-to-buy measure as the marketing group wanted to understand how to properly position the Viper. Table 19.9 presents the aggregate total sample corresponding regression analysis on these PCA scores for the nine derived components. Table 19.9 indicates that the total regression equation is significant, with $R^2 = 0.522$. As seen in Table 19.9, only the first and eighth components (financial concerns and traditionalism) are not statistically significant. Overall, this model suggests that the major psychographic constructs explaining likelihood of purchase are adventurous, style, and optimism. As Malhotra (1996) concludes, marketing communications should be designed to appeal to these three major traits (see also DeSarbo & Hausman, 2006).

Latent Structure Regression Results

Table 19.10a presents the various heuristics for solutions $k = 1, 2, 3, 4$ using the latent structure regression procedure. While the log-likelihood steadily improves with increasing k, we do not see consistent heuristics. The entropy, AIC, and AIC3 point to $K^* = 2$, while the BIC and CAIC indicate $K^* = 1$. Looking at R^2 alone, $K^* = 4$ appears most appropriate. As mentioned earlier, many times (unlike the synthetic data results), these heuristics do not all render consistent selection of the optimal K^*. Given the dominance of $K^* = 2$ on half of these measures, let's examine that solution's results. Table 19.10b presents the parameter estimates for the $K^* = 2$ solution. As seen, Segment 1 has approximately 91 consumers in it, and Segment 2 has 309

Table 19.7 Psychographic Items in Malhotra's (1996) Dodge Viper Study

1.	I am in very good physical condition.	(V1)
2.	When I must choose between the two, I usually dress for fashion, not comfort.	(V2)
3.	I have more stylish clothes than most of my friends.	(V3)
4.	I want to look a little different from others.	(V4)
5.	Life is too short not to take some gambles.	(V5)
6.	I am not concerned about the ozone layer.	(V6)
7.	I think the government is doing too much to control pollution.	(V7)
8.	Basically, society today is fine.	(V8)
9.	I don't have time to volunteer for charities.	(V9)
10.	Our family is not too heavily in debt today.	(V10)
11.	I like to pay cash for everything I buy.	(V11)
12.	I pretty much spend for today and let tomorrow bring what it will.	(V12)
13.	I use credit cards because I can pay the bill off slowly.	(V13)
14.	I seldom use coupons when I shop.	(V14)
15.	Interest rates are low enough to allow me to buy what I want.	(V15)
16.	I have more self-confidence than most of my friends.	(V16)
17.	I like to be considered a leader.	(V17)
18.	Others often ask me to help them out of a jam.	(V18)
19.	Children are the most important things in a marriage.	(V19)
20.	I would rather spend a quiet evening at home than go out to a party.	(V20)
21.	American-made cars can't compare with foreign-made cars.	(V21)
22.	The government should restrict imports of products from Japan.	(V22)
23.	Americans should always try to buy American products.	(V23)
24.	I would like to take a trip around the world.	(V24)
25.	I wish I could leave my present life and do something entirely different.	(V25)
26.	I am usually among the first to try new products.	(V26)
27.	I like to work hard and play hard.	(V27)
28.	Skeptical predictions are usually wrong.	(V28)
29.	I can do anything I set my mind to.	(V29)
30.	Five years from now, my income will be a lot higher than it is now.	(V30)

Source: Case and data adapted from Malhotra (1996).

consumers. The derived regression coefficients for Segment 2 are quite similar to the aggregate regression function displayed in Table 19.9, while the smaller segment's estimated regression coefficients are quite different. For Segment 1, higher intention to purchase the Viper would correspond to higher adventurism, opinion leadership, and societal concerns and lower intensity, style consciousness, financial concerns, and traditionalism. For Segment 2, higher intention to purchase the Viper would relate to higher adventurism, style consciousness, optimism, patriotism, and societal concerns. Thus, the analysis renders some common factors (adventurism and societal concerns) as well as segment-unique factors (Segment 1: opinion leadership, intensity, style consciousness, financial concerns, traditionalism; Segment 2: style consciousness, optimism, and patriotism) upon which to promote the vehicle.

Table 19.8 PCA Varimax-Rotated Components

Rotated Component Matrix[a]

	Component								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
V2	–.013	.907	.022	.054	.005	.072	.004	.001	.006
V3	.012	.905	.035	.069	.013	.053	–.009	.008	–.009
V4	.040	.826	.007	–.002	.094	.036	.027	.043	.015
V5	.005	.648	.006	–.093	.129	.036	.072	–.034	–.045
V6	.005	.162	.196	.156	.164	.445	.019	–.003	–.236
V7	.000	.029	.764	–.027	.051	.165	.073	.011	–.392
V8	.034	–.011	.837	–.022	.039	.131	.032	–.016	–.284
V9	.020	.030	.859	.050	–.041	–.087	.034	–.004	.367
V10	.011	.041	.858	.048	–.032	–.095	.011	–.011	.356
V11	.896	–.005	.013	.001	–.038	.037	.004	–.012	–.021
V12	.902	–.007	.006	–.010	–.021	.002	–.028	–.002	.007
V13	.937	–.008	.005	–.011	–.021	.011	.012	–.027	.003
V14	.937	.019	.029	–.004	–.010	.013	.010	.004	.013
V15	.871	.027	.020	–.007	.052	–.036	–.028	.015	.009
V16	.758	.025	–.010	.017	.061	–.023	–.028	.079	.065
V17	.001	.037	.035	.080	.011	.063	.903	–.016	–.001
V18	–.025	.011	.042	.056	.082	.016	.935	–.032	–.015
V19	–.028	.054	.032	.006	.020	–.045	.877	.028	–.001
V20	.026	.016	.003	.006	–.015	–.015	.001	.901	–.056
V21	.025	–.004	–.018	.009	–.040	.018	–.017	.900	.035
V22	–.048	–.014	.027	.955	.020	.074	.042	.003	–.003
V23	.001	.011	.010	.955	.054	.082	.070	.006	–.016
V24	.035	.010	.004	.915	.055	.056	.034	.008	.009
V25	–.044	.061	.033	.079	.043	.912	.031	–.005	.038
V26	–.004	.061	–.015	.063	.064	.923	.008	–.015	.048
V27	.045	–.002	–.036	–.010	.101	.708	–.020	.030	.328
V28	.053	–.023	.045	–.006	.105	.221	–.005	–.021	.618
V29	.005	.101	.036	.041	.950	.117	.034	–.037	.020
V30	.018	.070	.002	.028	.955	.126	.038	–.028	.047
V31	.003	.098	–.021	.062	.896	.047	.046	.000	.048

Note: Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalization.

a. Rotation converged in seven iterations.

One other aspect in Table 19.10 deserves mention and may reveal some insights as to these two segments. Segment 1's estimate of σ^2 is much smaller than that for Segment 2, indicating that there appears to be substantial homogeneity among these 91 respondents. Unfortunately, demographic, behavioral, and present car ownership data were not available to profile these derived segments.

Discussion

The finite mixture latent structure regression model approach has proven to be a powerful approach to analyzing marketing data. Its popularity has been propelled by the wide availability of (commercial) software such as GLIMMIX (Wedel, 2001) and Latent Gold (Vermunt & Magdison, 2005). Heterogeneity is of key importance in

Table 19.9 Aggregate Regression Model Fit on PCA Scores

ANOVA[b]

Model		*Sum of Squares*	df	*Mean Square*	F	*Sig.*
1	Regression	1395.267	9	155.030	47.401	.000[a]
	Residual	1275.531	390	3.271		
	Total	2670.797	399			

a. Predictors: (Constant), PCA9, PCA8, PCA7, PCA6, PCA5, PCA4, PCA3, PCA2, PCA1.
b. Dependent Variable: PREF.

Model Summary

Model	R	R-*Square*	*Adjusted* R-*Square*	*Standard Error of the Estimate*
1	.723[a]	.522	.511	1.808

a. Predictors: (Constant), PCA9, PCA8, PCA7, PCA6, PCA5, PCA4, PCA3, PCA2, PCA1.

Coefficients[a]

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		B	*Std. Error*	*Beta*	t	*Sig.*
1	(Constant)	5.023	.090		55.544	.000
	PCA1	.111	.091	.043	1.225	.221
	PCA2	.774	.091	.299	8.550	.000
	PCA3	.385	.091	.149	4.249	.000
	PCA4	.388	.091	.150	4.282	.000
	PCA5	.762	.091	.294	8.413	.000
	PCA6	1.371	.091	.530	15.143	.000
	PCA7	.201	.091	.078	2.225	.027
	PCA8	.096	.091	.037	1.063	.288
	PCA9	–.278	.091	–.108	–3.073	.002

a. Dependent Variable: PREF.

marketing and has since long been the very topic of its investigation. Finite mixture models provide a flexible and relatively simple representation of heterogeneity that ties in very well with the notion of the existence of market segments (for a discussion, see Wedel & Kamakura, 2000; Wedel et al., 1999). The ability to capture unobserved heterogeneity with a finite mixture model depends on the volume and quality of information available from each consumer. An important distinction between the formulations of the normal latent structure regression presented in (A2) and the generalized linear latent structure model in (A29) is that the latter formulation considers replications within each consumer. With more replications per consumer, additional information is available regarding individual behavior, leading to a clearer definition of the latent segments.

Table 19.10 Latent Structure Regression Results for the Viper PCA Scores

(a) Results

K	*ln* L	*# Parameters*	*AIC*	*AIC3*	*BIC*	*CAIC*	R^2	*Entropy*	*# Iterations*
1	–799.51	11	1621.01	1632.01	1664.92	1675.92	0.52	0.00	2
2	–772.99	23	1591.98	1614.98	1683.78	1706.78	0.61	0.91	242
3	–765.70	35	1601.39	1636.39	1741.09	1776.09	0.74	0.74	260
4	–754.85	47	1603.69	1650.69	1791.29	1838.29	0.94	0.58	200

(b) Parameter Estimates for* K *= 2 Solution

	k = *1*	k = *2*
*b*0	6.482**	4.904**
*b*1	–6.05**	0.149
*b*2	–0.209	0.829**
*b*3	0.685**	0.323**
*b*4	–0.190	0.453**
*b*5	0.293**	0.753**
*b*6	1.140**	1.389**
*b*7	0.910**	0.163
*b*8	–0.340**	0.164
*b*9	–1.757**	–0.177
$\hat{\sigma}^2$	0.100	1.701
λ	0.077	0.923

Note: AIC = Akaike information criterion; AIC3 = modified Akaike information criterion, where $d = 3$; BIC = Bayesian information criterion; CAIC = consistent Akaike information criterion.

$^*p < .05$. $^{**}p < .01$.

Despite its conceptual and computational attractions, the finite mixture model approach is not without its limitations. The most important ones have been cited by Allenby and Rossi (1999). The main critiques against the finite mixture model approach are that it assumes that, within each given segment, all consumers behave exactly in the same manner (i.e., have precisely the same values of the regression coefficients in the linear predictor). Whether this is an acceptable assumption depends on the manager's purpose. If her or his purpose is to classify the population of consumers into fewer market segments containing relatively homogeneous (compared across segments) consumers, the assumption of homogeneity within each segment might be adequate. However, some marketing researchers use the results from finite a mixture model to obtain individual-level estimates, leading to the "convex hull" problem: Individual-level predictions of finite mixture models are a weighted combination of the segment-level regression functions, weighted with the posterior membership probabilities. Therefore, predictions are constrained to lie within the boundaries provided by 0/1 posterior membership weights. This would imply that finite mixture models might do less well in capturing the extremes of distributions of heterogeneity and predicted values. For those interested in obtaining individual-level estimates from a finite mixture model, Kamakura and Wedel (2004) propose a procedure for integrating out the asymptotic distribution of the parameter estimates, rather than using point estimates, which alleviates (but does not eliminate) the "convex hull" problem.

A commonly proposed alternative is continuous mixtures where the regression parameters follow a normal distribution (Allenby & Rossi, 1999). Such a specification would not suffer from the two problems mentioned above. However, the assumption that parameters themselves have a normal distribution might not be valid. Therefore, it seems that the extent to which the finite mixture model versus the continuous mixture model fits well is an empirical issue (Wedel et al., 1999). Indeed, a recent series of extensive studies by Andrews and coauthors (e.g., see Andrews, Ainslie, & Currim, 2002) reveals that there is little empirical evidence that the continuous heterogeneity approach dominates the finite mixture approach. These studies revealed that the differences are small and that there is little reason to prefer one approach to the other. One important exception to this general rule is the case of discrete choice data with few observations per subject (small R above); here it appeared that the continuous mixture does not do very well and is inferior to the discrete heterogeneity approach.

Whereas these results should give applied researchers confidence that the finite mixture approach is valid if one deals with heterogeneous regression models, there remain caveats. In all analyses, one needs to exercise great care in checking the model specification. Distributional assumptions can be incorrect, important variables may have been omitted in the linear predictors, the number of assumed segments is too small or too large, or within-segment heterogeneity can be nonnegligible. The marketing and statistics literatures have provided a range of model checks and test statistics to investigate these issues (see McLachlan, 2000; Wedel & Kamakura, 2000), and we strongly suggest that these tests and checks be used routinely in practice. In case model violations are discovered, the model must be extended and adapted.

A host of extended finite mixture models are now available that allow for demographic variables to explain segment membership simultaneously (concomitant variable mixtures), allow segment composition to change over time according to a first-order process and joint segmentation based on sets of disparate variables (hidden Markov models), accommodate latent variables in the form of multidimensional scaling (MDS) and/or factor representations (Wedel & DeSarbo, 1996), and allow for within-segment (normal) heterogeneity (Allenby, Arora, & Ginter, 1998). In addition, there are interesting applications of finite mixture models to problems such as data fusion (Kamakura & Wedel, 1997). While these approaches seem as yet to be less often used in the practice of marketing research, we believe that they may hold great promise and are worth considering.

Appendix

The Normal Latent Structure Regression Algorithm (DeSarbo & Cron, 1988) and Its Mixture GLM (Wedel & DeSarbo, 1995) Extension

A. The Normal Latent Structure Regression Algorithm

Given a sample of I independent subjects/observations, one can thus form a log-likelihood:

$$L = \prod_{i=1}^{I} \left[\sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{X}_i \mathbf{b}_k)^2}{2\sigma_k^2} \right] \right] \tag{A1}$$

or

$$\ln L = \sum_{i=1}^{I} \ln \left[\sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{X}_i \mathbf{b}_k)^2}{2\sigma_k^2} \right] \right] \tag{A2}$$

Given K, $\mathbf{y}$, and $\mathbf{X}$, one wishes to estimate λ_k, σ_k^2, and b_{jk} to maximize L or $\ln L$ (as is done in ordinary multiple regression using MLE, as developed earlier), where

$$0 < \lambda_k < 1, \tag{A3}$$

$$\sum_{k=1}^{K} \lambda_k = 1, \tag{A4}$$

$$\sigma_k^2 > 0. \tag{A5}$$

The maximum likelihood estimates of λ_k, $\mathbf{b}_k$, σ_k^2, and p_{ik} are found by initially forming an augmented log-likelihood function to reflect the λ_k constraints in expressions (A3) to (A5), and

$$\Phi = \sum_{i=1}^{I} \ln\left[\sum_{k=1}^{K} \lambda_k f_{ik}(y_i|X_{ij}, \sigma_k^2, b_{jk})\right] - \mu\left(\sum_k \lambda_k - 1\right). \tag{A6}$$

The resulting maximum likelihood stationary equations are obtained by equating the first-order partial derivatives of the augmented log-likelihood function in (A2) to 0:

$$\frac{\partial \Phi}{\partial \lambda_k} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} f_{ik}(*) - \mu = 0, \tag{A7}$$

$$\frac{\partial \Phi}{\partial \sigma_k^2} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} \lambda_k \frac{\partial f_{ik}(*)}{\partial \sigma_k^2} = 0, \tag{A8}$$

$$\frac{\partial \Phi}{\partial b_{jk}} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} \lambda_k \frac{\partial f_{ik}(*)}{\partial b_{jk}} = 0, \tag{A9}$$

where $f_{ik}(*)$ is used for $f_{ik}(y_i \mid X_{ij}, \sigma_k^2, b_{jk})$. To estimate μ, we multiply both sides of Equation (A2) by λ_k and then sum both sides over *k:*

$$\sum_{i=1}^{I} \frac{\sum_k \lambda_k f_{ik}(*)}{\sum_k \lambda_k f_{ik}(*)} - \mu \sum_k \lambda_k = 0 \tag{A10}$$

or

$$\hat{\mu} = I. \tag{A11}$$

To estimate λ_k, we multiply both sides of Equation (A2) by λ_k and simplify:

$$\sum_{i=1}^{I} \frac{\lambda_k f_{ik}(*)}{\sum_k \lambda_k f_{ik}(*)} - \lambda_k \mu = 0, \tag{A12}$$

or

$$\sum_{i=1}^{I} \hat{p}_{ik} - \lambda_k I = 0, \tag{A13}$$

and

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{I} \hat{p}_{ik}}{I}. \tag{A14}$$

To estimate σ_k^2 and b_{jk}, we use the definition of $\hat{p}_{ik}$ in (13) and reexpress (A8) and (A9) as:

$$\frac{\partial \Phi}{\partial \sigma_k^2} = \sum_{i=1}^{I} \hat{p}_{ik} \frac{\partial \log f_{ik}(*)}{\partial \sigma_k^2} = 0, \tag{A15}$$

$$\frac{\partial \Phi}{\partial b_{jk}} = \sum_{i=1}^{I} \hat{p}_{ik} \frac{\partial \log f_{ik}(*)}{\partial b_{jk}} = 0. \tag{A16}$$

Thus, the maximum likelihood equations for estimating the parameters σ_k^2 and b_{jk} are weighted averages of the maximum likelihood equations $\frac{\partial \log f_{ik}(*)}{\partial \theta} = 0$, where θ reflects the parameter of interest arising from each component separately, and the weights are the posterior probabilities of membership of the subjects/observations in each cluster. These posterior membership probabilities are obtained via the application of Bayes rule:

$$\hat{p}_{ik} = \frac{\hat{\lambda}_k f_{ik}(y_i|X_{ij}, \hat{\sigma}_k^2, \hat{b}_{jk})}{\sum_{k=1}^{K} \hat{\lambda}_k f_{ik}(y_i|X_{ij}, \hat{\sigma}_k^2, \hat{b}_{jk})} \tag{A17}$$

This particular structure is equivalent to a two-stage E-M algorithm (Dempster et al., 1977) for the estimation of these parameters (see Hosmer, 1974). In the E stage, one estimates the class sizes λ_k and the membership probabilities for each case p_{ik} via expressions (A7) and (A17). In the M stage, one estimates b_{jk} and σ_k^2 via K weighted least squares regressions, using the class membership probabilities p_{ik} as weights. To show this M stage, we expand (A15) and (A16):

$$\frac{\partial \Phi}{\partial \mathbf{b}_k} = \sum_{i=1}^{I} \frac{1}{\sum \lambda_k f_{ik}(*)} \cdot \lambda_k (2\pi\sigma_k^2)^{-1/2} \times \exp\left[\frac{-(y_i - \mathbf{X}_i \mathbf{b}_k)^2}{2\sigma_k^2}\right].$$

$$\frac{2(y_i - \mathbf{X}_i\mathbf{b}_k)X_i}{2\sigma_k^2} = 0$$

$$= \sum_{i=1}^{I} \hat{p}_{ik}(y_i - \mathbf{X}_i\mathbf{b}_k)\mathbf{X}_i = 0, \tag{A18}$$

which are identical to the stationary equations derived by solving the weighted least squares problem, where $\mathbf{y} = (y_i)$ and $\mathbf{X}$ are each weighted by $\hat{p}_{ik}^{1/2}$ Thus, the entire set of $\mathbf{b}_k$ is derived by performing K separate weighted least squares analyses. Once this is done, the estimates of σ_k^2 follow:

$$\frac{\partial \Phi}{\partial \sigma_k^2} = \sum_{i=1}^{I} \frac{1}{\sum_k \lambda_k f_{ik}(*)} \left[\lambda_k \exp\left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2}\right]\right.$$
$$(-1/2(2\pi\sigma_k^2)^{-3/2}2\pi)$$
$$+ \lambda_k(2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^2}\right]$$
$$\left.\frac{1/2(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^4}\right] = 0$$

$$= \sum_{i=1}^{I} \hat{p}_{ik}\left[\frac{-1}{2\sigma_k^2} + \frac{(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{2\sigma_k^4}\right] = 0. \tag{A19}$$

Multiplying both sides of (A19) by $2\sigma_k$ and simplifying, one obtains:

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{I} \hat{p}_{ik}(y_i - \mathbf{X}_i\mathbf{b}_k)^2}{\sum_{i=1}^{I} \hat{p}_{ik}}. \tag{A20}$$

Thus, $\hat{\sigma}_k^2$ can be obtained during the K weighted least squares procedures for estimating $\mathbf{b}_k$. Note that because (A1) becomes unbounded as $\sigma_k^2 \to 0$, $\hat{\sigma}_k^2$ is set to a default small positive value (.01) if it becomes small during these iterations.

Thus, the computation of the maximum likelihood estimates is facilitated by the use of this E-M algorithm. Given starting values of the parameters, the expectation (E phase) and maximization (M phase) steps of this algorithm are alternated until convergence of a sequence of log-likelihood values is obtained. Dempster et al. (1977) prove that:

$$\Phi(\theta^{(m+1)}) \geq \Phi(\theta^{(m)}), \tag{A21}$$

where m is the iteration counter, indicating that the E-M algorithm provides monotone increasing values of the objective function. Given the constraint $\sigma_k^2 \geq .01$ one can show that Φ is bounded from above, and convergence to at least a local maximum can be established (cf. Titterington, Smith, & Makov, 1985). While several authors (e.g., Everitt & Hand, 1981; Redner & Walker, 1984) have documented the potentially slow convergence rate of E-M procedures for estimating the parameters of unconditional mixture distributions, we find that this E-M procedure typically converges in 300 or less iterations. Moreover, the iterations are processed much faster than if a gradient-based procedure had been used. Acceleration procedures discussed by Peters and Walker (1978) and Louis (1982) can also be implemented.

One of the appealing properties of maximum likelihood estimators is that, under typical regularity conditions, these estimators are asymptotically normal. Define $\mathbf{b}$ as a vector of all the ($\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K$) estimated coefficients in a maximum likelihood context and B as the corresponding vector of unknown population parameters ($\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_K$). Then, according to Theil (1971),

$$\sqrt{\boldsymbol{I}(\mathbf{b} - \mathbf{B})} \to^d N(\mathrm{O}, \lim(R(\mathbf{B})/I)^{-1}), \tag{A22}$$

where:

$$R(\mathbf{B}) = -E\left[\frac{\partial^2 \Phi}{\partial \mathbf{B} \partial \mathbf{B}'}\right], \tag{A23}$$

the information matrix. According to Judge, Griffiths, Hill, Lütkepohl, and Lee (1985), replacing $\lim(R(\mathbf{B})/I)$ by a consistent estimator does not change the asymptotic distribution of the test statistics or confidence intervals for **b.** Here, the consistent estimator used is:

$$F = \frac{1}{I}\left[\sum_{i=1}^{I}\left(\frac{\partial \Phi}{\partial \mathbf{b}^*}\right)\left(\frac{\partial \Phi}{\partial \mathbf{b}^*}\right)'\right]_{\mathbf{b}^*=\mathbf{b}}, \quad \text{(A24)}$$

and the asymptotic variances of **b** can be defined as the main diagonal elements of $\mathbf{F}^{-1}$, the asymptotic variance covariance matrix. From (A22) to (A24), it follows that an asymptotic $(1 - \alpha)$ 100% confidence interval for $\mathbf{B}_n$ is given by:

$$(b_n - Z_{\alpha/2}\sqrt{f_{nn}^{-1}}, b_n + Z_{\alpha/2}\sqrt{f_{nn}^{-1}}), \quad \text{(A25)}$$

where $Z_{\alpha/2}$ is the central value of a normal distribution with mean 0 and variance 1, and f_{nn}^{-1} is the asymptotic estimate of the variance of b_n.

B. The Latent Structure Generalized Linear Model Algorithm

For the extension to mixtures of generalized linear models, we assume that the conditional probability density function of y_{ir}, given that y_{ir} comes from segment *k,* takes the general exponential family form:

$$f_{ir|k}(y_{ir}|\Theta_{kir}, \alpha_k) = \exp\{(y_{ir}\Theta_{kir} - b(\Theta_{kir}))/\gamma(\alpha_k) + c(y_{ir}, \alpha_k)\} \quad \text{(A26)}$$

for specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ that determine the specific distribution as a member of the exponential family. Conditional upon segment *k,* the y_{ir} are independently distributed according to this distribution, with canonical parameters Θ_{kir} and means μ_{kir}. The parameter α_k is called the dispersion parameter—relevant for some members of the exponential family only such as the normal and the gamma distributions, and it is assumed to be constant over observations in segment *k,* while $\gamma(\alpha_k) > 0$. If α_k is known, then the distribution is a member of the exponential family with canonical parameter Θ_{kir}. The distribution may or may not be a member of the exponential family if α_k is unknown (cf. McCullagh & Nelder, 1989; e.g., the negative binomial may not be a member of the exponential family for such unknown dispersion parameters). We specify a linear predictor η_{kir} and a link function $g(\cdot)$ that captures the structural part of consumer response that one wishes to describe, such that in segment *k*:

$$\eta_{kir} = g(\mu_{kir}), \quad \text{(A27)}$$

where the linear predictor is produced by *J* covariates $X_1, \ldots, X_J$ $(X_J = (X_{irj}))$, $j = 1, \ldots, J$, and the parameter vectors $\beta_k = (\beta_{ij})$ in segment *k:*

$$\eta_{kir} = \sum_{j=1}^{J} X_{irj}\beta_{kj}. \quad \text{(A28)}$$

Thus, conditional upon segment *k,* a generalized linear model is formulated with the specification of the distribution of the random variable, y_{ir}; a linear predictor η_{kir}; and a function $g(\cdot)$, which links the random and systematic components (so-called canonical links occur when $\Theta_{kir} = \eta_{kir}$ for the normal, Poisson, binomial, gamma, and inverse Gaussian distributions; see McCullagh & Nelder, 1989). The unconditional probability density function of an observation vector y_i can therefore be expressed in the finite mixture form (McLachlan & Basford, 1988):

$$f_i(y_i|\Phi) = \sum_{k=1}^{K} \lambda_k \prod_{r=1}^{R} f_{ir|k}(y_{ir}|\beta_k, \alpha_k), \quad \text{(A29)}$$

where:

$$\Phi' = ((\lambda', \beta', \pi'); \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'; \boldsymbol{\beta} = (\beta_1', \ldots, \beta_K')'; \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)').$$

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

$$L(\Phi; y) = \prod_{i=1}^{n} f_i(y_i|\Phi). \quad \text{(A30)}$$

An estimate of Φ can be obtained by maximizing the likelihood Equation (A30) with respect to Φ, subject to the restrictions in (A3) to (A5) using an E-M algorithm (Dempster et al., 1977), or through direct numerical maximization using algorithms such as Newton-Raphson. Once an estimate of Φ has been obtained, estimates of the posterior probability, p_{ki}, that observation i comes from latent segment k can be calculated for each observation vector y_i by means of Bayes theorem, where, as before, this posterior probability is given by:

$$p_{ki}(y_i, \Phi) = \frac{\lambda_k \prod_{r=1}^{R} f_{ir|k}(y_{ir}|\beta_k, \alpha_k)}{\sum_{k=1}^{K} \lambda_k \prod_{r=1}^{R} f_{ir|k}(y_{ir}|\beta_k, \alpha_k)}. \quad \text{(A31)}$$

The entropy of the posterior classification is provided by:

$$E_k = 1 - \frac{\sum_{i=1}^{I} \sum_{k=1}^{K} p_{ik} \ln p_{ik}}{I \ln K}. \quad \text{(A32)}$$

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control, AC-19,* 716–723.

Allenby, G. M., Arora, N., & Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research, 35,* 384–389.

Allenby, G. M., & Rossi, P. (1999). Marketing models of heterogeneity. *Journal of Econometrics, 89,* 57–78.

Andrews, R., Ainslie, A., & Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research, 39,* 479–487.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Bozdogan, H. (1994). Mixture model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan (Ed.), *Multivariate statistical modelling* (Vol. 2, pp. 69–113). Dordrecht, The Netherlands: Kluwer Academic.

Celeux, G., & Soromenho, G. (1996). An entropy based criterion for assessing the number of cluster in a mixture model. *Journal of Classification, 13,* 195–212.

Dempster, A. P., Laird, N. M., & Rubin, R. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society, B39,* 1–38.

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification, 5,* 249–282.

DeSarbo, W. S., & Hausman, R. A. (2006). *Restricted components analysis* (Working paper). University Park: Pennsylvania State University.

DeSarbo, W. S., & Ramaswamy, V. (1995). Marketing segmentation with choice based conjoint analysis. *Marketing Letters, 6,* 137–148.

DeSarbo, W. S., Wedel, M., Vriens, K., & Ramaswamy, V. (1992). Latent class metric conjoint analysis. *Marketing Letters, 3,* 273–288.

DeSoete, G., & DeSarbo, W. S. (1991). A latent class probit model for analyzing pick any/n data. *Journal of Classification, 8,* 45–63.

Dias, J. (2004). *Finite mixture models: Review, applications and computer intensive methods.* Thesis, SOM Research School, University of Groningen, The Netherlands.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distribution.* New York: Chapman & Hall.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215–231.

Hosmer, D. W. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics, 3,* 995–1006.

Johnston, J. (1984). *Econometric methods* (3rd ed.). New York: McGraw-Hill.

Judge, G. C., Griffiths, W. E., Hill, R. C., Lütkepohl, M., & Lee, T. C. (1985). *Theory and practice of econometrics.* New York: John Wiley.

Kamakura, W. A., & Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research, 26,* 339–390.

Kamakura, W. A., & Wedel, M. (1997). Statistical datafusion for cross-tabulation. *Journal of Marketing Research, 34,* 485–498.

Kamakura, W. A., & Wedel, M. (2004). An empirical Bayes procedure for improving individual-level estimates and predictions from finite mixture logit models. *Journal of Business and Economic Statistics, 22,* 121–125.

Louis, T. A. (1982). Finding the observed information matrix when using the E-M algorithm. *Journal of the Royal Statistical Society, B44,* 226–233.

Maddala, G. S. (1976). *Econometrics.* New York: McGraw-Hill.

Malhotra, N. K. (1996). *Marketing research: An applied orientation* (2nd ed.). London: Prentice Hall.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* New York: Chapman & Hall.

McLachlan, G. J. (2000). *Finite mixture models.* New York: John Wiley.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering.* New York: Marcel Dekker.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A135,* 370–384.

Peters, B. C., & Walker, H. F. (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal of Applied Mathematics, 35,* 362–378.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood, and the E-M algorithm. *SIAM Review, 2,* 195–239.

Theil, H. (1971). *Principles of econometrics.* New York: John Wiley.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* New York: John Wiley.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent Gold 4.0.* Belmont, MA: Statistical Innovations.

Wedel, M. (2001). GLIMMIX: Software for estimating mixtures and mixtures of generalized linear models. *Journal of Classification, 18*(1), 129–137.

Wedel, M., & DeSarbo, W. S. (1993). A latent class binomial logit methodology for the analysis of paired comparison choice data: An application reinvestigating the determinants of perceived risk. *Decision Sciences, 24,* 1157–1170.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification, 12,* 1–35.

Wedel, M., & DeSarbo, W. S. (1996). An exponential family mixture MDS methodology for simultaneous segmentation and product positioning. *Journal of Business and Economic Statistics, 14*(4), 447–459.

Wedel, M., DeSarbo, W. S., Bult, J. R., & Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics, 8,* 397–411.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation, methodological and conceptual foundations* (2nd ed.). Dordrecht, The Netherlands: Kluwer.

Wedel, M., Kamakura, W. A., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., et al. (1999). Discrete and continuous representation of heterogeneity. *Marketing Letters, 10*(3), 217–230.

Additional Readings

Cooper, P. W. (1964). Non-supervised adaptive signal detection and pattern recognition. *Information and Control, 7,* 416–444.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika, 56,* 463–474.

Dynkin, E. B. (1961). *Selected translations in mathematical statistics and probability.* Providence, RI: American Mathematical Society.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification, 17,* 273–296.

Teicher, H. (1961). Identifiability of finite mixtures. *Annals of Mathematical Statistics, 34,* 1265–1269.

Teicher, H. (1961). Identifiability of mixtures. *Annals of Mathematical Statistics, 32,* 244–248.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5,* 329–350.

Yakowitz, S. J. (1970). Unsupervised learning and the identifiability of finite mixtures. *IEEE Transactions on Information Theory and Control, IT-16,* 330–338.

Yakowitz, S. J., & Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics, 39,* 209–214.