CHAPTER 8

The Semiparametric MLE

Although there are certainly examples where interest in the mixture NPMLE method has focused on the estimated latent distribution itself or functionals of it, the largest class of investigations and applications has occurred in the arena of semiparametric estimation, in which the latent distribution is included in the model to allow extra heterogeneity, but the focus is on a set of auxiliary parameters, generally of the regression type.

The earliest extensive applied investigation of semiparametric mixture maximum likelihood was by Heckman and Singer (1984), who were investigating the effects of latent distribution misspecification and were comparing, therefore, the nonparametric and parametric approaches to modeling the latent distributions. Among the many further investigations of this type we might count Follman and Lambert (1989), Brännäs and Rosenqvist (1994), Butler and Louis (1992), and Davies (1993). See Lindsay and Lesperance (1995) for a survey of the results in this area.

Since there exists a substantial literature on the implementation of this methodology, we will focus herein instead on special simplifying structures that can exist when we have a likelihood with both latent and auxiliary parameters. We will consider three different semiparametric models. One is an exponential family random effects model, of which the Rasch model is an illustration. The second is a measurement error problem, with the additional complication of case-control sampling. The final problem is an outlier distribution model for contingency tables that leads to a new method for assessing the fit of a parametric model.

We start with a general optimization result that is important because it leads to a simplification and clarification of what would otherwise be numerically more difficult problems. It will be used in the examples.

8.1. An equivalence theorem. In this section, we will use the notation β and G to designate two arbitrary "parameters" upon which our model depends, but in our applications, β will be real-valued parameters of interest and G will be the latent distribution. Our interest is in whether the β parameters

we would obtain from maximizing a conditional likelihood will be the same as we would obtain from maximizing a simpler joint likelihood.

The setting: We wish to maximize a target likelihood L_c with ratio structure:

$$L_c(\beta, G) = \frac{L_J(\beta, G)}{L_m(\beta, G)}.$$

We let $(\hat{\beta}_c, \hat{G}_c)$ be maximizers of L_c and we wish to compare them with $(\hat{\beta}_J, \hat{G}_J)$, which will come from maximizing the simpler likelihood L_J. In the settings in which this is applied, the likelihood we wish to maximize, L_c, is a *conditional* likelihood. Then the numerator likelihood L_J is the *joint* likelihood, which is of the desired product form, while the denominator likelihood is the *marginal* likelihood L_m of the conditioning statistics.

Our first key assumption is that the denominator likelihood has multinomial structure; that is, there are constants $n_1, \ldots, n_K$ and nonnegative functions $p_j(\beta, G)$ satisfying

$$\sum p_j(\beta, G) = 1$$

such that the likelihood can be written

$$L_m(\beta, G) = \prod p_j(\beta, G)^{n_j}.$$

Let $n = \sum n_j$. This structure will clearly arise if we are conditioning on a set of discrete statistics.

The next assumption specifies a form of *nonidentifiability* of the parameter G in the conditional likelihood. That is, we assume that for every pair (β, G) in the parameter space there exists G^* such that

$$L_c(\beta, G) = L_c(\beta, G^*) \tag{8.1}$$

and

$$p_j(\beta, G^*) = \frac{n_j}{n} \qquad \forall j. \tag{8.2}$$

That is, no matter which β we start with, we can find G^* that shifts the multinomial probabilities under the model to equal the observed sample proportions without altering the target likelihood. We note that this is where the nonparametric nature of G is important, so that there are "sufficient degrees of freedom" to solve these equations.

With these assumptions, we are done, because as far as the parameter of interest β is concerned, we can equally well maximize either L_J or L_c; the solutions are the same.

PROPOSITION 29. *Under the preceding assumptions, if* $(\hat{\beta}_c, \hat{G}_c)$ *maximizes the target likelihood* L_c, *then* $(\hat{\beta}_c, \hat{G}_c^*)$ *maximizes* L_J. *Conversely, if* $(\hat{\beta}_J, \hat{G}_J)$ *maximizes* L_J, *then* $(\hat{\beta}_J, \hat{G}_J)$ *satisfies* (8.2) *and is in an equivalence class that maximizes* L_c.

PROOF. We write the likelihoods in the product form

$$L_J(\beta, G) = L_c(\beta, G)L_m(\beta, G).$$

We note that by the assumption (8.1), if $(\hat{\beta}_c, \hat{G}_c)$ maximizes L_c, then so must $(\hat{\beta}_c, \hat{G}_c^*)$. Further, $(\hat{\beta}_c, \hat{G}_c^*)$ also maximizes L_m because of its multinomial form and (8.2). The sample proportions give, in fact, the maximum possible marginal likelihood over all possible models. As to the converse, if (8.2) was not satisfied at the maximum, we could increase the likelihood component L_m without changing that of L_c by choosing G^* to make it so, a contradiction. If $(\hat{\beta}_J, \hat{G}_J)$ did not maximize the other component L_c, then we could increase L_c without decreasing L_m by the same device. □

8.2. Exponential response models. We now consider a simple class of exponential response models. A sample is taken from a population, with the measurement being a vector $x = (x_1, \ldots, x_r)'$, with x_i, being the "response" to the ith out of r "items." There will be item parameters $\theta = (\theta_1, \ldots, \theta_r)$ that determine the distribution of responses and a parameter ϕ_i that reflects the latent propensities of the sampled unit. The density for the ith sampling unit, conditional upon latent variable $\Phi = \phi_i$, will have the exponential form

$$f(x_i; \theta, \phi_i) = \exp[\theta' x + \phi_i s(x) - k(\theta, \phi_i)],$$

where $s(x)$ is the sufficient statistic for the variable ϕ_i.

For the ensuing discussion, it will be essential that $s(x)$ has a finite discrete distribution, say with sample space $\{0, \ldots, K\}$. [A subject for investigation would be its approximate validity when $s(x)$ is continuously distributed.]

Two other features of the exponential form that are important in the theory are:

1. The statistic $s(x)$ is complete and sufficient for ϕ when θ is fixed;
2. The marginal distribution of $s(x)$ is an exponential family when θ is fixed.

8.2.1. *Example: Rasch model.* A model of the type we are interested in that has received a great deal of attention is the *Rasch model*, in which the responses X_{ij} are binary variables. We start with a logistic model for the ith subject's response to the jth item, conditional on the subject's latent variable $\Phi = \phi_i$:

$$\Pr[X_{ij} = 1 \mid \Phi_i = \phi_i] = \frac{\exp(\theta_j + \phi_i)}{1 + \exp(\theta_j + \phi_i)}.$$

We further assume that conditional on the latent variable, all the responses of an individual are independent, so that conditional on the latent variable the density has the exponential response form

$$\Pr[X_i = x_i \mid \Phi = \phi_i] = \exp\left(\sum_j \theta_j x_{ij} + \phi_i x_{i\cdot} - \kappa(\theta, \phi_i)\right),$$

where the sufficient statistic for ϕ_i is the response total $x_{i\cdot}$ of the ith subject. The model is overparameterized, so to make things identifiable a constraint must be used. We will here specify that the last item parameter, θ_K, is zero.

8.2.2. *Type I conditional models.* A reason for particular interest in the exponential response model is that if the focal parameters are the θ's, then there exists a natural competitor to using semiparametric mixture likelihood methods. Because of the structure of the model, one can form a *conditional likelihood* for each subject that depends strictly on the parameters of interest, namely,

$$L_{i,\text{cond}}(\theta) = \Pr[X_i = x_i \mid s(X_i) = s_i].$$

We can now estimate the focal parameters without regard to the structure of the latent variables, whether they are treated as nuisance parameters or as a sample from an unknown distribution. The resulting conditional maximum likelihood estimators (or maximum conditional likelihood estimators) are very generally consistent and highly efficient relative to maximum likelihood [e.g., Liang (1984)]. Important early work on this method was done by Anderson (1973).

8.2.3. *The two-item example.* A simple example that illustrates our situation is the *paired Bernoulli* problem, which can also be described as the Rasch model with two items. We have from each unit (subject or paired subjects) two binary variates (X_{i1}, X_{i2}). We might think of the first as corresponding to a response under treatment and the second as the response under control. We apply the Rasch model, where we have one item parameter $\theta = \theta_1$ (since $\theta_2 = 0$), which corresponds to the log odds ratio for the success parameters for the pair, and so represents a common treatment effect. There are only four possible responses for each pair, namely, $(0,0)$, $(1,0)$, $(1,0)$ and $(1,1)$. We let $N(a,b)$ denote the number of pairs of responses with pattern (a,b).We can therefore easily write the conditional densities for the data given the sufficient statistic $X_{i1} + X_{i2} = S_i$:

$$\begin{aligned}
\Pr[X = (0,0) \mid S = 0] &= 1, \\
\Pr[X = (1,0) \mid S = 1] &= \frac{\exp(\theta)}{1+\exp(\theta)}, \\
\Pr[X = (0,1) \mid S = 1] &= \frac{1}{1+\exp(\theta)}, \\
\Pr[X = (1,1) \mid S = 2] &= 1.
\end{aligned}$$

The conditional likelihood for the problem is therefore

$$L_{\text{cond}}(\theta) = \left[\frac{\exp(\theta)}{1+\exp(\theta)}\right]^{N(1,0)} \left[\frac{1}{1+\exp(\theta)}\right]^{N(0,1)}$$

This is easily maximized for θ and provides a method of consistent estimation even though the fixed effects MLE is inconsistent. However, it offers a curiosity

in that it only uses the information from the *discordant pairs*; that is, those in which the binary variates were not equal. The estimator does not use $N(1,1)$ or $N(0,0)$, the numbers of concordant pairs.

8.2.4. *Efficiency theorem.* My early interest in the mixture problem arose around the statistical question of the efficiency of the conditional approach to eliminating nuisance parameters. Addressing the question of efficiency in the case of the number of nuisance parameters going to infinity is quite difficult. My approach was to consider the efficiency of the conditional approach within the semiparametric mixture model, where the nuisance parameters ϕ_i were assumed to come from a latent distribution Q. The problem is still difficult, but there is sufficiently more structure so that one can, in a natural way, extend the efficiency ideas from the parametric model to the nonparametric, and answer this question.

The answer [Lindsay (1983c)] is that the conditional method is generally fully efficient within the mixture setting. This last paper was written when the appropriate semiparametric theory was nascent; for an up-to-date description of the necessary optimality theory, we refer the reader to the book by Bickel, Klaassen, Ritov and Wellner (1993).

We note that there are some technical issues concerning the boundaries of the parameter space. If the true latent distribution is, for example, degenerate, then the conditional MLE is still best asymptotically normal, but there may be estimators, such as the mixture MLE, that will gain superior performance at the sacrifice of asymptotic normality.

8.2.5. *Equivalence theorem for mixture MLE.* If the conditional MLE is fully efficient in the mixture setting, then one must ask "what about the semiparametric mixture MLE?"

Since logic suggests that the MLE for the specified model should be efficient therein, this suggests that perhaps the mixture MLE for the parameters of interest will be asymptotically equivalent to the conditional MLEs.

We are now in a position to address this question using our equivalence theorem. We return to that setting with $\theta = \beta$, $Q = G$, and L_J, L_c and L_m are the mixture likelihood, the conditional likelihood and the marginal likelihood of S, respectively. We check our conditions for equivalence of estimation. Recall that for each (β, G) we need to find G^* with certain properties. The first,

$$L_c(\beta, G) = L_c(\beta, G^*),$$

is trivial here because the conditional likelihood does not depend on G. Thus all we need to do is to find G^* satisfying

$$p_j(\beta, G^*) = \frac{n_j}{n} \qquad \forall j, \tag{8.3}$$

where $p_j(\beta, G) = \Pr[S = j]$, the marginal distribution of the sufficient statistic S.

We now summarize the extensive treatment of this problem in Lindsay, Clogg and Grego (1991).

We have already developed the tools in Chapter 2 to study questions of existence such as posed by (8.3). The subject at hand is the mixture density under fixed β of the statistic S, and so we can plot the unicomponent models as a curve $\{\mathbf{p}(\beta, \phi)\colon \phi \in \Omega\}$ in the simplex, form the convex hull of mixture densities and then ask the question: Does the vector of observed proportions $\mathbf{d}$, with coordinates n_j/n, lie inside this mixture set? If it does, then there exists a solution G^* to (8.3); otherwise, there does not.

Thus, in fact, the answer depends on the observed empirical distribution of the variable S. If it is equal to a mixture density vector, then the conditional and mixture maximum likelihood estimators of the item parameters are *identical*, and otherwise not necessarily. It is important from a conceptual point of view, however, that when the model is correct, then the estimators are identical, with probability 1, for a large enough sample size.

For completeness we add that this asymptotic result relies on the distribution G having enough support points that the vector of true probabilities $\Pr[S=j]$ does not lie on the boundary of the mixture set. If this occurs, the semiparametric MLE of θ is not asymptotically equivalent to the conditional MLE, nor even normally distributed. This question can be addressed quite explicitly in the two-item model. The only boundary situation here is when the true latent distribution has a single point of support. In this case the asymptotic distribution of the maximum likelihood estimator of θ is that of a preliminary test estimator in which one chooses between the conditional estimator and the unicomponent estimator based on a preliminary test of heterogeneity.

In all the data sets examined by Lindsay, Clogg and Grego (1991), the conditional and mixture approaches resulted in the same estimator. Among the other issues addressed therein were algorithms and the identifiability of parameters in the mixture model. The authors note that a key advantage of the mixture approach is that it leads quite naturally to empirical Bayes assessment of a subject's latent parameter value.

8.3. Errors-in-variables and case-control studies. As a second example of the use of the equivalency theorem, we consider the problem of estimating logistic regression parameters in a case-control sampling framework. The presentation here represents the pertinent part of a study by Roeder, Carroll and Lindsay (1993). We will first replicate a famous result of Prentice and Pyke by putting it into the framework of the equivalency theorem, and then we extend it into the case of a measurement error model.

8.3.1. *The joint sampling model.* We start with a sample of data (D_i, X_i) in which D_i is a binary response variable, such as diseased (1) and nondiseased (0), and X_i is a vector of potential explanatory variables. We model the sample with a prospective logistic regression that specifies a parametric conditional distribution of Y given X:

$$\Pr[Y = 1 \mid X = x] = \frac{\exp(\alpha + x'\beta)}{1 + \exp(\alpha + x'\beta)}.$$

The second part of the model, the marginal distribution of X, denoted G, is left completely unspecified. We will assume it is discrete, with density $g(x)$. If we do this, the joint likelihood splits apart into two terms, one depending on α and β alone and the other on G alone:

$$L_J(\alpha, \beta, G) = \prod_{i=1}^{n} \Pr[D = d_i \mid X = x_i; \alpha, \beta] \prod_{i=1}^{n} g(x_i).$$

Thus joint maximum likelihood is easy. The nonparametric MLE of G is the empirical CDF of the x data, and α and β are estimated from the prospective logistic regression model.

8.3.2. *The retrospective model.* We next suppose that the sampling was not carried out randomly, but rather we took a sample from each of the two populations: the *cases* ($D = 1$) and the *controls* ($D = 0$). Let n_1 and n_0 be the size of the samples from the two populations and let $n = n_1 + n_0$. Then, in truth we are sampling from the conditional densities $\Pr[X = x \mid D = d]$ under two different values of d, and the likelihood we should maximize is the conditional likelihood:

$$L_c(\alpha, \beta, G) = \prod\prod \Pr[X = x_i \mid D = d_i].$$

The corresponding marginal likelihood for the data is

$$L_m(\alpha, \beta, G) = \Pr[D = 1]^{n_1} \Pr[D = 0]^{n_0}.$$

8.3.3. *Prentice and Pyke's equivalency.* Prentice and Pyke (1979) established that one could obtain the parameters $\hat{\beta}_c$ that maximized the retrospective likelihood L_c by maximizing the joint likelihood L_J, which amounts to maximizing the prospective logistic regression over β and α. Moreover, the solutions obtained this way satisfied

$$\Pr[D = 1; \hat{\alpha}, \hat{\beta}, \hat{G}] = \frac{n_1}{n}.$$

This last result is a clue that this result is an application of the equivalency theorem in Section 8.2.

To apply the theorem, we must establish that for each set of parameters (α, β, G), there exists another set (α^*, β, G^*) such that the retrospective distributions are the same,

$$\Pr[X = x \mid D = d; \alpha, \beta, G] = \Pr[X = x \mid D = d; \alpha^*, \beta, G^*],$$

but the marginal disease distribution equals the observed proportions exactly:

$$\Pr[D = 1; \alpha^*, \beta, G^*] = \frac{n_1}{n}. \tag{8.4}$$

Let $\Pr[D = 1; \alpha, \beta, G] = p$. For the proof of this claim the reader should check that

$$\alpha^* = \alpha + \log[n_1(1 - p)/pn_0]$$

and

$$dG^*(x) \propto \frac{[1+\exp(\alpha^*+\beta x)]}{[1+\exp(\alpha+\beta x)]} dG(x)$$

do the trick. One key point here is that unlike our earlier application, we can satisfy (8.4) for *every* case-control sampling fraction n_1/n that is not zero. The Prentice–Pyke result follows directly.

8.3.4. *The measurement error extension.* The extension we desire to make incorporates one further level of difficulty. Suppose that the desired regressor variable is measured with error. That is, instead of measuring X directly, we observe the surrogate variable $S_i = X_i + \text{error}$. More precisely, we suppose knowledge of a density

$$f(s|x) = \Pr[S = s \mid X = x],$$

possibly with unknown parameters, for the distribution of S_i given $X_i = x$. In our terms, the variable X_i is the latent variable, and what we desire is the latent logistic regression of y on x. The mechanism generating the errors is assumed to be independent of the regression of interest, so that

$$\Pr[D = d \mid S = s,\ X = x] = \Pr[D = d \mid X = x].$$

The joint distribution of the observables therefore has the mixture form

$$\Pr[D = d,\ S = s] = \int \Pr[D = d \mid X = x] \Pr[S = s \mid X = x]\, dG(x). \tag{8.5}$$

8.3.5. *The extended equivalency result.* The question now arises: Can we simply maximize the *joint* likelihood

$$L_J(\alpha, \beta, G) = \prod\prod \Pr[D = d_i, S = s_i],$$

which from (8.5) has a standard mixture form, when we want to find the regression parameters β that maximize the *retrospective* likelihood:

$$L_c(\alpha, \beta, G) = \prod\prod \Pr[S = s_i \mid D = d_i]?$$

To maximize the latter, we would have to maximize over a nonstandard likelihood containing ratios of mixture probabilities.

The answer follows directly from the identifiability result we used in the no-measurement-error example. If (α, β, G) are the original parameters, then let (α^*, β, G^*) be as in the previous result. These parameters still give a distribution for D that fits the proportions of cases and controls, because the marginal distribution of D does not depend on the measurement of S. Moreover, if these parameters give identical distributions for $X|D$, then they must give identical distributions for $S|D$.

Furthermore, when we maximize L_J we achieve a perfect fit of the marginal proportions of cases and controls, as in the Prentice–Pyke result. We note that it is critical to this result that G be modeled nonparametrically, so that the equivalency equations (8.1) and (8.2) have a solution.

Roeder, Carroll and Lindsay (1993) have an extended version of this problem in which both X and S are measured on a subsample, as well as considerably more on the practical problem and the efficiency of the maximum likelihood procedure.

8.4. A mixture index of fit. As a last semiparametric example, we will introduce a very different kind of model that arises from a mixture point of view. It provides an interesting extension of some of the techniques we have discussed. The original source for this analysis is Rudas, Clogg and Lindsay (1994).

8.4.1. *The problem.* We suppose that there exists a simple *baseline* probability model $f(x;\beta)$ for the data X that we wish to use to make inferences for a population. It might, for example, be the model of independence for a multiway contingency table. However, we know that it cannot be a perfectly correct description of the population. We would like a simple interpretable measure of how close the baseline model is to being correct—one that we could easily use to evaluate the predictive capabilities of the baseline model. As an illustration, consider the data in Table 8.1, taken from Diaconis and Efron (1985).

We have a simple cross-classification of variables. The chi-squared goodness-of-fit for the independence model gives us a test statistic of 138.29 on 9 degrees of freedom, so we certainly reject that model. However, from this information we cannot tell if it is a reasonably close description of the data anyway, and our rejection arose because we have a relatively large sample size.

8.4.2. *The concept.* We let $g(x)$ be the true distribution of the data. We suppose that it can be written as a mixture of a baseline model density and a second completely arbitrary density $q(x)$:

$$g(x) = (1-\pi)f(x;\beta) + \pi q(x). \tag{8.6}$$

Note that the density $q(x)$ represents a *lack-of-fit* or *outlier* distribution. [Instead of letting $q(x)$ be arbitrary, we could here specify that $q(x)$ is an element of some large class of densities that includes all the model densities and that is also sure to include the true density.]

TABLE 8.1
Cross-classification of eye color and hair color

	Hair color			
Eye color	**Black**	**Brunette**	**Red**	**Blonde**
Brown	68	119	26	7
Blue	20	84	17	94
Hazel	15	54	14	10
Green	5	29	14	16

The variables π and $q(x)$ in the mixture representation (8.6) are not unique, because once a representation is given using π and q, one can construct others by the action of "moving some of the baseline model into the lack-of-fit density" as follows:

$$(8.7)\quad g(x) = (1 - \pi - \varepsilon)f(x;\beta) + (\pi + \varepsilon)\left[\frac{\pi}{\pi + \varepsilon}q(x) + \frac{\varepsilon}{\pi + \varepsilon}f(x;\beta)\right],$$

provided that ε is sufficiently small that the new mixture weight $\pi + \varepsilon$ is in [0, 1]. However, one can turn the parameter π into something well defined and interpretable by letting π^*, the *lack-of-fit index* for the density g, be the *smallest* π one could use in such a representation:

$$\pi^*(\mathbf{g}) = \inf\{\pi: g(x) = (1 - \pi)f(x;\beta) + \pi q(x)\}.$$

We will assume that the class of baseline models is closed, so that there exists a representation of the form

$$g(x) = (1 - \pi^*)f(x;\beta^*) + \pi^* q^*(x).$$

The parameter π^* has a simple interpretation: It is the smallest fraction of the population that must be removed before the baseline model would fit perfectly. Additionally, we can interpret $(1 - \pi^*)$ as the maximal fraction of the population that can be described exactly by the baseline model, and so the population fraction to which the model-based inference applies.

8.4.3. *Application to the multinomial.* This modeling scheme leads directly to a method of estimation if we are in a contingency table setting. We continue as before, but substitute the variable t for x to remind us that it is an index for the cells of a multinomial contingency table. We also recall that in this case, we have the nonparametric multinomial MLE, which is just $d(t)$, the observed cell proportions. Hence we can use $\pi^*(\mathbf{d})$ to estimate $\pi^*(\mathbf{g})$.

To better understand how this works, we first consider all multinomial densities $p(t)$ such that their lack-of-fit $\pi^*(\mathbf{p})$ is less than some fixed value π_0:

$$\mathscr{H}_{\pi_0}(\text{model}) = \{p(t): \pi^*(\mathbf{p}) \le \pi_0\}.$$

We can think of $\mathscr{H}_{\pi_0}(\text{model})$ as representing the set of all multinomial densities whose lack-of-fit is acceptably small. We note that because of the shuffling property (8.7) we can also describe $\mathscr{H}_{\pi_0}(\text{model})$ as being the set of multinomial probability vectors for which there exists *some* π_0 representation:

$$\mathscr{H}_{\pi_0}(\text{model}) = \{p(t) = (1 - \pi_0)f(t;\beta) + \pi_0 q(t), \text{for some } \beta \text{ and } q(\cdot)\}.$$

In the simplest case, where the baseline model is just a single known distribution $f(t)$, this is just a simplex of the form

$$\mathscr{H}_{\pi_0}(\mathbf{f}) = \text{conv}\{(1 - \pi_0)\mathbf{f} + \pi_0 \mathbf{e}_t: t = 1, \ldots, T\}.$$

We invite the reader to verify this statement. [The key here is that an arbitrary multinomial distribution $q(t)$ can be written as $\mathbf{q} = \sum q(t)\mathbf{e}_t$.]

The next step is to note that we can find $\mathscr{H}_{\pi_0}(\text{model})$ simply by taking the union of the simplices generated by individual baseline model elements:

$$\mathscr{H}_{\pi_0}(\text{model}) = \bigcup_{\beta} \mathscr{H}_{\pi_0}(\mathbf{f}_\beta).$$

It follows that $\mathscr{H}_{\pi_0}(\text{model})$ is itself not necessarily a convex set. However, it is true that as we increase π, say from π_0 to π_1, we increase the set of acceptable models:

$$\mathscr{H}_{\pi_0}(\text{model}) \subset \mathscr{H}_{\pi_1}(\text{model}). \tag{8.8}$$

We also note that $\mathscr{H}_0(\text{model})$ is just the family of baseline models and that $\mathscr{H}_1(\text{model})$ is the full set of multinomial models.

8.4.4. *Maximum likelihood estimation.* This said, we can find the maximum likelihood estimator of the true multinomial density $g(t)$ under the constraint that $\mathscr{H}_{\pi_0}(\text{model})$ is our class of acceptable models. We can think of this as estimating the baseline model $f(t;\beta)$, but allowing for a contamination fraction of up to π_0 from some other distribution. Maximum likelihood on this set is fairly easy, since we can hold π_0 *fixed* and maximize over β and $q(\cdot)$ from the class of all models of the form $(1-\pi_0)f(t;\beta) + \pi_0 q(t)$.

Rudas, Clogg and Lindsay (1994) describe an EM algorithm approach that works, but is quite slow. Xi and Lindsay (1995) give a more efficient sequential quadratic programming method.

If we let the *profile* log likelihood $L^*(\pi)$ be the value of the log likelihood after we maximize over $\mathscr{H}_{\pi_0}$, then it is clear from the nesting of the models (8.8) that $L^*(\pi)$ is increasing in π. Let $\hat{L}$ be the likelihood of $\mathbf{d}$, the nonparametric multinomial MLE. We can visualize our acceptable model sets $\mathscr{H}_\pi$ growing in π until just as $\hat{\pi}^* = \pi^*(\mathbf{d})$, we find that $\mathbf{d}$ is in the boundary of $\mathscr{H}_{\hat{\pi}^*}$. From that value of π on, $\mathbf{d}$ is the maximum likelihood estimator from the model set $\mathscr{H}_\pi$, and so $L^*(\pi) = \hat{L}$. We can construct a natural measure of the adequacy of a value of π_0 by the likelihood ratio statistic

$$\text{lrs}(\pi_0) = 2\left[\ln \hat{L} - \ln L^*(\pi_0)\right].$$

This statistic becomes zero when π_0 is sufficiently large.

An analysis of the data in the Table 8.1 can be carried out in this manner, and we arrive at the information in Table 8.2. We can see that about 30% of the data would have to be discarded for the independence model to fit the data exactly. The column labeled X^2 gives the corresponding Pearson chi-squared values. The statistic $\hat{\pi}^*_L$ in the table is our next topic.

8.4.5. *Inference on the lack-of-fit index.* One inferential problem that we face is that the parameter estimator $\hat{\pi}^*$ is biased upward. This is most evident when the baseline model is correct, because the true value of π^* is zero, but the sampled values are necessarily positive. Fortunately, when the true value of π_0 is *not* zero, we can use asymptotics to construct a lower confidence limit for it, and so rescue some assurance that a large value of $\hat{\pi}^*$ is truly atypical.

TABLE 8.2
Fit statistics for the semiparametric mixture model applied to the data in Table 8.1

	X^2	lrs(π)
0.00	138.29	146.44
0.10	47.35	48.67
0.15	23.74	24.36
0.20	8.55	8.75
0.236 ($= \hat{\pi}^*_L$)	2.57	2.66
0.25	1.38	1.44
0.26	0.83	0.87
0.27	0.42	0.43
0.28	0.16	0.16
0.29	0.02	0.02
0.298 ($= \hat{\pi}^*$)	0.00	0.00
$\pi \geq 0.298$	0.00	0.00

We therefore consider using the likelihood ratio test statistic lrs(π_0) to test the hypothesis $\pi^*(\mathbf{g}) \leq \pi_0$ against $\pi^*(\mathbf{g}) > \pi_0$.

The asymptotics of the situation can be handled by the techniques described in Chapter 4. When π_0 is correct, the null density vector $\mathbf{g}$ sits on the boundary of the set $\mathscr{H}_{\pi_0}$. This set, from its earlier description, is full dimensional within the simplex.

To aid the geometric imagination, we first consider the simple case where the baseline model is a single density $\mathbf{f}$ and $\mathscr{H}_{\pi_0}$(model) is just the simplex $\mathscr{H}_{\pi_0}(\mathbf{f})$ containing $\mathbf{f}$. After transforming to the dagger space, the simplex is still a simplex, and as long as $\mathbf{g}$ is on one of the flat full-dimensional faces of this simplex, then the model cone generated by the models in $\mathscr{H}_{\pi_0}(\mathbf{f})$ are an entire half space. All directions that one can move while staying in the simplex face correspond to "nuisance score" directions, and any vector which is E_0 orthogonal to the face corresponds to the corrected score for π. Since if we go in one direction we go into the model and the other way we leave it, we are in a setting where the asymptotic distribution of likelihood ratio test statistic for $\pi^*(\mathbf{g}) \leq \pi_0$ versus $>$ has the limiting distribution $0.5\chi^2_0 + 0.5\chi^2_1$.

Of course, when the true distribution is not on a face of the simplex $\mathscr{H}_{\pi_0}(\mathbf{f})$, but on an edge or corner or other lower dimensional surface, the analysis becomes considerably more difficult. The geometry shows that the test statistic will have a chi-bar-squared distribution that is stochastically larger than $0.5\chi^2_0 + 0.5\chi^2_1$. Unfortunately, since this makes us more likely to reject if the true distribution is such a boundary point and therefore if one uses the distribution $0.5\chi^2_0 + 0.5\chi^2_1$, then the test procedure is anticonservative. However, we do note that the region of parameter values where the test based on $0.5\chi^2_0 + 0.5\chi^2_1$ is anticonservative has Lebesgue measure zero and might be presumed to have prior probability of zero.

These arguments extend beyond the single distribution model $\mathscr{H}_{\pi_0}(\mathbf{f})$ to the general case $\mathscr{H}_{\pi_0}$(model) by using the fact that for a general baseline

model, the π-model set is a union of simplices and so has an open interior in the neighborhood of each boundary point. If the surface is smooth at the null hypothesis density, then the half-space conal geometry described above still holds true. However, the problem with "edges" is now more difficult to analyze, and one could lose the convex cone structure that guaranteed that using $0.5\chi_0^2 + 0.5\chi_1^2$ would be anticonservative.

Despite these technical difficulties, we think a reasonable procedure is the use of the distribution $0.5\chi_0^2 + 0.5\chi_1^2$ as a guide for constructing tests and confidence intervals. Inverting this test gives an upper confidence interval of the form

$$\{\pi: \mathrm{lrs}(\pi) \leq \chi_1^2(2\alpha)\} = [\pi_L^*, 1].$$

If we examine Table 8.2, we see that the 95% lower limit for π^* in this data is about 24%, still a rather large lack-of-fit fraction.