CHAPTER 7

# Extending the Method

In this chapter we present some simple extensions of the NPMLE theorem that solve problems that are similar, but not identical, in structure. We consider three situations: first, a class of problems in which the unknown latent distribution appears in the likelihood in a ratio form; second, the question of maximizing a mixture likelihood with linear constraints on the latent distribution; third, the problem of estimating the latent distribution with a continuous density function.

**7.1. Problems with ratio structure.** We start with a simple example that illustrates a problem in which the unknown latent distribution shows up in the likelihood in a ratio form.

7.1.1. *Example: Size bias.* Suppose that $X_1, \ldots, X_n$ are positive-valued random variables, but they arose from a population that was sampled not randomly, but with probabilities that are proportional to some positive function $w(x)$ of the variable of interest. That is, suppose the underlying distribution of the variable $X$ is $G$, with density $g$, but the sampling is from the density proportional to $w(x)g(x)$.

A classic example of this type would be if we were to sample vacationers in a hotel lobby and ask how long they were staying in the hotel. The vacationers who have longer stays are more likely to be included in the sample.

The nonparametric MLE problem is then to find the underlying distribution $G$ given knowledge of the sampling weights $w(x)$. We can write the likelihood, for discrete $G$, as

$$L(G) = \prod \frac{w(x_i)G(\{x_i\})}{\int w(x)\, dG(x)} = \prod \frac{\int w(x_i)\mathcal{I}[\phi = x_i]\, dG(\phi)}{\int w(x)\, dG(x)}.$$

The question is, how do we maximize such a likelihood, which now has the latent distribution in both numerator and denominator?

7.1.2. *NPMLE with ratio structure.* We first generalize the problem. We desire to solve a problem of the form

$$\sup_G \prod_{i=1}^{n} \left[ \frac{\int h_i(\phi)\, dG(\phi)}{\int H(\phi)\, dG(\phi)} \right]. \tag{7.1}$$

We require that $H(\phi)$ be positive-valued. The first step is to reparameterize the problem by forming a reweighted version of $G$:

$$dQ(\phi) = \frac{H(\phi)\, dG(\phi)}{\int H(\phi)\, dG(\phi)}. \tag{7.2}$$

If this is done, check that we can rewrite the original problem as

$$\sup_Q \prod \left[ \int \frac{h_i(\phi)}{H(\phi)}\, dQ(\phi) \right]. \tag{7.3}$$

Now we note that this is exactly of the mixture NPMLE form, where we use the likelihood kernels $L_i(\phi) = h_i(\phi)/H(\phi)$. Thus we can solve this problem to find a maximizing $\hat{Q}$. To find the NPMLE for the original problem, we must undo the transformation (7.2), and we obtain the following result:

PROPOSITION 27. *If $\hat{Q}$ solves the modified problem* (7.3)*, then*

$$d\hat{G}(\phi) = \frac{[1/(H(\phi))]\, d\hat{Q}(\phi)}{\int [1/(H(\phi))]\, d\hat{Q}(\phi)}.$$

*solves the original problem* (7.1).

7.1.3. *Example: Size bias.* In the size bias problem, we have $H(\phi) = w(\phi)$ and

$$h_i(\phi) = w(\phi) \mathcal{I}[\phi = x_i].$$

It follows that the likelihood kernels for the transformed problem are

$$L_i(\phi) = \frac{h_i(\phi)}{H(\phi)} = \mathcal{I}[\phi = x_i].$$

Therefore the transformed problem is exactly that for which the empirical CDF, mass $n^{-1}$ at each $x_i$, is the solution $\hat{Q}$. Following through the next step, we obtain the standard solution to the weighted sampling problem, that $\hat{G}$ has the form

$$\sum_i \left[ \frac{w(x_i)^{-1}}{n \sum_j w(x_j)^{-1}} \right] \Delta_{x_i}.$$

7.1.4. *Example: Weibull competing risks.* We introduce, with a motivating Weibull example, another class of problems in which the maximum likelihood solution requires solving a ratio problem. Suppose we have a piece of machinery with an unknown number of independent parts that are each subject to failure with a Weibull lifetime distribution. We observe the first time to failure. Thus we are in a setting where there are an unknown number of competing risks, assumed to be independent. We also do not observe the cause of failure.

We let $T_1, \ldots, T_v$ be the latent failure times of the competing sources of risk and so

$$X = \min\{T_1, \ldots, T_v\}$$

is the observed random variable. Since the latent times are independent, the *cumulative hazard function* of the observed variable, say $H(t)$, is the sum of the individual cumulative hazards for the latent variables, in the Weibull case, $H_r(t) = \lambda_r t^{\alpha_r}$. That is,

$$H(t) = \sum_{r=1}^{v} H_r(t) = \sum_{r} \lambda_r t^{\alpha_r}.$$

We can put this into a mixture format by setting

$$\Lambda = \sum \lambda_r$$

and defining the discrete distribution $G$ to have mass $\lambda_r/\Lambda$ at $\alpha_r$. If this is done, we can write the cumulative hazard function in the form

$$H_X(t) = \Lambda \cdot \int t^{\alpha}\, dG(\alpha).$$

Thus we can see that if we have a problem with an unknown number of competing Weibulls, we have a *mixed hazard function.* We note also that if there are multiple latent failure variables with the same value of $\alpha$, we cannot identify the separate $\lambda_i$ from this hazard function because they show up in the hazard only through their total. We also note that the hazard intensity rate has a mixed form:

$$h_X(t) = \frac{d}{dt} H_X(t) = \Lambda \cdot \int \alpha t^{\alpha-1}\, dG(\alpha).$$

We would like to maximize the likelihood of a sample of observations as a function of the unknown $(\Lambda, G)$.

7.1.5. *Mixed hazards NPMLE.* There is an obvious generalization of the Weibull formulation in which we have a family of distributions whose cumulative hazard can be expressed as a mixture of kernel cumulative hazards,

$$H_X(t) = \Lambda \cdot \int K(t;\alpha)\, dG(\alpha) = \Lambda \cdot K(t;G),$$

and whose hazard rate therefore has the form

$$h_X(t) = \Lambda \cdot \int k(t;\alpha)\, dG(\alpha) = \Lambda \cdot k(t;G),$$

where $k$ is the derivative of $K$ with respect to $t$. Such a formulation can arise from a competing risk framework, as in the Weibull example.

However, this formal structure can also arise if we partition the time axis into regions $A_r$ of unknown but constant hazard, writing kernel hazards of the form $k(t;r) = \mathcal{I}[t \in A_r]$, and obtaining the mixed hazard model

$$h_X(t) = \sum \lambda_r \mathcal{I}[t \in A_r].$$

In this case the "latent variable" $\alpha$ is discrete, corresponding to the index $r$ of the interval, and the setting is parallel to the mixture problem with known component distributions.

We now consider the likelihood for the problem, referring the reader to the paper by Hsi, Lindsay and Lynch (1992) for details on how to incorporate censoring. We recover the density for $X$ from the hazard specification as

$$f_X(t) = h_X(t) \exp(-H_X(t)).$$

It follows that the likelihood for a sample has the form

$$L(\Lambda, G) = \exp\left(-\Lambda \sum K(x_i; G)\right) \Lambda^n \prod k(x_i; G).$$

We next fix $G$ and maximize the likelihood over $\Lambda$ to find

$$\hat{\Lambda}_G = n \left( \sum K(x_i; G)\right)^{-1} = \bar{K}_G^{-1}.$$

It follows that the profile likelihood for the unknown distribution $G$ has the form

$$L(\hat{\Lambda}_G, G) = e^{-n} \prod \left[ \frac{k(x_i; G)}{\bar{K}_G} \right].$$

However, this problem is exactly of the ratio type we have described before and so we can solve it from the NPMLE theorem by transformation.

We note that Hsi, Lindsay and Lynch (1992) obtained this solution through a more difficult route, and so this presentation has its value in showing the simple structure that lies behind the problem. See the cited paper for further details on the applications of this model.

**7.2. NPMLE with constraints on $Q$.** There are a number of circumstances in which we might wish to maximize the nonparametric mixture likelihood under constraints on the unknown latent distribution $Q$. We might, for instance, need for identifiability reasons to constrain it to have mean 0 and variance 1. Another situation in which we might want to employ constraints is as follows.

7.2.1. *Profile likelihood.* Suppose we wish to form a profile likelihood based on some function of $Q$ that is of interest. To be specific, let us say the mean value of the latent distribution is of interest:

$$\tau(Q) = \int \phi \, dQ(\phi).$$

To form a profile likelihood $\mathbb{L}$ in such a nonparametric setting, we calculate for each fixed value of $\tau_0$, the solution to a constrained maximum likelihood problem,

$$\mathbb{L}(\tau_0) = \sup\{L(Q): Q \ni \tau(Q) = \tau_0\}. \tag{7.4}$$

The results of Owen (1988) suggest that such a *nonparametric profile likelihood* might well give us a method of performing tests and constructing confidence intervals for sufficiently smooth functions $\tau(Q)$ of the latent distribution.

Although the asymptotic theory is not yet available, we believe that profile likelihood intervals will provide a value tool for understanding which features of the latent distribution are trustworthy and which are poorly determined by the data. If profile likelihood intervals are carried out with a good optimization routine, they may be substantially more time efficient than bootstrapping and also provide a more natural way to construct confidence sets in more than one dimension.

Thus we have a modified optimization problem, and our goal here is to provide a theory for its solution.

7.2.2. *Linear constraints.* In this chapter we will consider only linear constraint problems. One type of such constraint is the *linear equality constraint*, by which we mean there are a set of functions $h_1(\phi), \ldots, h_a(\phi)$ and a set of constants $h_1^o, \ldots, h_a^o$ such that we wish to maximize the likelihood subject to the following restrictions on $Q$:

$$\begin{aligned} \int h_1(\phi)\, dQ(\phi) &= h_1^o \\ &\vdots \qquad\quad \vdots \\ \int h_a(\phi)\, dQ(\phi) &= h_a^o. \end{aligned} \tag{7.5}$$

We can add yet more flexibility in fitting profile likelihoods by allowing an additional set of *linear inequality constraints*:

$$\begin{aligned} \int k_1(\phi)\, dQ(\phi) &\le k_1^o \\ &\vdots \qquad\quad \vdots \\ \int k_b(\phi)\, dQ(\phi) &\le k_b^o. \end{aligned} \tag{7.6}$$

7.2.3. *Examples with linear constraints.* As examples of the equality constraint, the moments of the distribution of $Q$ are the most obvious parameters of interest. However, we can also construct a profile likelihood for the distribution function of $Q$ at a particular fixed value, say $Q(\phi_0)$, by using an indicator function

$$h_1(\phi) = \mathcal{I}[\phi \le \phi_0]$$

to construct the linear equality constraint.

Here is another example that is less obvious: Suppose we are interested in constructing a confidence interval for a posterior empirical Bayes function, such as

$$E[g(\Phi) \mid X = x_0] = \frac{\int g(\phi) f(x_0; \phi)\, dQ(\phi)}{\int f(x_0; \phi)\, dQ(\phi)}.$$

(Note that $x_0$ is a constant here.) Although this posterior mean is not a linear function of $Q$, when its value is fixed at some constant $c$, the maximization takes place over a set of $Q$ defined by the following linear equality constraint:

$$\int [g(\phi) f(x_0; \phi) - c f(x_0; \phi)]\, dQ(\phi) = 0.$$

If we wish to do inference on the quantiles of $Q$, we can turn to the linear inequality constraints version of the problem. For example, if we fix the median of $Q$ to be a specified value, say $c$, then we can maximize the likelihood subject to the two simultaneous linear inequalities:

$$\int \mathcal{I}[\phi \le c]\, dQ(\phi) \ge 0.5,$$
$$\int \mathcal{I}[\phi \ge c]\, dQ(\phi) \ge 0.5.$$

In this fashion, one could construct the profile likelihood of the median of $Q$.

7.2.4. *The constrained NPMLE.* We now consider the properties of the NPMLE if there are equality and inequality constraints in the form (7.5) and (7.6). Because the constraints are linear, we can extend our previous geometric analysis to allow for their consideration. We construct an extended likelihood vector in $(D + a + b)$-dimensional Euclidean space,

$$\mathbf{L}^*(\phi) := \begin{pmatrix} \mathbf{L}(\phi) \\ \mathbf{h}(\phi) \\ \mathbf{k}(\phi) \end{pmatrix},$$

and we let its convex hull be $\mathcal{M}^*$. The elements of $\mathcal{M}^*$ are of the form

$$\begin{pmatrix} \mathbf{L}(Q) \\ \mathbf{h}(Q) \\ \mathbf{k}(Q) \end{pmatrix}.$$

Let

$$\mathbf{x} = \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \\ \mathbf{r} \end{pmatrix}$$

be an arbitrary vector of dimension $D + a + b$, and define the log likelihood objective function on this space by

$$l(\mathbf{x}) = \sum_{i=1}^{D} n_i \ln(p_i),$$

noting that the last $a+b$ coordinates are given zero weight, so that the objective function is not strictly concave any longer.

Next, we define the suitable set over which the optimization will take place. The set of latent distributions that give extended likelihood vectors that meet our constraints lie in the set

$$\mathcal{M}^*_{\text{cons}} = \mathcal{M}^* \cap \{\mathbf{x}: \mathbf{q} = \mathbf{h}^o\} \cap \{\mathbf{x}: \mathbf{r} \leq \mathbf{k}^o\}.$$

The observation to be made here is that $\mathcal{M}^*_{\text{cons}}$ is a convex set, and so we can establish some results directly about this optimization problem.

PROPOSITION 28. *If the likelihood vector curve is closed and bounded and the mixture set contains points of positive likelihood, then:*

1. *If there are only linear equality constraints, then there exists a unique maximizing vector* $\hat{\mathbf{L}}$.
2. *If there are both equality and inequality constraints, then the maximum likelihood vector may not be unique, but there does exist a convex set of maximum likelihood solutions.*
3. *In either case, all solutions can be represented as a mixture with* $D + a + b$ *or fewer components.*

PROOF. For part 1, we note that the equality constraints form $\mathcal{M}^*_{\text{cons}}$ by taking a slice through $\mathcal{M}^*$. The resulting cross section set is still of dimension $D$ and the objective function is still strictly concave on this set. The strict concavity gives the uniqueness. Concavity implies the presence of the solution on the boundary of $\mathcal{M}^*_{\text{cons}}$ and hence the boundary of $\mathcal{M}^*$.

For part 2, we lose strict concavity of the objective function on $\mathcal{M}^*_{\text{cons}}$ when we have inequality constraints, because the coordinates of $\mathbf{x}$ corresponding to the inequality constraints can now be varied without affecting the objective function. Under nonstrict concavity, we can only make the weaker statement.

Part 3 follows from the fact that the solutions must lie in the boundary of the set $\mathcal{M}^*$, so we can apply Carathéodory's theorem with the dimension reduced by one. □

7.2.5. *A simple algorithm.* Problems of this type can be solved by the technique of Lagrange multipliers. We will describe here the case where all constraints are of the linear equality type. Our problem is then to jointly maximize over $Q$ and the Lagrange multipliers $\lambda_1, \ldots, \lambda_a$ the objective function

$$\sum n_i \ln(L_i(Q)) + \lambda_1[h_1(Q) - h_1^o] + \cdots + \lambda_a[h_a(Q) - h_a^o].$$

Our approach will be to treat the Lagrange multipliers $\boldsymbol{\lambda}$ as fixed, and to maximize just over $Q$ at first. We can readily find the gradient function for this new criterion to be

$$D_{Q,\lambda}(\phi) = D_Q(\phi) + \sum \lambda_r[h_r(\phi) - h_r(Q)]. \tag{7.7}$$

With this gradient function, we are ready to solve the problem. We can either use a straightforward gradient algorithm from this point, or we can use the EM algorithm with constraints, and just use this gradient function to

check the convergence. Our output will be a mixture solution $\hat{Q}_{\boldsymbol{\lambda}}(\phi)$. Note that the Lagrange gradient (7.7) does *not* depend on the initial conditions specified, but only the multipliers $\boldsymbol{\lambda}$ themselves, and so the same is true of $\hat{Q}_{\boldsymbol{\lambda}}(\phi)$.

Our next observation is that even though this $\hat{Q}_{\boldsymbol{\lambda}}(\phi)$ need not satisfy our initial constraints, we can easily solve for the set of linear constraints that it actually does satisfy:

$$\begin{aligned} \int h_1(\phi)\, dQ(\phi) &= \int h_1(\phi)\, d\hat{Q}_{\boldsymbol{\lambda}}(\phi) := h_1^*(\boldsymbol{\lambda}) \\ \vdots \qquad & \qquad \vdots \qquad \vdots \\ \int h_a(\phi)\, dQ(\phi) &= \int h_a(\phi)\, \hat{Q}_{\boldsymbol{\lambda}}(\phi) := h_a^*(\boldsymbol{\lambda}). \end{aligned}$$

Checking backward, we find that if we had started with $h_1^*(\boldsymbol{\lambda}), \ldots, h_a^*(\boldsymbol{\lambda})$ as our initial constraints $h_1^o, \ldots, h_a^o$, then $\hat{Q}_{\boldsymbol{\lambda}}(\phi)$ and $\lambda_1, \ldots, \lambda_a$ would have been the solutions to the Lagrange multiplier problem.

That is, we did not obtain the value $\mathbb{L}(h_1^o, \ldots, h_a^o,)$ of the profile likelihood with our original $h_1^o, \ldots, h_a^o$, but we do end up with another value of the profile likelihood, namely,

$$\mathbb{L}(h_1^*, \ldots, h_a^*).$$

It follows that reconstruction of the profile likelihood can be carried out by interpolation over a selected region of Lagrange multiplier values, chosen to give $h^*$ values in the right region. This operation is particularly simple if there is a single constraint function $h(\phi)$ which is nonnegative, because then $h^*(\lambda)$ is monotonically decreasing in $\lambda$. [*Exercise.*]

**7.3. Smooth estimates of $Q$.** This author is not, on the whole, in favor of the idea of using continuous densities to estimate the latent distribution. Although the argument that nature is continuous has some compelling features, in most models the level of information about the latent distribution is simply too small to consider any discrimination about the form of this distribution. In essence, as indicated in Chapter 1, one can obtain reasonable estimates or intervals only for the smoothest of functionals of the latent distribution, and the goal of actually discerning the true density is typically impossible for all practical purposes. However, there are certainly some examples, especially in image analysis, where the prior information that the picture is relatively smooth is important in performing a useful analysis.

Therefore, some further references in this regard will be given for the sake of readers with a particular statistical interest in a smooth estimator. In addition, at the end there will be a suggestion of my own about how to directly use the nonparametric likelihood approach and still obtain a smooth estimator by maximum likelihood.

7.3.1. *Roughening by smoothing.* One approach that has been taken is to take a very smooth initial estimator, such as normal latent distribution, and apply the EM algorithm to it for a few steps. The EM algorithm formula

generalizes very simply to the gradient updating of density functions by the formula

$$q_{\text{em}}(\phi) = q_c(\phi)[1 + n^{-1} D_Q(\phi)].$$

If we start with a normal density, the first step takes us to a mixture of $n$ normal densities, each with smaller variance. Laird and Louis (1991) call this "smoothing by roughening." A similar approach is recommended by Vardi and Lee (1993).

7.3.2. *Deconvolution.* Another approach is to extend the idea of kernel-based density estimation into the domain of the latent distribution. There have been a number of papers in this regard, of which we might mention Fan (1991). The method usually relies on the convolution type mixture and is found in the literature under the keyword deconvolution. The most important lesson from this literature is that the best possible rates of convergence are extremely poor, and therefore density estimation is practically impossible.

7.3.3. *Series expansion.* Another set of workers have developed analogues of series expansions to use for fitting the latent distribution smoothly. Gallant and Nychka (1987) called this a semi-nonparametric approach. A similar approach is carried out by Walter and Hamedani (1991) in the context of empirical Bayes estimation.

7.3.4. *A likelihood method.* We can easily extend the nonparametric maximum likelihood idea to construct estimates with smooth densities possessing any prespecified degree of smoothness. In particular, we can choose the estimated density to have a likelihood nearly that of the global nonparametric maximum likelihood estimator.

We start with a family of densities $g(\phi; \theta, \tau)$ on the parameter space $\Omega$. The parameter $\theta$ is assumed to determine the central location of this density and $\tau$ is a dispersion parameter, with the distribution concentrating about $\theta$ as $\tau \to 0$. Of course, the normal density is such a family, but in many situations the natural conjugate density family might be more suitable because it would avert numerical integration problems.

If we have a known component model, there might be a natural way to construct a kernel family that gives some target smoothness to the probabilities over physically neighboring components. For example, in positron emission tomography, one could construct a discrete distribution over the sites $\phi$ that are neighbors to site $\theta$, with the dispersion parameter $\tau$ reflecting the amount of mass spread to the neighbors.

Ideally this construction is done so that we can explicitly calculate the marginal distribution of $X$ when $g$ is the latent distribution. That is, we desire

$$f^*(x; \theta_j, \tau) = \int f(x; \phi) g(\phi; \theta_j, \tau)\, d\phi$$

to be readily calculable. Note that under our specifications, as $\tau \to 0$ this density should go to the unicomponent density $f(x; \theta_j)$, and we so assume.

Suppose we replace our basic discrete class of latent distributions $\sum \pi_j \Delta_{\phi_j}$ with arbitrary convex combinations of latent distributions of the form

$$\sum \pi_j g(\phi; \theta_j, \tau).$$

The resulting class of mixture densities for $X$ can now be expressed as convex combinations of new family of basic densities, namely,

$$X \sim \sum \pi_j f^*(x; \theta_j, \tau) = \int f^*(x; \theta, \tau)\, dH(\theta).$$

Now we can describe a strategy. For each fixed $\tau$ we can calculate the NPMLE for this new family of mixtures. If the answer is $\hat{H}_\tau$, then we have a resulting smooth density estimator for the original problem, namely,

$$q_\tau(\phi) = \int g(\phi; \theta, \tau)\, d\hat{H}_\tau(\phi),$$

with corresponding distribution $\hat{Q}_\tau$. The selection of $\tau$ can be based on likelihood considerations. The NPMLE for the problem necessarily has higher likelihood, but if we target a fixed difference

$$\ln L(\hat{Q}) - \ln L(Q_\tau) = \delta,$$

then consistency will follow from the Kiefer–Wolfowitz (1956) result.

Moreover, if we set $\delta$ sufficiently small, say 0.005, the arguments of Chapter 6 imply that the resulting estimator will differ very little from the NPMLE in inference for nonparametric functionals. However, this will come at the cost of losing much of the smoothness of the estimator.