Chapter 1
Frailty, Profile Likelihood and Medfly Mortality

Roger Koenker and Jiaying Gu

Abstract Unobserved heterogeneity is an increasingly common feature of statistical survival analysis where it is often referred to as frailty. Parametric mixture models are frequently used to capture these effects, but it is sometimes desirable to consider nonparametric mixture models as well. We illustrate the latter approach with a reanalysis of the well-known large scale medfly mortality study of Carey, Liedo, Orozco, and Vaupel (1992). Recent developments in convex optimization are exploited to expand the applicability of the Kiefer-Wolfowitz nonparametric maximum likelihood estimator for mixture models. Some ensuing problems of profile likelihood are also addressed.

1.1 Introduction

The notion of frailty to describe unobserved heterogeneity of population risks has become a familiar feature of demographic analysis since Vaupel, Manton, and Stollard (1979), and has gradually spread to other statistical domains. A valuable early exposition of the impact of frailty in models of treatment evaluation is provided by Shepard and Zeckhauser (1980). Often, as in the aforementioned sources, parametric models are posited for the frailty effects, but it is usually difficult to justify such assumptions given the unobserved nature of the frailty components. Recent progress in estimation and inference for general, non-parametric mixture models has opened the way to a more flexible approach. We will illustrate some features of such an approach with a reanalysis of the influential Carey, Liedo, Orozco, and Vaupel (1992) study of medfly mortality.

Roger Koenker
University of Illinois, Urbana, IL, 61801, USA, e-mail: `rkoenker@uiuc.edu`

Jiaying Gu
University of Illinois, Urbana, IL, 61801, USA, e-mail: `gu17@uiuc.edu`

1.2 Data

In the largest of the three experiments reported in Carey, Liedo, Orozco, and Vaupel (1992), 1.2 million Mediterranean fruit flies (*Ceratitis Capitata*) were raised in a large facility in Mexico,

> "...Pupae were sorted into one of five size classes using a pupal sorter. This enabled size dimorphism to be eliminated as a potential source of sex-specific mortality differences. Approximately, 7,200 medflies (both sexes) of a given size class were maintained in each of 167 mesh covered, 15 cm by 60 cm by 90 cm aluminum cages. Adults were given a diet of sugar and water, ad libitum, and each day dead flies were removed, counted and their sex determined ..."

The primary objective of the experiment was to study the upper tail of the mortality distribution, an endeavor that revealed several surprising features.

1.3 Declining Mortality Rates

Prior to this experiment it was an article of faith throughout biology that within species mortality (hazard) rates were monotonically increasing with age. Indeed it was commonly suggested that each species had a species specific upper bound for age rendering the whole notion of investigating the "tail behavior" of the mortality distribution pointless. In Figure 1.1 we plot raw daily mortality rates from the experiment and superimpose a smoothed, geometric moving average curve. More explicitly, let y_t denote the number of flies alive (at risk) at the beginning of day t, then the raw mortality rates plotted in Figure 1.1 are, $r_t = 1 - y_{t+1}/y_t$, and the smoothed (geometric) weekly moving average. Contrary to the received wisdom, mortality rates actually declined after about age 60. This finding provoked an extensive reappraisal of the biology of aging. The observed decline in mortality offered no consolation to the 99.8 percent of the flies that were already dead by age 60, but to the remaining, more than 2000 less frail ones, it offered some hope of a prolonged retirement. The oldest flies in the experiment expired on day 172.

How should we interpret this remarkably long tail? One explanation, suggested by Vaupel and Carey (1998), was that the population under study was really a mixture of several subpopulations of varying frailties. Rather than assume a particular parametric form for the mixing distribution, Vaupel and Carey adopted a nonparametric mixture model. While their two page note in *Science* precluded a detailed description of their computational methods, we have been able to "reverse engineer" an approach that closely mimics the results reported in their Figure 1.

The first question is: What are we mixing? Here we follow Vaupel and Carey and consider both Gompertz and Weibull mixtures. The Gompertz model assumes that log hazard is linear in age, while the Weibull model assumes that log hazard is linear in log age. Figure 1.2 illustrates raw log-hazard rates plotted against age, and superimposed are two estimates of the baseline model. The dashed red line represents the estimated baseline Gompertz model fit to the data for the first 15 days
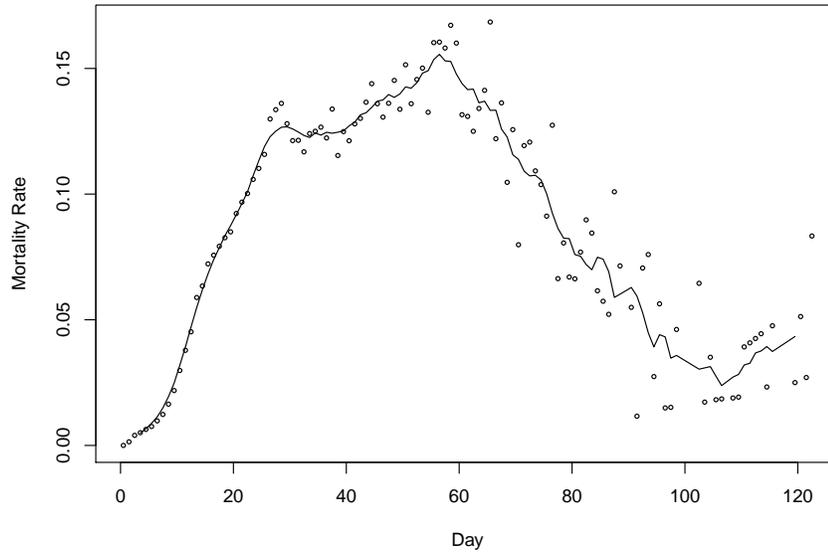
Fig. 1.1 Raw Daily Medfly Mortality Rates and Moving Average Smooth

of the experiment by weighted least squares with weights given by the relative frequencies of the daily counts. It appears that the first day is an outlier in this plot, however since few flies died on the first day it exerts little influence on the fitted line. The blue solid curve represents the baseline Weibull fit based on the first 20 days of the experiment. How many observations to use to estimate the parameters of the baseline model is obviously somewhat debatable, in this respect the problem is somewhat similar to the notorious controversies over how to choose k in the Hill estimator of the Pareto exponent. We won't indulge in further speculation about these choices, but simply remark that our k selection yields baseline Gompertz hazard of $h(t) = 0.002 \exp(0.24t)$, while Vaupel and Carey use $h(t) = 0.003 \exp(0.3t)$, and for the Weibull model we obtain $h(t) = 0.0004t^{1.85}$, against Vaupel and Carey's $h(t) = 0.001t^2$. The intercept in these models is not crucial, since the estimated mixture distribution is scale equivariant it simply fixes a normalization. The shape parameter is more important, but in both cases our approach of fitting the left tail of the distribution yields rather similar estimates to those employed by Vaupel and Carey. An intriguing, open theoretical and practical question remains: can likelihood methods be brought to bear to estimate these shape parameters. We will return to this question when we consider profile likelihoods.

Given our estimated baseline models it is now time to address the problem of estimating the mixing, or frailty, distribution. There is a long history and exten-
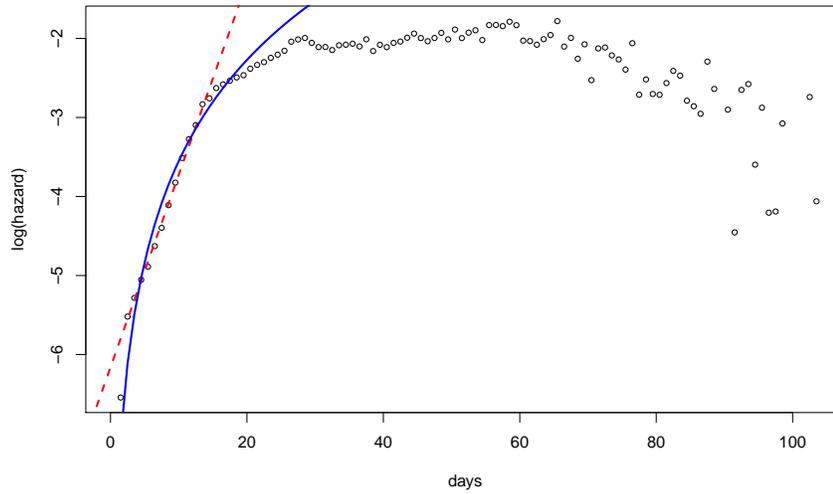
Fig. 1.2 Estimated Baseline Gompertz and Weibull Hazard Models: Linear (Gompertz) and log linear (Weibull) fits to the initial k observations of raw daily log mortality rates.

sive literature on this subject, Lindsay (1995) provides a thorough overview. Kiefer and Wolfowitz (1956) demonstrated that such mixture models were consistently estimable under weak conditions by maximum likelihood. If we write the baseline density as $\varphi(x, \theta)$ and the mixture density as,

$$g(x) = \int \varphi(x, \theta) dF(\theta),$$

then given iid observations, $x_1, \cdots, x_n$ from g, we wish to solve

$$\max_{F \in \mathscr{F}} \sum_{i=1}^{n} \log(g(x_i)).$$

Following Laird (1978), the EM algorithm, or a variant of it, has been employed to solve such problems. However, EM is notoriously slow to converge. Koenker and Mizera (2011) proposed an alternative computational strategy based on convex optimization. Let, $t_0 < t_1 < \cdots < t_m$ denote a grid of values for the potential mass points of the distribution F, and let f_i denote the mass associated with the ith grid interval. Then, we can rewrite the MLE problem as,

$$\max_{f \in \mathbb{R}^m} \{ \sum \log(g(x_i)) \mid g = Af, \ \sum f_i \Delta t_i = 1, \ f \geq 0 \},$$

where A denotes the n by m matrix with typical element, $\varphi(x_i, t_j)$, and g denote the n vector with typical element $g(x_i)$. This is a garden variety convex optimization problem that can be efficiently solved by modern interior point methods. We employ Mosek, Andersen (2010), for this purpose. The R package, REBayes, Koenker (2012), implements a variety of related problems, all of the computational results reported here were carried out in this environment.
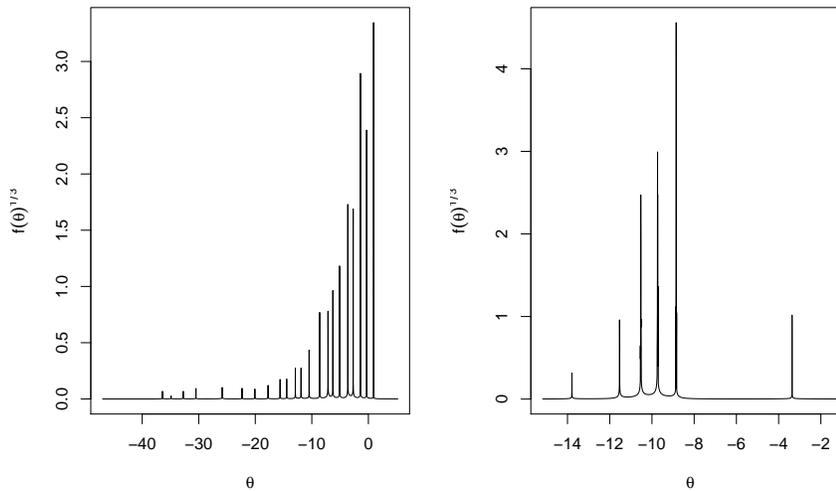
Fig. 1.3 Estimated Mixing Distributions for the Gompertz (left) and Weibull (right) Models

In Figure 1.3 we plot the two mixing distributions estimated by the Kiefer-Wolfowitz maximum likelihood procedure. Note that the vertical axis in these plots is the cube root of the density to exaggerate the smaller mass points that are nearly invisible on the original $f(\theta)$ scale. The Kiefer-Wolfowitz estimator is known to deliver a discrete distribution, here represented by a "density" with a small number of "almost" point masses. The Weibull model is considerably more parsimonious in this respect with only six distinct points of support. The implied hazard functions for the two estimated mixture densities are shown in Figure 4, superimposed over the raw mortality rates. Fewer mass points in the Weibull model translates to much smoother behavior of the hazard function, but this is ultimately traceable back to the forms of the base density, the Gompertz being more sharply peaked and consequently generating a rougher mixture. In both cases the mixing parameter θ functions as a scale parameter, but the mixing distribution is estimated on the $\log \theta$ scale, so we can interpret the mixing as convolution as with the familiar kernel density estimator.
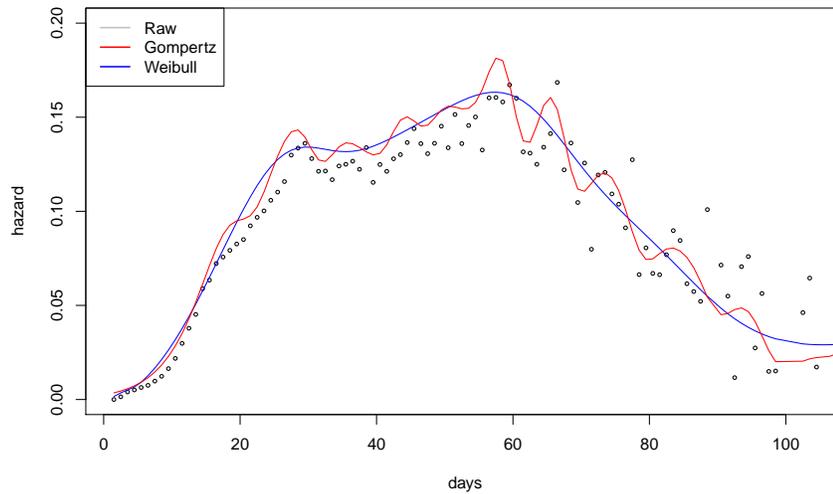
Fig. 1.4 Hazard Functions for the Estimated Gompertz and Weibull Models

1.4 Gender Crossover

An obvious source of *observed* heterogeneity is gender differences. Again, the Carey et al experiment revealed some surprising new facts. When we repeat our prior exercise fitting separate baseline Weibull models for males and females we obtain the results appearing in Figure 5. The Weibull model fits considerably better in both of these plots than in the previous aggregated plot, and considerably better than the corresponding Gompertz plots, so we will restrict attention henceforth to the Weibull model. Given the baseline models the Kiefer-Wolfowitz estimates of the mixture model yields the gender specific hazard functions of Figure 6. Several features of this plot are worth noting. Until about age 20 female mortality is higher than that of males, but after age 20 female mortality is substantially below male mortality. This crossover of the hazard functions clearly contradicts the proportional hazard assumption that is frequently made in survival analysis. The second crossover of the estimated hazard curves at about age 75 probably shouldn't be taken too seriously, but the initial crossing is quite precisely estimated and induces a crossing of the estimated gender specific survival functions at about age 36. It is impossible to resist noting that this pattern reverses the typical finding for human populations for which males are more frail than females with a possible crossover only at very advanced ages.
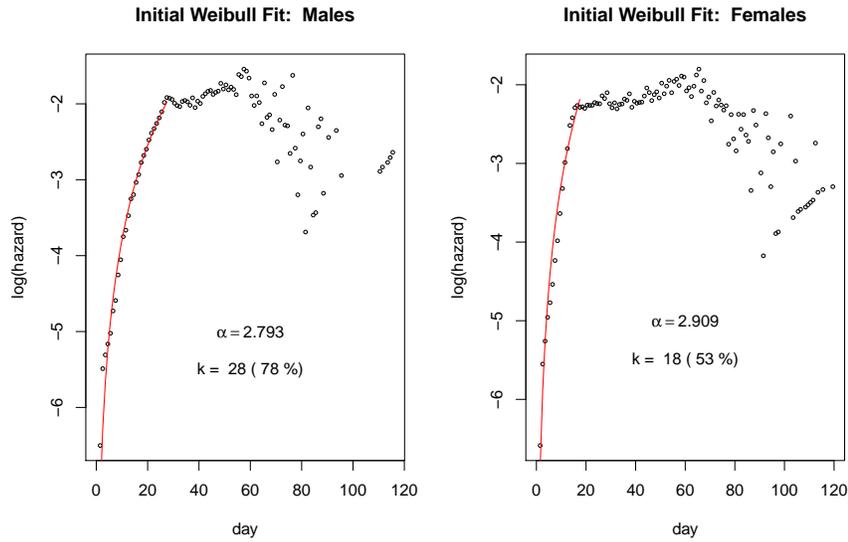
Fig. 1.5 Gender Specific of Baseline Weibull Models: Weighted least squares fitting of the initial *k* observations on daily mortality rates. The percentage of the sample population dead by day *k* is given in parentheses. The estimated shape parameter of the baseline Weibull model is α.

1.5 Profile Likelihood and Covariate Effects

If nonparametric maximum likelihood estimation of frailty effects were restricted to univariate survival models it would still be a very valuable addition to the statistical repertoire, but it would be much more useful if it could be extended to semiparametric applications including covariate effects. Of course we already have the proportional hazard model for this purpose, however frailty offers another valuable perspective. Factorization of the likelihood makes the proportional hazard assumption especially convenient from a computational viewpoint; the Weibull mixture model has no comparable factorization, nevertheless it is possible to employ a profile likelihood formulation to elaborate the model to include covariate effects.

From the beginning a controversial aspect of the Carey experiment was the effect of cage density. Critics claimed that flies raised in more crowded cages would be more likely to die earlier. Carey, Curtsinger, and Vaupel (1993) responded that the cage density was quite low after 60 days, only 16 flies per cage, on average, survived beyond this age, so it seemed difficult to attribute differences in mortality rates in elderly medflies to differences in crowding. To investigate whether differences in initial cage density had a significant impact on mortality we considered a model in which it entered as as a linear multiplicative scale shift in the Weibull model, that is the baseline Weibull scale becomes $\theta_0 \exp(d_i \beta)$ where d_i denotes initial cage den-
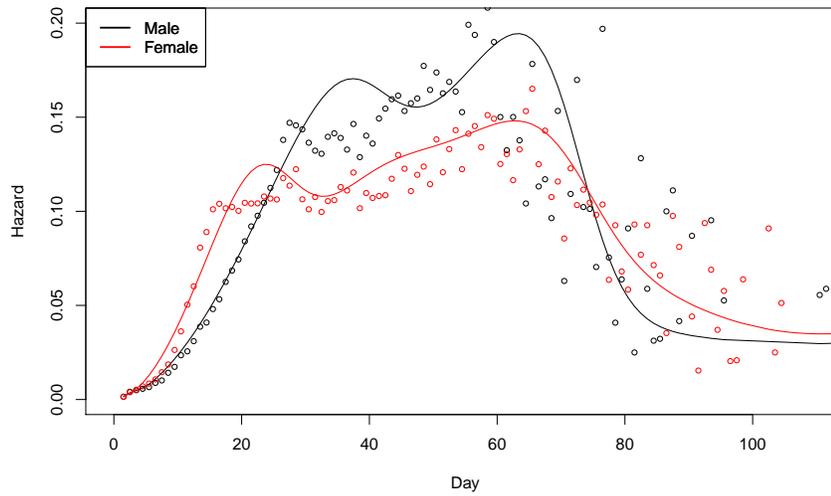
Fig. 1.6 Gender Specific of Hazard Functions for the Weibull Mixture Model: Raw daily mortality rates are plotted in black for males and red for females, superimposed are the estimated hazard functions for the Weibull mixture models using the baseline models shown in Figure 1.6.

sity. To estimate the density effect parameter, β, we simply evaluated the profiled likelihood on a grid of values on the interval $[-1,1]$, yielding Figure 1.7. This exercise yields a point estimate of about $\hat{\beta} = -0.5$ that is quite precise, at least if we are to believe the confidence bounds implied by the classical Wilks, $2\log\lambda \rightsquigarrow \chi^2_1$, theory. Leaving the reliability of such intervals to future investigation, we conclude simply that the negative estimated coefficient implies that higher density shifts the survival distribution to the right, thus prolonging lifetimes, and directly contradicting the conjecture of the Carey critics. This finding is confirmed by other methods, see for example Koenker and Geling (2001) where similar results are reported for both the Cox model and several quantile regression models.

The success of profile likelihood in a few cases prompts one to wonder how far similar methods can be extended to other semiparametric mixture settings. There is a considerable literature on this topic, pioneered by Lindsay. When profiling leads to fully adaptive estimation of structural parameters, not only do we get efficient estimates of those parameters, as a by-product we also get valid inference from the profiled likelihood ratio statistic, see Murphy and Van der Vaart (2000). The latter bonus is sometimes referred to as the Wilks phenomenon, see e.g. Fan, Zhang, and Zhang (2001).

But profiling is not always so perceptive; sometimes it can lead the unwary toward disaster. To illustrate this less benign side of profile likelihood for mixture models we would like to briefly reconsider estimation of the Weibull shape param-
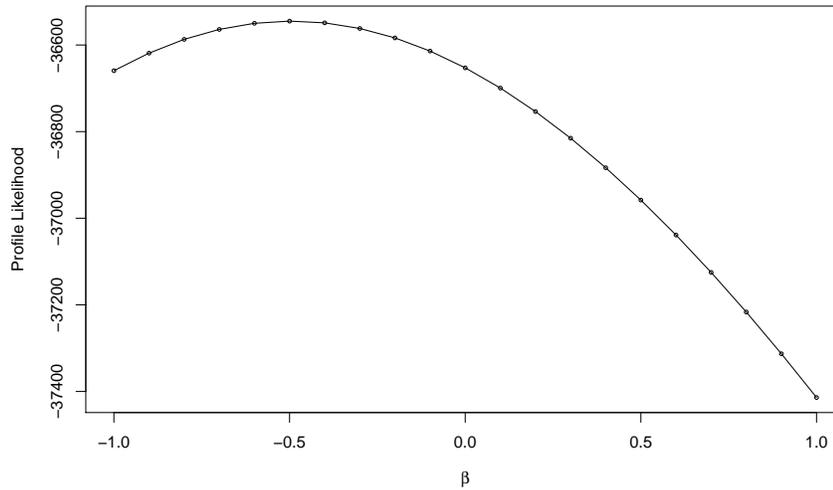
Fig. 1.7 Profile Likelihood for the Initial Cage Density Effect in the Weibull Mixture Model

eter, α based on the medfly data. In Figure 1.8 we show the profile likelihood for α based on the male medfly data. Based on our earlier results we know that $\alpha \approx 2.8$ fits the initial portion of the log hazard plot quite well. What does the profile likelihood have to say about it? The message is a bit confusing: the profile likelihood increases sharply up to about $\alpha = 2.8$, and then dramatically flattens out. In fact, closer examination reveals that the profile likelihood continues to increase beyond this value, but very, very gradually. Indeed, as $\alpha \to \infty$, the profile likelihood also tends to infinity. To understand this better it is helpful to consider how the estimated mixture distribution responds to changes in α. For small α, the estimated mixture distribution has only a single mass point, and this single mass point persists for a while, by the time we get to α between 2.5 and 3.0 though we have 5 or 6 mass points as in Figure 1.3. As α becomes larger we get more and more mass points, eventually yielding positive mass corresponding to virtually all the distinct observed values. This is reminiscent of the familiar Dirac catastrophe produced by kernel density bandwidths chosen by maximum likelihood. Indeed, the situation is quite similar, as α becomes large the effective bandwidth of the baseline Weibull model becomes narrower and more mass points are needed in the mixture distribution to mimic the density of the observed data.

So profile likelihood has failed us. Now what? There is a familiar litany of circumstances in which naive adherence to the principle of maximum likelihood leads to absurd results: various Gaussian examples in which driving variance parameters to zero yields unbounded likelihood at unlikely places in parameter space, estima-
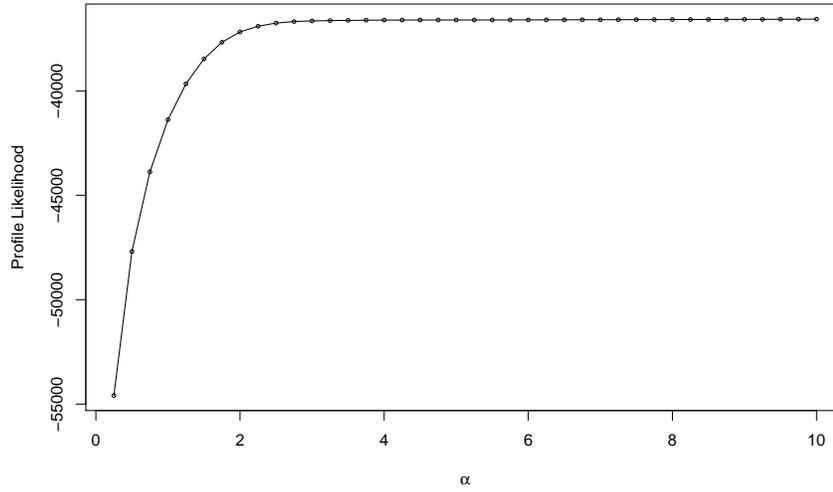
Fig. 1.8 Profile Likelihood for the Initial Cage Density Effect in the Weibull Mixture Model

tion of the threshhold parameter of the three parameter lognormal distribution, and many others. One approach that has proven successful in such situations is the maximum product spacing methods introduced by Cheng and Amin (1983) and Ranneby (1984). Roeder (1990) describes an application of this approach in astronomy that although based on Gaussian assumptions is qualitatively quite similar to our Weibull problem.

Log product spacings optimization can be viewed as a discretization of classical maximum likelihood. Let $G(x, \theta)$ denote the distribution function of a parametric model for a scalar random variable, X. Given a sample, $X_1, \cdots, X_n$ of identical copies of X, let

$$\Delta G_i(\theta) = G(X_{(i)}, \theta) - G(X_{(i-1)}, \theta),$$

for $i = 1, \ldots, n+1$ with $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$ and $X_{(i)} : i = 1, \cdots, n$ denoting the order statistics of the original sample. Since $G(X, \theta_0)$ is uniform when evaluated at the true parameter, θ_0 of the model, the $\Delta G_i(\theta_0)$ constitute a sample of uniform spacings for which there is an extensive theory. Considering

$$R_n(\theta) = \frac{1}{\sqrt{n+1}} \sum_{i=1}^{n+1} (\log(\Delta G_i(\theta)(n+1)) + \gamma)/(\pi^2/6 - 1)^{1/2}$$

with $\gamma \approx 0.577216$, the Euler constant, we have a normalized sum that satisfies a central limit theorem with a standard normal limiting distribution. Maximizing $R_n(\theta)$ with respect to θ requires computing the Kiefer-Wolfowitz mixture distribu-

tion, $\hat{G}(x, \theta)$, at each θ to obtain the profile log product spacing objective function. The function $R_n(\theta)$ behaves like the usual log-likelihood; this is to be expected since the summands can be viewed as difference quotient approximations of $g(\tilde{x}_i, \theta)$ for $\tilde{x}_i \in (X_{(i-1)}, X_{(i)})$. However, by avoiding the direct evaluation of densities we circumvent the pathological behavior of the log likelihood.

An important feature of the maximum product spacing method noted by Roeder (1990), is that for given θ, it selects an $\hat{G}(x, \theta)$ that is asymptotically equivalent to the mixture distribution estimated by nonparametric maximum likelihood that we have focused on thus far. For θ taking various values, we get a profiled objective function similar to the profiled nonparametric likelihood. Yet unlike the problematic profiled likelihood, the limiting form of $R_n(\theta)$ yields an estimating function centered at zero for the true parameter and a simple confidence interval construction for the structural parameter. Further details regarding the maximum product spacing method can be found in Roeder (1990), Roeder (1992) and Ekström (2008).

We have seen already that an α parameter that fits the left tail of the survival distribution can be estimated well by a simple regression of log hazards on log event times using data from the first few days of the experiment. This assumes that flies that only survive for the first k days are all from a homogeneous parametric survival model. When we move on to the semiparametric mixture model using all the observations, a natural question becomes how reasonable is it to assume a global value for α while allowing scale heterogeneity with frailty. We employ a first-order form of the log-product-spacing method and find that the test strongly rejects the mixture models with $\alpha = 2.85$. However, when we use only the observations surviving up to 50 days, a subsample that actually contains 99.5% of the full sample, we obtain a test statistic of only 0.33 and the model is not rejected. Similar conclusions are drawn when we estimate gender specific models. The message seems to be that the Weibull semiparametric mixture model fits the majority of the data quite well, but fails to perform adequately for the extreme right tail.

This conclusion may simply reassert that estimating a fixed shape parameter in the Weibull mixture model is an extremely difficult task; this is indeed the impression one gets from the prior literature. Hahn (1994) shows that the information matrix is singular for mixed Weibull proportional hazard model. When there are no covariates, the score function for α is identically zero, hence also the Fisher information. This means that the Weibull parameter can not be estimated at a root-n rate. Various estimation strategies for α are nevertheless available, for example Honoré (1990) Honoré (1997) and Ishwaran (1996). We would like to highlight what seems to be a somewhat neglected paper by Ishwaran (1999) discussing the information loss phenomenon for a class of semi parametric mixture models. Ishwaran shows that for the Weibull mixture model, there is information loss for α's bigger than the true value α_0, so that with $\alpha > \alpha_0$, one can find a mixing distribution that produces a model that is arbitrarily close to the true model in the sense of Hellinger distance. This corresponds to the flat region in our profile likelihood. On the other hand, as he notes, it is curious that the same information loss phenomenon does not occur for α's that are smaller than α_0. Whether one could take advantage of this asymmetric behavior for estimation of α is left for future investigations.

References

Andersen ED (2010) The MOSEK optimization tools manual, version 6.0, available from `http://www.mosek.com`

Carey J, Liedo P, Orozco D, Vaupel J (1992) Slowing of mortality rates at older ages in large medfly cohorts. Science 258:457–61

Carey J, Curtsinger J, Vaupel J (1993) Fruit fly aging and mortality. response to letters to the editor. Science 260:1567–1569

Cheng R, Amin N (1983) Estimating parameters in continuous univariate distributions with a shifted origin. Journal of the Royal Statistical Society Series B (Methodological) pp 394–403

Ekström M (2008) Alternatives to maximum likelihood estimation based on spacings and the kullback–leibler divergence. Journal of Statistical Planning and Inference 138:1778–1791

Fan J, Zhang C, Zhang J (2001) Generalized likelihood ratio statistics and wilks phenomenon. The Annals of Statistics 29:153–193

Honoré B (1990) Simple estimation of a duration model with unobserved heterogeneity. Econometrica 58:453–473

Honoré B (1997) A note on the rate of convergence of estimators of mixtures of weibulls, preprint, Princeton

Ishwaran H (1996) Uniform rates of estimation in the semiparametric weibull mixture models. Annals of Statistics 24:1560–1571

Ishwaran H (1999) Information in semiparametric mixtures of exponential families. Annals of Statistics 27:159–177

Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. The Annals of Mathematical Statistics 27:887–906

Koenker R (2012) Rebayes: An R package for empirical bayes methods, available from `http://www.econ.uiuc.edu/~roger/research/ebayes/ebayes.html`

Koenker R, Geling O (2001) Reappraising medfly longevity: A quantile regression survival analysis. J of Am Stat Assoc 96:458–468

Koenker R, Mizera I (2011) Convex optimization, shape constraints, compound decisions and empirical bayes rules, `http://www.econ.uiuc.edu/~roger/research/ebayes/ebayes.html`

Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 73:805–811

Lindsay B (1995) Mixture models: theory, geometry and applications. In: NSF-CBMS regional conference series in probability and statistics

Murphy SA, Van der Vaart AW (2000) On profile likelihood. Journal of the American Statistical Association 95:449–465

Ranneby B (1984) The maximum spacing method. an estimation method related to the maximum likelihood method. Scandinavian Journal of Statistics pp 93–112

Roeder K (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. Journal of the American Statistical Association 85:617–624

Roeder K (1992) Semiparametric estimation of normal mixture densities. Annals of Statistics 20:929–943

Shepard D, Zeckhauser R (1980) Long-term effects of interventions to improve survival in mixed populations. J Chronic Disease 33:413–433

Vaupel J, Carey J (1998) Compositional interpretations of medfly mortality. Science 260:1666–1667

Vaupel J, Manton K, Stollard E (1979) The impact of heterogeneity in individual frailty on the
      dynamics of mortality. Demography 16:439–454