REBAYES: AN R PACKAGE FOR EMPIRICAL BAYES MIXTURE METHODS

ROGER KOENKER AND JIAYING GU

Abstract. Models of unobserved heterogeneity, or frailty as it is commonly known in survival analysis, can often be formulated as semiparametric mixture models and estimated by maximum likelihood as proposed by Robbins (1950) and elaborated by Kiefer and Wolfowitz (1956). Recent developments in convex optimization, as noted by Koenker and Mizera (2014b), have led to dramatic improvements in computational methods for such models. In this vignette we describe an implementation contained in the R package **REBayes** with applications to a wide variety of mixture settings: Gaussian location and scale, Poisson and binomial mixtures for discrete data, Weibull and Gompertz models for survival data, and several Gaussian models intended for longitudinal data. While the dimension of the nonparametric heterogeneity of these models is inherently limited by our present gridding strategy, we describe how additional fixed parameters can be relatively easily accommodated via profile likelihood. We also describe some nonparametric maximum likelihood methods for shape and norm constrained density estimation that employ related computational methods.

1. Introduction

Empirical Bayes methods as conceived by Robbins (1956) are enjoying a robust revival stimulated by more bountiful data sources and new theoretical developments exemplified by Efron (2010). Mixture models have played a central role in this revival, and this has sparked renewed interest in the Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator (NPMLE) for mixtures. Relatively recent developments in convex optimization have dramatically improved computational methods for the Kiefer-Wolfowitz NPMLE, as described in Koenker and Mizera (2014b). To make these methods accessible to the research community we have developed an R package **REBayes** that incorporates a wide variety of nonparametric mixture models and provides Kiefer-Wolfowitz procedures for each of them.

The simplest univariate mixture model takes the form,

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

where φ is a known density, that we will refer to as the base density, and F is an unknown distribution function that we would like to estimate, given an iid sample from

Version: February 21, 2019. This research was partially supported by NSF grant SES-11-53548.

the mixture density g. The most familiar example would be the Gaussian location model with φ standard Gaussian, so,

$$g(x) = \int \varphi(x - \mu) dF(\mu).$$

This is the standard Gaussian sequence model and has been studied in many simulation experiments, including Johnstone and Silverman (2004), Martin and Walker (2014) and Castillo and van der Vaart (2012), and employed in many – typically genomic – applications. The objective of such analyses is a compound decision problem: Given an exchangeable sample, $X_1, ...X_n$ estimate the corresponding $\mu_1, ...\mu_n$ subject to quadratic loss. As noted by Robbins (1956) this yields the optimal Bayes rule,

$$\mathbb{E}(\mu|x) = x + g'(x)/g(x). \tag{1}$$

Efron (2011) calls this Tweedie's formula since Robbins attributes it to M.C.K. Tweedie, however it appears earlier in Dyson (1926) who credits it to the English astronomer Arthur Eddington. To turn this into a practical shrinkage formula we obviously need to choose an estimator for the mixture density g. Much of the earlier literature on this problem may be viewed as offering parametric empirical Bayes proposals in which F is specified up to a finite dimensional vector of hyperparameters. Prominent examples of this parametric strategy would be the **EbayesThresh** package described in Johnstone and Silverman (2004) and Johnstone and Silverman (2005), and the recent work of Efron (2016) and Efron (2010). In Gu and Koenker (2016a) we have made some comparisons with nonparametric Bayes procedures based on the Dirichlet process prior using the **DPpackage** of Jara et al. (2011). In our limited experience this leads to similar estimates of the mixing distribution as those of the NPMLE provided that the concentration parameter of the Dirichlet is small. See Liu (1996) for another comparison of Dirichlet and NPMLE methods.

More recently interest has focused on nonparametric estimation of the mixing distribution as in Efron (2011), Brown and Greenshtein (2009) and Jiang and Zhang (2009). The latter authors proposed using the Kiefer-Wolfowitz NPMLE to estimate F, and thereby g, and then to use the Tweedie formula. The main drawback of this proposal was the painfully slow convergence of the fixed point iteration of the EM algorithm used to compute the NPMLE. Koenker and Mizera (2014b), observing that the discretization suggested by Jiang and Zhang (2009) produced a convenient, finite dimensional convex optimization problem showed that the NPMLE could be implemented much more efficiently by standard interior point methods. In the next section we will briefly describe this implementation, and then turn to descriptions of various applications. Other recent applications of the **REBayes** package may be found in Dicker and Zhao (2016) and Jiang and Zhang (2015).

The extensive literature on estimating finite mixture models, that is models with a prespecified number of parametric components, faces a number of challenging problems: potentially unbounded and multi-modal likelihoods, lack of identifiability, as

well as selection of the number of components. The Kiefer and Wolfowitz NPMLE enjoys several important advantages over these finite mixture models. Because it is formulated as a discretized convex optimization problem, it is automatically assured to produce a unique solution with both the location and associated mass of the mixture components determined by the optimization of the likelihood. Positivity of the mass associated with mixture components also ensures a strong form of parsimony as we shall see, so the mixture distribution is encoded by a relatively simple discrete mixing distribution. In the next section we will describe our implementation of the Kiefer-Wolfowitz NPMLE, further details are provided in Koenker and Mizera (2014b). As a practical matter the optimization requires an algorithmic approach capable of dealing with a quite general class of additively separable likelihoods optimized subject to both linear equality and inequality constraints. For this purpose we have found the Mosek environment of Andersen (2010), and the associated R interface **Rmosek** of Friberg (2012) to be highly efficient and reliable. Koenker and Mizera (2014a) provide a broader survey of convex optimization methods for the R environment, including a brief mention of some basic **REBayes** functionality and further details regarding the general capabilities of Mosek. Installation of Mosek and **Rmosek** are described in detail in the Readme file in the "inst" directory of the **REBayes** package. Various options controlling Mosek optimizing behavior can be passed via the **REBayes** fitting functions. Among these `rtol` and `verb` that control the convergence tolerance and the verbosity of the optimization printed output are most frequently useful. While we have endeavored to choose sensible default values for these and other parameters some experimentation may be required in unusual cases.

2. Computation of the Kiefer-Wolfowitz NPMLE

It is easy to see that the primal problem

$$\min_{F\in\mathcal{F}}\{-\sum_{i=1}^{n}\log g(x_i) \mid g(x_i) = \int \varphi(x_i,\theta)dF(\theta),\ i = 1,...,n\}, \tag{2}$$

where $\mathcal{F}$ denotes the set of all mixing distributions, is a convex program. We seek to minimize a strictly convex objective function subject to linear equality constraints over the convex set, $\mathcal{F}$. The dual formulation of the problem is also illuminating.

Theorem 1. *(Koenker and Mizera (2014b)) The solution, $\hat{F}$, of* (2) *exists, and is an atomic probability measure, with not more than n atoms. The locations, $\hat{\mu}_j$, and the masses, $\hat{f}_j$, at these locations can be found via the following dual characterization: the solution, $\hat{\nu}$, of*

$$\max\{\sum_{i=1}^{n}\log\nu_i \mid \sum_{i=1}^{n}\nu_i\varphi(Y_i,\mu) \le n \text{ for all } \mu\} \tag{3}$$

satisfies the extremal equations (n equations in less than n variables)

$$\sum_j \varphi(Y_i, \hat{\mu}_j)\hat{f}_j = \frac{1}{\hat{\nu}_i}, \tag{4}$$

and $\hat{\mu}_j$ are exactly those μ where the dual constraint is active—that is, the constraint function in (3) *is equal to n.*

The dual formulation reduces the objective function to a simple finite dimensional sum, albeit now with an infinite dimensional constraint. The upper bound of n on the number of atoms, established under slightly stronger conditions by Lindsay (1983), encourages us in the quest for a discrete formulation. We should hasten to add that we have no assurances about where these atoms occur, in particular it is clear already from an example in Laird (1978) that they need not occur at the observed x_i. Laird (1978) proposed using the EM algorithm to solve a discretization of the primal problem (2) and subsequent authors, notably Heckman and Singer (1984) and Jiang and Zhang (2009), have followed her lead. However, as has been frequently observed, EM can be quite lethargic in its pursuit of the optimum. Koenker and Mizera (2014b) describe some comparisons of a fixed point EM algorithm with the interior point method implemented in Mosek. For a relatively small Gaussian location mixture problem with $n = 200$ and a grid of 300 points for the mixing distribution for μ, the interior point method produced a very precise solution in about 1 second and 15 iterations, while after 10 minutes and 100,000 iterations the EM algorithm was still struggling to obtain the same accuracy as the interior point solution.

In our discrete formulation we consider a fixed grid, $\{u_1, ..., u_m\}$, of potential support points for the mixing distribution, F. Typically, m is a few hundred, and the grid is equally spaced, but this can be easily adapted to particular applications. We denote by A an n by m matrix, with the elements $\varphi(Y_i, u_j)$ in the i-th row and j-th column. The discrete version of the primal problem is then,

$$\min_{f \in \mathbb{R}^m} \{ -\sum_{i=1}^n \log(g_i) \mid Af = g,\ f \in \mathcal{S}\},$$

where $\mathcal{S}$ denotes the unit simplex in $\mathbb{R}^m$, i.e., $\mathcal{S} = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geq 0\}$. So f_j denotes the estimated mixing density estimate $\hat{f}$ evaluated at the grid point u_j, and g_i denotes the estimated mixture density estimate, $\hat{g}$, evaluated at Y_i. In our experience it is somewhat more efficient to solve the corresponding dual problem,

$$\max_{\nu \in \mathbb{R}^n} \{ \sum_{i=1}^n \log \nu_i \mid A^\top \nu \leq n 1_m,\ \nu \geq 0\},$$

and subsequently recover the primal solution. In the **REBayes** package we have implemented this dual solution method for a wide variety of mixture problems that we will describe in subsequent sections. It is frequently convenient to consider weighted MLE formulations so **REBayes** fitting functions make some provision for weights.

The implementation relies heavily on the Mosek optimization software of Andersen (2010) and its R interface package **Rmosek**, Friberg (2012).

3. Gaussian mixture models

Gaussian mixture models are a natural point of departure for application of the foregoing methods. We will begin by describing usage in the simplest Gaussian sequence models. Some connections to multiple testing are described in the following subsection. Gaussian scale mixtures are then considered, followed by some brief remarks on Gaussian longitudinal models where heterogeneity in both location and scale comes into play. The section concludes with a cautionary parable concerning Gaussian location-scale mixtures in non-longitudinal settings.

3.1. **Needles in haystacks.**

To illustrate our methods in the simplest possible setting, consider the simulation framework of Johnstone and Silverman (2004): we have $X_i \sim \mathcal{N}(\mu_i, 1),\ i = 1, ..., n$, with s of the $\mu_i = \mu_0 \neq 0$ and the remainder, $\mu_i = 0$. When s is reasonably large relative to n and μ_0 is well separated from zero, then it should be easy to distinguish the two mass points of the mixture. Suppose we take $n = 1000$ and $s = 100$ with $\mu_0 = 2$ then the mixture density looks like that illustrated in in the left panel of Figure 1. In the middle panel of the figure we plot the NPMLE estimate of the mixing "density," which puts most of the mass near zero, and the remainder at a value slightly greater than two. The reader is encouraged to repeat this exercise to gauge the reliability of the NPMLE procedure with the R code reproduced below. Finally, in the right panel we illustrate the Bayes rule for predicting μ_i given observations at various values between -5 and +6. It may be noted that not only are observations below zero shrunken aggressively toward zero, but observations above two are also shrunken toward the estimated prior mass point near two. Observations between zero and two are, given the estimated mixing distribution, more ambiguous and the Bayes rule must account for both mass points in computing its conditional expectation.

```
R> # A simple Gaussian mixture model
R> par(mfrow = c(1,3))
R> x <- seq(-5, 6, by = 0.05)
R> plot(x, 0.9 * dnorm(x,0) + 0.1 * dnorm(x,2), type = "l",
+      xlab = "x", ylab = expression(g(x)), main = "")
R> y <- rep(c(0,2), times = c(900,100)) + rnorm(1000)
R> z <- GLmix(y)
R> plot(z, xlab = expression(mu), ylab = expression(f(mu)), main = "")
R> plot(x, predict(z,x), type = "l", ylab = expression(delta(x)))
```

The Tweedie shrinkage strategy depicted in Figure 1 is effective not only in shrinking the observations with $\mu_i = 0$ toward zero, but also in shrinking the non-null $\mu_i = 2$ observations toward two. This helps to explain the good performance of the NPMLE

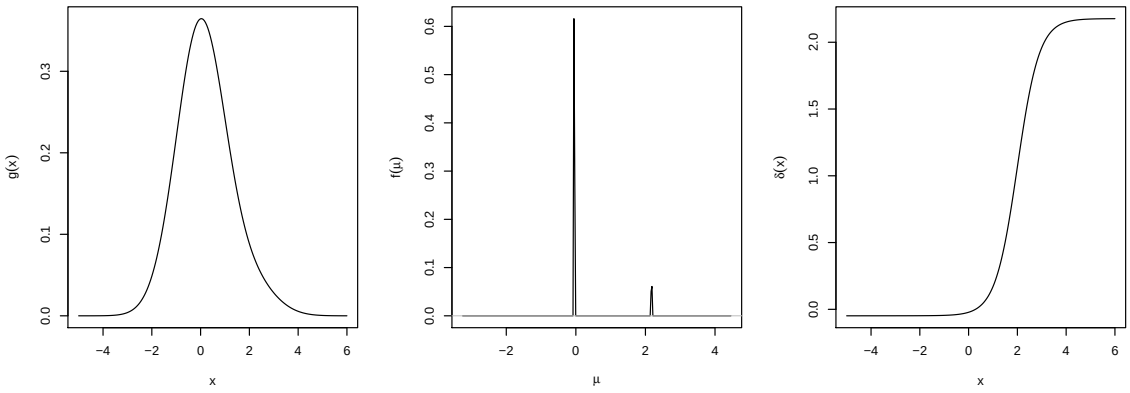

FIGURE 1. Kiefer Wolfowitz Estimation of a Gaussian Location Mixture: The left panel is the (unknown) two component mixture density, the middle panel is the estimated NPMLE mixing density and the right panel is the estimated Bayes rule for predicting $\hat{\mu} = \delta(x)$ based on seeing an observation x.

described in Koenker (2014) relative to the thresholding and parametric empirical Bayes procedures of Johnstone and Silverman (2004), Martin and Walker (2014) and Castillo and van der Vaart (2012). These competitors are quite good at shrinking the null observations toward zero, unlike the NPMLE they *know* that there is mass at zero, but they tend to leave the non-null observations alone and this tends to inflate their mean squared error. This observation raises the natural question how would the NPMLE do when the non-null observations came from a more diffuse distribution?

In Figure 2 we illustrate similar performance for a Gaussian location mixture in which 200 of the 1000 observations have μ_i's drawn from a $\mathcal{N}(2, 1)$ distribution. The true mixture density looks quite similar to the prior example, but the NPMLE now identifies three distinct mass points, one large one near zero, a smaller one near two and a very small mass point at about 4.5. The Bayes rule is still quite sure that negative x_i should be pulled toward zero, and observations near two are nudged toward two. But despite its small mass the upper mass point of the estimated mixing distribution exerts a substantial effect. Only when we see extremely large observations bigger than 4.5 are they pulled back toward this largest mass point. This example is considerably more challenging than the previous one, but nevertheless the empirical Tweedie formula produced by the NPMLE provides a reasonable approach.

```
R> # Another simple Gaussian mixture model
R> par(mfrow = c(1,3))
R> x <- seq(-5, 7, by = 0.05)
R> plot(x, 0.8 * dnorm(x,0) + 0.2 * dnorm(x,2,sqrt(2)), type = "l",
```

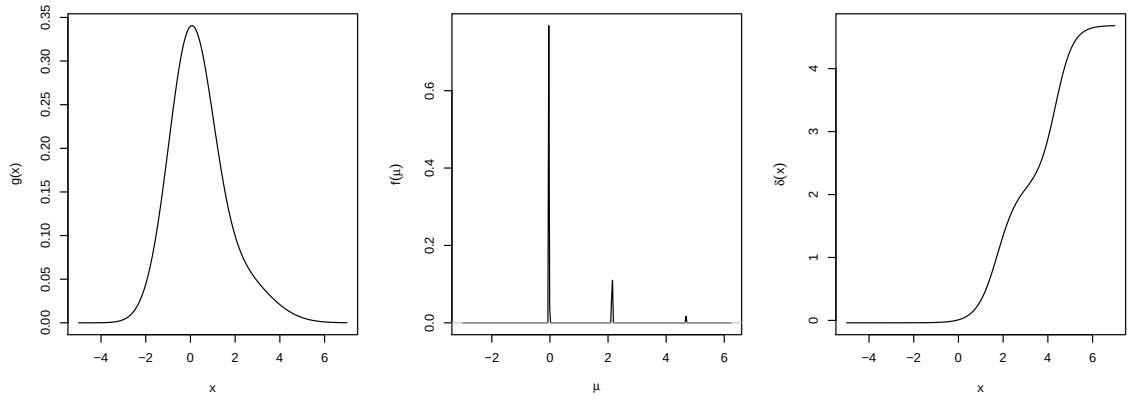

FIGURE 2. Kiefer Wolfowitz Estimation of a Gaussian Location Mixture: The left panel is the (unknown) mixture density, the middle panel is the estimated NPMLE mixing density and the right panel is the estimated Bayes rule for predicting $\hat{\mu} = \delta(x)$ based on seeing an observation x.

```
+      xlab = "x", ylab = "g(x)", main = "")
R> y <- c(rep(0,800), rnorm(200, 2)) + rnorm(1000)
R> z <- GLmix(y)
R> plot(z, xlab = expression(mu), ylab = expression(f(mu)), main = "")
R> plot(x, predict(z,x), type = "l", ylab = expression(delta(x)))
```

In the foregoing examples we have employed the posterior mean as a prediction assuming implicitly that we faced quadratic loss, however it is straightforward to adopt other loss functions and provide alternative predictions. For `GLmix` fitted objects `predict.GLmix` allows the user to specify posterior median or posterior modal prediction by setting the argument `Loss` equal to 1 or 0, respectively. If `Loss` is specified as a value τ between zero and one, predictions return the posterior τth quantile. Empirical Bayes posterior quantile prediction is considered in Mukherjee et al. (2016) although they restrict attention to linear shrinkage rules. They reference a large literature on the so-called "newsvendor" problem going back to Edgeworth (1888) that motivates the quantile loss function. Analogous `predict` functions are also available for Poisson, i.e. `Pmix`, and binomial, i.e. `Bmix`, fitted objects.

3.2. **Gaussian mixtures and multiple testing.** Robbins (1951) introduced compound decision making with the following (deceptively) simple problem. Suppose we observe,

$$Y_i = \theta_i + u_i, \quad i = 1, \cdots, n, \tag{5}$$

with $\{u_i\}$ iid standard Gaussian, and we know that the θ_i take values ± 1. The objective is to estimate the n-vector, $\theta \in \{-1, 1\}^n$ subject to ℓ_1 loss,

$$L(\hat{\theta}, \theta) = n^{-1} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|.$$

He observes that when $n = 1$ the least favorable version of the problem occurs when we assume that the θ_i's are drawn as independent Bernoulli's with probability $p = 1/2$ that $\theta_i = \pm 1$, and then he proceeds to show that this remains true for the general "compound decision" problem with $n \geq 1$. The minimax decision rule is thus,

$$\delta_{1/2}(y) = \operatorname{sgn}(y)$$

and yields constant risk,

$$R(\delta_{1/2}, \theta) = \mathbb{E}L(\delta_{1/2}(Y), \theta) = \Phi(-1) \approx 0.1586,$$

irrespective of p. And yet, something feels wrong with this procedure. If we saw mostly positive Y_i's wouldn't we begin to think that $p \neq 1/2$? Why are we so attached to the worst case scenario? Exploiting the common structure of the n problems, Robbins suggests *estimating* p by $\hat{p} = (\bar{y} + 1)/2$. Given this method of moments estimate of p, he suggests plugging it into the decision rule,

$$\delta_p(y) = \operatorname{sgn}(y - 1/2 \log((1 - p)/p)),$$

a procedure that follows immediately from the requirement that,

$$P(\theta = 1|x, p) = \frac{p\varphi(x - 1)}{p\varphi(x - 1) + (1 - p)\varphi(x + 1)},$$

exceeds one half, that is, that the posterior median of θ be 1. This prototype empirical Bayes procedure sacrifices a little in performance when p is really near $1/2$, but achieves substantial gains in performance when p differs substantially from $1/2$. Of course, when n is large, $\hat{p} \to p$, so we have a form of asymptotic optimality.

The link to the multiple testing literature for the Robbins problem is immediately clear since estimation of $\theta \in \{-1, 1\}^n$ is essentially a testing problem in which we have weighed false discovery and false non-discovery equally. If we treat $\theta = -1$ as the null hypothesis and $\theta = 1$ as the alternative, a p-value procedure based on $T_i = 1 - \Phi(X_i + 1)$ with cutoff $\Phi(-1)$ the decision rule,

$$\delta_p(T) = \operatorname{sgn}(\Phi(-1) - T)$$

is equivalent to the minimax rule, $\delta(x) = \operatorname{sgn}(x)$. If, instead, we would like to fix the marginal false discovery rate (mFDR) at some level and optimize marginal false nondiscovery rate (mFNR) a modified p-value cutoff can be constructed, and this would be equivalent to replacing our symmetric ℓ_1 loss for the estimation/classification problem by an asymmetric linear loss.

A p-value testing procedure that is equivalent to the empirical Bayes rule estimator described earlier for the Robbins problem can also be constructed. Under the null that $X_i \sim \mathcal{N}(-1,1)$, $T_i = 1 - \Phi(X_i + 1) \sim U[0,1]$, while if $X_i \sim \mathcal{N}(1,1)$,

$$\mathbb{P}(T_i < u) = \mathbb{P}(X_i + 1 > \Phi^{-1}(1-u)) = 1 - \Phi(\Phi^{-1}(1-u) - 2).$$

Thus, under the null, the density of T is $f_0(t) \equiv 1$, and under the alternative,

$$f_1(t) = \varphi(\Phi^{-1}(1-t) - 2)/\varphi(\Phi^{-1}(1-t)),$$

and the posterior probability of $\theta_i = 1$ given t_i and assuming for the moment that the unconditional probability, $p = \mathbb{P}(\theta_i = 1)$ is known, is given by,

$$\mathbb{P}(\theta = 1|t,p) = \frac{p f_1(t)}{p f_1(t) + (1-p) f_0(t)}.$$

Under symmetric loss we were led to the posterior median so $\hat{\theta}_i = 1$ if $\mathbb{P}(\theta_i = 1|T_i, p) > 1/2$, which is equivalent to the p-value rule,

$$T_i < 1 - \Phi(1 + 0.5 \log((1-p)/p)).$$

Again, we are led back to the problem of estimating p. In these two point mixture problems ℓ_1 loss is equivalent to $0-1$ loss since the median and the mode are identical.

In Gu and Koenker (2016b) we explore some extensions of this simple setting to several other multiple testing problems. We first consider a grouped setting in which we have,

$$Y_{ij} = \theta_{ij} + u_{ij}, \quad i = 1, \cdots, n, \;\; j = 1, \cdots, m,$$

with $\{u_{ij}\}$ iid standard Gaussian as before, and $\theta_{ij} = 1$ with probability p_i and $\theta_{ij} = -1$ with probability $1 - p_i$, and independent over $j = 1, \cdots, m$. In this framework we can consider "group specific" p_i that vary within the full sample yielding a nonparametric mixture problem. In the multiple testing context this grouped model has been considered by Efron (2008) and Sun and Cai (2007) among others. This formulation leads us back to the Kiefer and Wolfowitz NPMLE. We also consider abandoning the rather implausible assumption that we know the support points of the θ's. This allows us to consider multiple testing rules for more realistic settings with both composite null and alternatives. Comparing performance of these rules with the empirical characteristic function procedures of Sun and McLain (2012) shows very favorable performance.

3.3. **Gaussian scale mixtures.** Gaussian scale mixtures can be estimated in much the same way that we have described for location mixtures. Suppose we now observe an unbalanced panel,

$$y_{it} = \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n$$

with $u_{it} \sim \mathcal{N}(0,1)$. Sufficiency reduces the sample to n observations on $S_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it}^2$, and thus S_i has a gamma distribution with shape parameter, $r_i = m_i/2$,

and scale parameter θ_i/r_i, i.e.,

$$\gamma(S_i|r_i, \theta_i/r_i) = \frac{1}{\Gamma(r_i)(\theta_i/r_i)^{r_i}} S_i^{r_i-1} \exp\{-S_i r_i/\theta_i\},$$

and the marginal density of S_i when the θ_i are iid from F is

$$g(S_i) = \int \gamma(S_i|r_i, \theta/r_i) dF(\theta).$$

Estimation of F proceeds as in the location mixture setting except that now the matrix A has typical element $\gamma(S_i|\theta_j)$ with θ_j's constituting a fine grid covering the support of the sample S_i's. This can be implemented in **REBayes** with the function `GVmix`, which may be seen as a general procedure for scale mixtures of χ^2. A yet more general procedure for scale mixtures of gamma random variables is provided by the function `gammamix`. Robbins (1982) contains an early discussion of parametric empirical Bayes methods for scale mixture of Gaussians, van der Vaart (1996) considers the semiparametric efficiency for the same model with an additional unknown location parameter. The scale mixture of Gaussians is also a crucial building block for the more general location-scale mixture we have considered in the longitudinal setting.

An application of the Gaussian scale mixture procedure is described in Koenker (2013) where a simple bivariate linear regression model,

$$Y_i = \beta_0 + x_i\beta_1 + U_i$$

is considered. The u_i are assumed to be generated iidly from a scale mixture of Gaussians, so U_i^2 have mixture density,

$$g(v) = \int_0^\infty \gamma(v|\theta) dF(\theta)$$

where $\theta = \sigma^2$, and γ is the $\chi^2(1)$ density with free scale parameter θ,

$$\gamma(v|\theta) = \frac{1}{\Gamma(1/2)\sqrt{2\theta}} v^{-1/2} \exp(-v/(2\theta))$$

Given a preliminary estimate of the β parameters we can estimate the mixing distribution F based on the sample of $\hat{u}_i^2$'s, and this in turn can be used to estimate the score function,

$$\hat{\psi}(u) = (-\log \hat{g}(u))' = \frac{\int u\varphi(u/\sigma)/\sigma^3 d\hat{F}(\sigma)}{\int \varphi(u/\sigma)/\sigma d\hat{F}(\sigma)},$$

used to reestimate β. Iterating this procedure may be seen as our first encounter with Kiefer-Wolfowitz profile likelihood and can be shown to achieve an asymptotically fully efficient regression estimator for the class linear models with iid scale mixture of Gaussian errors.

3.4. **Longitudinal Gaussian models.** Longitudinal data allow us to explore heterogeneity in both location and scale for Gaussian Models. Let's begin by considering the model,

$$y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n$$

with $u_{it} \sim \mathcal{N}(0,1)$. We will provisionally assume that $\alpha_i \sim F_\alpha$ and $\theta_i \sim F_\theta$ are independent. Again, we have sufficient statistics:

$$\bar{y}_i | \alpha_i, \theta_i \sim \mathcal{N}(\alpha_i, \theta_i / m_i)$$

and

$$S_i | r_i, \theta_i \sim \gamma(S_i | r_i, \theta_i / r_i),$$

where $r_i = (m_i - 1)/2$, $S_i = (m_i - 1)^{-1} \sum_{t=1}^{m_i} (y_{it} - \bar{y}_i)^2$, and the log likelihood becomes,

$$\ell(F_\alpha, F_\theta | y) = K(y) + \sum_{i=1}^{n} \log \int \int \gamma(S_i | r_i, \theta / r_i) \sqrt{m_i} \phi(\sqrt{m_i}(\bar{y}_i - \alpha_i)/\sqrt{\theta}) / \sqrt{\theta} dF_\alpha(\alpha) dF_\theta(\theta).$$

Since the scale component of the log likelihood is additively separable from the location component, we can solve for $\hat{F}_\theta$ in a preliminary step, as in the previous subsection, and then solve for the $\hat{F}_\alpha$ distribution. In fact, under the independent prior assumption, we can re-express the Gaussian component of the likelihood as Student-t and thereby eliminate the dependence on θ in the Kiefer-Wolfowitz problem for estimating F_α. An implementation is available in the function `WTLVmix` of **REBayes**. Gu and Koenker (2016a) describe an application to predicting baseball batting averages in which following Brown (2008) averages are transformed to normality, and the θ's reflect either under or over dispersion relative to the standard binomial model. Again, profile likelihood is used to explore covariate effects embedded in this model of heterogeneity. In particular we estimate an age profile for batting prowess as a quadratic effect that peaks at age 27. Comparing predictive performance for this model we find that the independent prior NPMLE performs considerably better than its more naive competitors.

It is also possible to relax the independence assumption on the location and scale effects. In Gu and Koenker (2016c) we use longitudinal data from the Panel Study on Income Dynamics (PSID) to study models of income dynamics with an arbitrary joint distribution of location and scale heterogeneity. In these models we estimate an AR(1) effect by profile likelihood. The implementation for these models uses the function `WGLVmix` and requires a bivariate gridding strategy for the mixing distribution. We find that there is a distinct *negative* dependence between the α (location) and θ scale effects indicating that low "ability" individuals also tend to be high income variability people. Accounting for heterogeneity in scale has an acute effect on the estimation of the AR(1) effect reducing what is often regarded as a unit root effect to a rather mild $\rho \approx 0.5$ effect. The Bayesian formulation of these models offers the significant additional advantage that it affords a convenient environment for forecasting future income trajectories.

3.5. **The parable of the crabs: A cautionary tale.** The first formal estimation of a mixture model in statistics seems to have been Karl Pearson's 1894 analysis of the ratio of "forehead breadth" to body length of 1000 crabs sampled from the Bay of Naples by the prominent biologist W.F.R. Weldon. Pearson estimated a two component normal mixture model by the method of moments, a truly heroic computational effort given the technology of the time. He allowed his two normal components to have distinct means and variances so together with the relative weight of the two components he had five parameters. Modern (EM) methods are capable of producing similar results, although they are quite sensitive to the choice of initial values. It is thus tempting to ask: Can the Kiefer-Wolfowitz NPMLE offer any further insight into such problems.

The short answer, unfortunately, is no. The immediate difficulty one encounters is that in contrast to our baseball application, or the income dynamics model, there is no longitudinal dimension to the data. All we have is a single sample, a basket of crabs. If we were to assume that we had simply a location mixture, or simply a scale mixture, it would be easy to estimate the mixing distribution with the NPMLE. But if we try to emulate Pearson and estimate a nonparametric location and scale mixture we are headed for a Dirac catastrophe. For each observation, we are entitled to assign a distinct mixing value $\mu_i = x_i$, corresponding to these μ_i we are also entitled to assign a $\sigma_i = 0$, and to each of these points $(\mu_i, \sigma_i) = (x_i, 0) \quad i = 1, ..., n$ we can assign mass $1/n$. The likelihood explodes and our mixing distribution has collapsed to the familiar empirical distribution.

The moral of this fable is this: Sorting a basket of crabs is tougher than it might seem. Kiefer and Wolfowitz knew a thing or two about this; the final paragraph of their 1956 paper points the fundamental difficulty of the location-scale Gaussian mixture model, and earlier they had already pointed out that the empirical distribution function was, itself, an MLE, of a sort. Teicher (1967) provides a more formal discussion.

4. Mixture models for counts

The Kiefer-Wolfowitz NPMLE can also be useful in analyzing discrete random variables such as count data where unobserved heterogeneity also arises naturally. Many applications involve count data as an object of interest: the number of patents across firms or industries, the number of hospital visits among patients, or the number of claims in insurance applications. The typical model for analyzing such data is Poisson regression. Often, however, even after accounting for observed covariates, there remains some over or under-dispersion in the data, indicating a need to introduce additional unobserved heterogeneity into the Poisson model. When handling this unobserved heterogeneity, a parametric model is typically imposed on the heterogeneity distribution in the literature. We illustrate below how the NPMLE provides a more flexible nonparametric approach for handling unobserved heterogeneity in Poisson

models based on a model for the number of claims for a group life insurance policy. We also point out some advantages of NPMLE over the linear credibility estimators that are widely used for experience rating of insurance contracts. For a detailed discussion of credibility theory in actuarial science see Bühlmann and Gisler (2005).

Our data, first analyzed in Norberg (1989), consists of a portfolio of Norwegian workmen's group life insurance policies. The original 1125 contracts are aggregated into 72 occupational categories and consists of the total number of deaths X_i (number of claims) and the total number of years exposed to risk E_i for $i = 1, \ldots, 72$ for each occupational group. This data is available from **REBayes** as `data("Norberg")`. Data on the 1125 individual contracts is only partially documented in Norberg (1989), so we resort to the 72 occupation group data that is documented in Haastrup (2000) and is provided in the dataset `Norberg` in the **REBayes** package. Figure 3 illustrates the histogram of the ratio of X_i and E_i.

Following Norberg (1989), we assume a Poisson model for X_i, so conditional on iid $\theta_i \sim G$,

$$X_i \sim \text{Poisson}(\theta_i E_i)$$

Here E_i is renormalized by a factor of 344 as in Haastrup (2000), and can be interpreted as the *à priori* expected number of claims in the period of contract. The multiplicative unobserved (occupational) specific factor θ_i then accounts for the fact that various occupations have different risk profiles that are not observed, but can be indirectly inferred by the observed number of claims. In classical credibility theory this leads to insurance premiums tailored to individual risk profiles based on the observed claims and exposures that have occurred. Rather than assuming that the distribution G belongs to a particular parametric class as in Norberg (1989) and Haastrup (2000), we adapt the Kiefer-Wolfowitz NPMLE to this task. Haastrup (2000) also conducts a nonparametric Bayesian analysis with a Dirichlet Process prior using Gamma distribution as a base, our methods serve as a nonparametric empirical Bayes contrast to his results.

```
R> # Parametric Gamma vs Poisson mixture models for insurance claims
R> data("Norberg")
R> E <- Norberg$Exposure/344
R> X <- Norberg$Death
R> hist(X/E, 90, freq = TRUE, xlab = "X/E", main = "", ylab = "Frequency")

R> # Maximum likelihood estimation of the Gamma model
R> logL<-  function(par, x, e){
+      f <- choose(x + par[1] - 1, x) *
+          (par[2]/(e + par[2]))^par[1] * (e/(e+par[2]))^x
+      -sum(log(f))
+      }
R> z <- optim(c(5,5), logL, x = X, e = E, hessian = TRUE)
R> sez <- sqrt(diag(solve(z$hessian)))
```

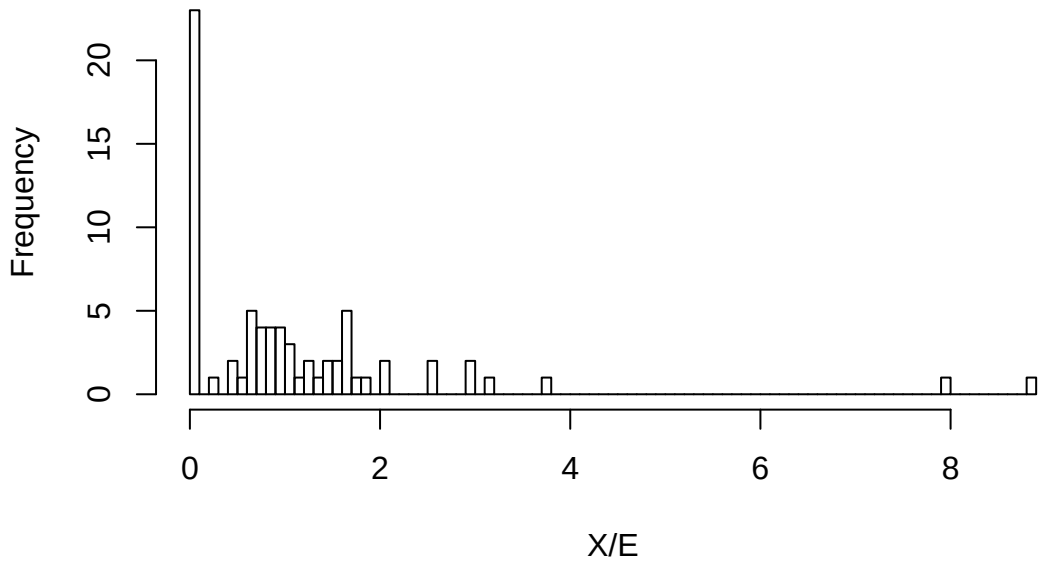

FIGURE 3. Histogram of Claims per Exposure for 72 occupation groups.

```
R> z <- z$par
R> # Estimation of the Poisson mixture model
R> f = Pmix(X, v = 1000, exposure = E, rtol = 1e-10)
R> # Now plot the comparison
R> par(mfrow=c(1,2))
R> plot(f$x,f$y/sum(f$y), type="l", xlab = expression(theta),
+      ylab = expression(f(theta)), ylim = c(0,1))
R> lines(f$x, dgamma(f$x, shape = z[1], rate = z[2]), col = 2)
R> plot(f$x,(f$y/sum(f$y))^(1/3), type="l", xlab = expression(theta),
+  ylab = expression(f(theta)^{1/3}), ylim = c(0,1))
```

Figure 4 contrasts the NPMLE estimator with the corresponding parametric empirical Bayes estimates assuming that G follows a Gamma distribution. The main reason for adopting the Gamma mixing distribution is analytical convenience. With

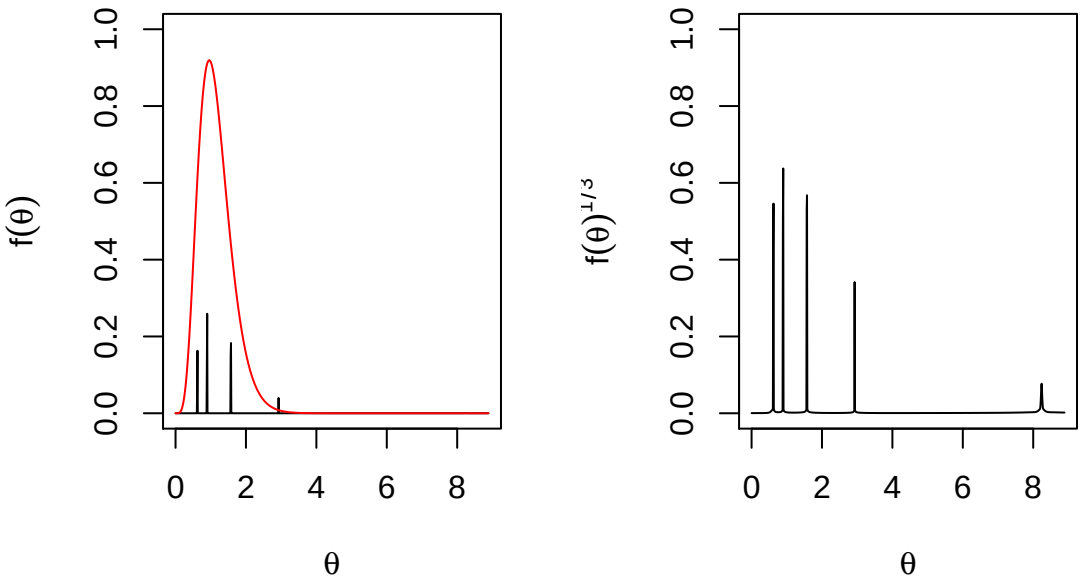

FIGURE 4. Estimated mixing distribution G for θ for the group insurance data. The left panel depicts to the Kiefer-Wolfowitz NPMLE estimator for G with 1000 grid points. The right panel depicts the cube root of the mass associated with support points around 8. The smooth curve superimposed in the left panel corresponds to the parametric maximum likelihood estimate of the mixing density assuming G follows a Gamma distribution.

$\theta_i \sim \text{Gamma}(\alpha, \beta)$, the marginal distribution of X_i follows a negative binomial distribution

$$\begin{aligned} g(X_i|E_i) &= \int \frac{(\theta E_i)^{X_i} \exp(-\theta E_i)}{X_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) d\theta \\ &= \binom{X_i + \alpha - 1}{X_i} \left(\frac{\beta}{E_i + \beta}\right)^\alpha \left(\frac{E_i}{E_i + \beta}\right)^{X_i} \end{aligned}$$

The maximum likelihood estimates are $\alpha = 6.02$ and $\beta = 5.25$. The credibility estimator of the risk per exposure leads to

$$\hat{\theta}_i = \delta_i X_i / E_i + (1 - \delta_i)\mathbb{E}(\theta)$$

with $\mathbb{E}(\theta) = \int \theta dG(\theta)$ and $\delta_i = \frac{\mathbb{V}(\theta)}{\mathbb{V}(\theta)+\mathbb{E}(\theta)/E_i}$. Under the parametric assumption that G is Gamma(α, β), it is easy to see that $\mathbb{E}(\theta) = \alpha/\beta$ and $V(\theta) = \alpha/\beta^2$, hence $\hat{\theta}_i = \frac{X_i+\alpha}{E_i+\beta}$, which is nothing but $\mathbb{E}(\theta|X_i, E_i)$ from the Poisson-Gamma mixture model. The Gamma assumption on G leads to a convenient analytical form for the credibility estimator, but since it may produce a rather unrealistic estimator of the underlying mixing distribution the premium calculation of the parametric credibility estimator may be questionable.

In Figure 4 we see that although the majority of the support points seem to situated under the "umbrella" of the Gamma density, the Gamma distribution fails to detect the two outliers (Group 13 and Group 53, with X/E ratios equal to 8.89 and 7.98 respectively) that account for the remote mass point around 8. In the right panel of Figure 4, we plot the cube root of the estimated mixing distribution and magnify the very small yet important mass point around 8. One may argue that these two occupational groups could be viewed as outliers and hence should not be allowed to influence our views about the distribution of the unobserved risk factor θ. However, an insurance company would ignore them at its peril.

For our general mixing distribution NPMLE estimator $\hat{G}$, the credibility estimator then becomes,

$$\hat{\mu} = \mathbb{E}(\theta|X_i, E_i) = \frac{\int \theta \frac{(\theta E_i)^{X_i} \exp(-\theta E_i)}{X_i!} d\hat{G}(\theta)}{\int \frac{(\theta E_i)^{X_i} \exp(-\theta E_i)}{X_i!} d\hat{G}(\theta)}$$

Figure 5 contrasts the $\hat{\theta}_i$ based on the parametric Poisson-Gamma empirical Bayes estimator and those based on the nonparametric Poisson mixture model. We can see that for most of the occupational groups, the two estimators agree closely except for the two most extreme case (Group 13 and 53), that have the largest X/E ratio. The nonparametric empirical Bayes procedure, relying on the mass point associated with a much larger support point, produces substantially larger credibility estimator for these "riskier" groups, thereby justifying a higher premium. The `Pmix` function produces the Bayes rule automatically with an output denoted as `dy`, as illustrated in the code below.

```
R> # Bayes rules for insurance application
R> PBrule <- (X + z[1])/(E + z[2])
R> NPBrule <- f$dy
R> plot(PBrule, NPBrule, cex = 0.5,
+       xlab = "P-EBayes", ylab = "NP-EBayes")
R> abline(c(0,1))
```

5. Frailty models in survival analysis

The notion of frailty to describe unobserved heterogeneity of population risks has become a familiar feature of demographic analysis since its introduction in Vaupel

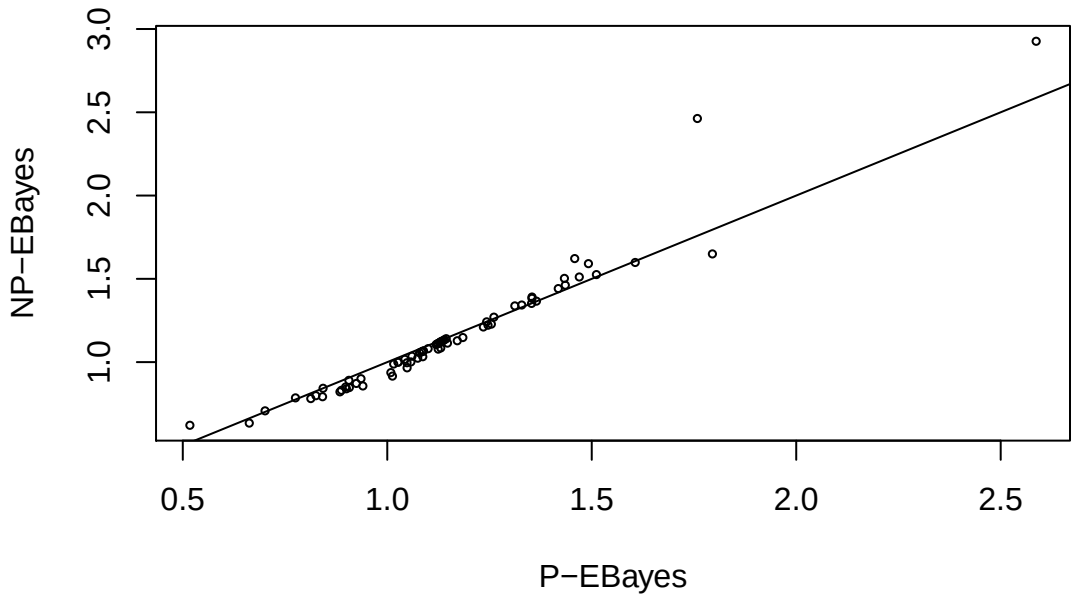

FIGURE 5. Comparison of the Parametric and the Nonparametric Empirical Bayes estimator of θ_i for 72 occupation groups. As indicated by the 45 degree line there is good agreement between the parametric and nonparametric Bayes rules except for the two groups appearing in the upper right corner of the plot.

et al. (1979), and has gradually spread to other statistical domains. It is tempting to begin a survival analysis by specifying a simple parametric model for the survival distribution, say the Weibull, and then on further reflection decide that a more flexible approach is necessary. One way to introduce such flexibility is to consider mixtures of the original simple model, for example by letting the scale parameter of the Weibull be random. This sort of thinking leads to deeper concerns about the nature of randomness touched upon by Aalen et al. (2008). Do we really believe that individual subjects are assigned a scale parameter and then fated to draw a survival date from the corresponding Weibull? Or should we instead just regard the population survival distribution as adequately approximated by the scale mixture? In the absence of further information to distinguish subpopulations it is difficult to see how to untangle these two interpretations, and we will not try to pursue this. Instead, we will illustrate what can be done with our Kiefer-Wolfowitz apparatus in a reanalysis of the

influential Carey et al. (1992) experiments on medfly mortality. The primary objective of these experiments was to characterize the upper tail of the medfly mortality distribution, an endeavor that revealed several surprising biological features.

- Mortality rates *declined* at advanced ages, contrary to conventional biological wisdom that ageing was an inexorable process of physical decline,
- The survival distribution had an extremely heavy tail, contrary to the common view that each species had an explicit upper bound on survival propects,
- Gender cross-over in mortality rates gave males an advantage at early ages and females an advantage at advanced ages, reversing expectations from other species.

In the largest of the three experiments reported in Carey et al. (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 ..."

Data from this experiment is available from the **REBayes** with further details documented there.

All three of the principle conclusions of the study are illustrated in Figure 6. As specified in the code fragment below we compute daily death counts by age and gender, allowing us to plot raw mortality rates by gender. We then estimate the Weibull mixture model using gender specific Weibull shape parameters as described in Koenker and Gu (2013). As illustrated in the displayed code, given the estimated mixing distribution it is easy to compute the hazard functions of the corresponding mixture distributions.

```
R> data("flies")
R> attach(flies)
R> # Weibull hazard function
R> hweibull <- function(s,alpha,lambda, f){
+      Lambda<-outer((lambda*s)^(alpha),exp(f$x))
+      Surv <- exp(-Lambda) %*% f$y/sum(f$y)
+      A <- matrix(0, length(s), length(f$x))
+      for (i in 1:length(s)){
+          for (j in 1:length(f$x))
+              A[i,j] <- dweibull(s[i],shape=alpha,
+                   scale = lambda^(-1) * (exp(f$x[j]))^(-1/alpha))
+         }
+      g <- A %*% f$y
```

```
+      g/(sum(g)*Surv)
+  }
R> ahat <- c(2.793, 2.909) # Gender specific Weibull shape parameters
R> counts <- tapply(num,list(age,female),"sum")
R> cols <- c("black","grey")
R> labs <- c("Male","Female")
R> # Plot raw and estimated hazard functions by gender
R> for(g in 1:2){
+     gc <- counts[!is.na(counts[,g]),g]
+     freq <- gc/sum(gc)
+     day <- as.numeric(names(gc))
+     atrisk <- rev(cumsum(rev(gc)))
+     h <- rev(diff(rev(c(atrisk,0))))/atrisk
+     fW <- Weibullmix(day, m = 5000, alpha = ahat[g], weight = freq)
+     hW <- hweibull(day, alpha = ahat[g], lambda = 1, fW)
+     if(g == 1){
+         plot(day[1:100],hW[1:100],type="l", xlim = c(0,110),
+           ylim = c(0,.20), xlab = "Day", ylab = "Hazard")
+         points(day[1:100], h[1:100], cex = 0.7)
+         }
+     else{
+         lines(day[1:120],hW[1:120],col = cols[2])
+         points(day[1:100], h[1:100], cex = 0.7, col = cols[2])
+         }
+     legend("topleft", labs, lty = rep(1,2), lwd = 1.5, col=cols)
+     }
```

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. To investigate whether differences in initial cage density had a significant impact on mortality we can consider a model in which density enters 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 density. To estimate the density effect parameter, β, we simply evaluate the profiled likelihood on a grid of values on the interval $[-1, 1]$, yielding Figure 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

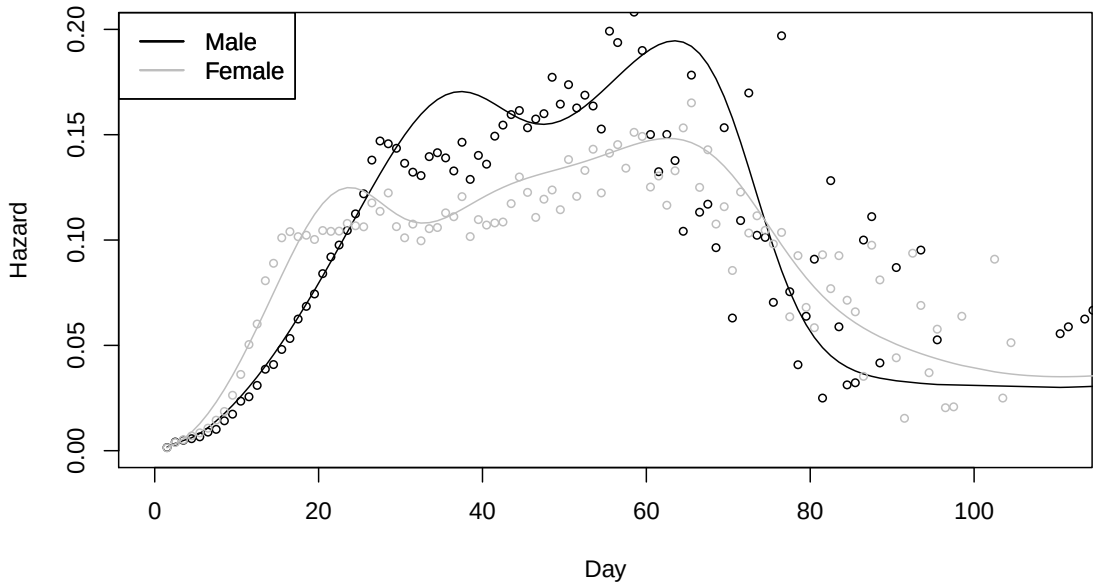

FIGURE 6. Raw and estimated mortality rates for Carey medflies by gender

regression models. Profile likelihood is not always so successful in models of this type, for a cautionary lesson involving estimation of the Weibull shape parameter see Koenker and Gu (2013).

```
R> # Profile likelihood estimation of initial cage density effect
R> counts <- tapply(num,list(age, begin),"sum")
R> freq <- c(counts)
R> day <- as.numeric(dimnames(counts)[[1]])
R> den <- as.numeric(dimnames(counts)[[2]])
R> day <- rep(day, 165)
R> den <- rep(den, each = 136)
R> s <- !is.na(freq)
R> day <- day[s]
R> den <- den[s]
R> freq <- freq[s]/sum(freq[s])
R> beta <- -10:10/10
R> logL <- beta
R> for(i in 1:length(beta)){
+     f <- Weibullmix(day, m = 500, alpha = 2.95,
+          lambda = exp(beta[i]*den), weight = freq)
```

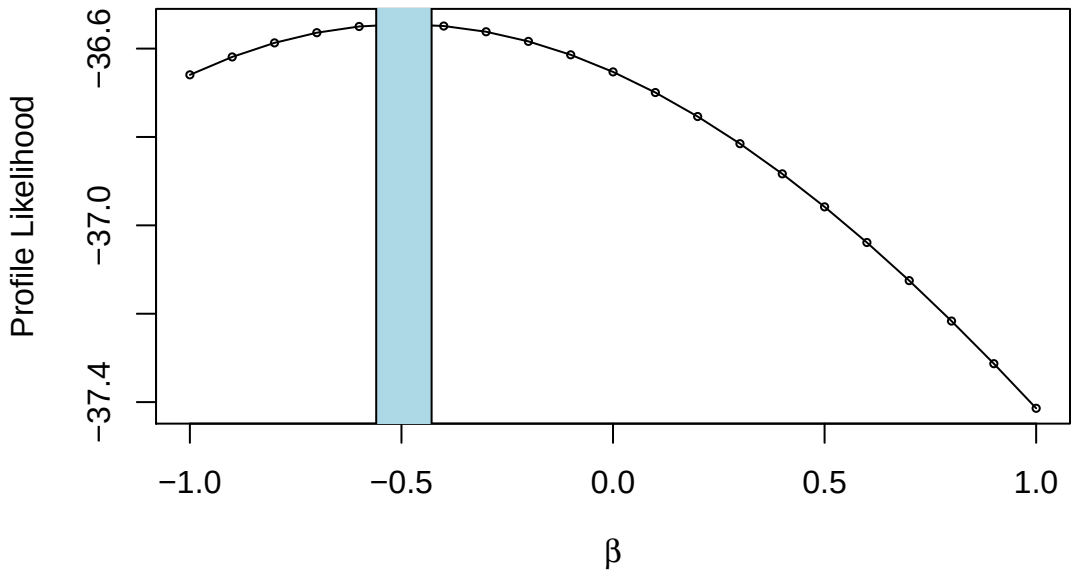

FIGURE 7. Initial Cage Density Effect in the Weibull Mixture Model: Profile Log Likelihood (in 1000's) for the cage density effect with 0.95 (Wilks) confidence interval in blue.

```
+     logL[i] <- f$logLik
+     }
R> plot(beta, logL/1000, cex = 0.5, xlab = expression(beta),
+      ylab = "Profile Likelihood")
R> lines(beta, logL/1000)
R> # Wilks interval for cage density effect
R> fsp <- splinefun(beta, max(logL) - logL - qchisq(.95,1)/2)
R> blo <- uniroot(fsp,c(-1,-.5))$root
R> bhi <- uniroot(fsp,c(-.5, 0))$root
R> polygon(c(blo,bhi,bhi,blo), c(-40,-40,-30,-30), col = "lightblue")
```

5.1. **MedLife: Fly-by-night insurance for Mediterranean fruit flies.** Imagine that you have been engaged by MedLife™ to design life insurance contracts for medflies of various ages. To keep things relatively simple, suppose that you are not allowed to discriminate on the basis of gender or other observable characteristics, like pupal size or initial cage density. How should we compute an actuarially fair premium for

a medfly of age T for a policy that pays 1, if the fly dies between T and $T + k$. We will resist speculating on who the beneficiaries of these policies might be or how double indemnity might be adjudicated. Instead, we will compare our nonparametric Weibull mixture approach with a more conventional parametric method that assumes gamma frailty for the Weibull model.

Let's begin by comparing hazard function estimates for the parametric and nonparametric specifications. When the frailty distribution is gamma, so,

$$h(z) = \frac{\nu^\eta}{\Gamma(\eta)} z^{\eta-1} e^{-\nu z},$$

it is convenient to restrict the mean frailty to be one, so $\nu = \eta$ and denote $\delta = 1/\eta$. Then for the Weibull base model with hazard function, $a(t) = (\alpha/\beta)(t/\beta)^{\alpha-1}$ and cumulative hazard, $A(t) = (t/\beta)^\alpha$, we can write the unconditional hazard and survival functions for the population as,

$$\lambda(t) = a(t)/(1 + \delta A(t))$$

and

$$S(t) = (1 + \delta A(t))^{-1/\delta}.$$

This yields the loglikelihood,

$$\ell(\alpha, \beta, \delta | t) = \sum_{i=1}^{n} \log a(t_i) - (1 + 1/\delta) \log(1 + \delta A(t_i)).$$

```
R> # Parametric Gamma frailty vs nonparametric Weibull mixture model
R> GammaFrailty <- function(pars, age, num, hazard = FALSE){
+      alpha <- pars[1]
+      beta <- pars[2]
+      delta <- pars[3]
+      a <- (alpha/beta) * (age/beta)^(alpha - 1)
+      A <- (age/beta)^alpha
+      if(hazard)
+              z <- a/(1 + delta * A)
+      else
+          z <- -sum(num * (log(a) - (1 + 1/delta)* log(1 + delta * A)))
+      z
+      }
R> pars <- c(5, 20, 1)
R> z <- optim(pars, GammaFrailty, age = age, num = num)
R> fitG <- z$par
R> fitW <- Weibullmix(day, m = 5000, alpha = 2.95, weights = freq)
R> s <- 1:110
R> day <- day[s]
R> hazard <- hazard[s]
```

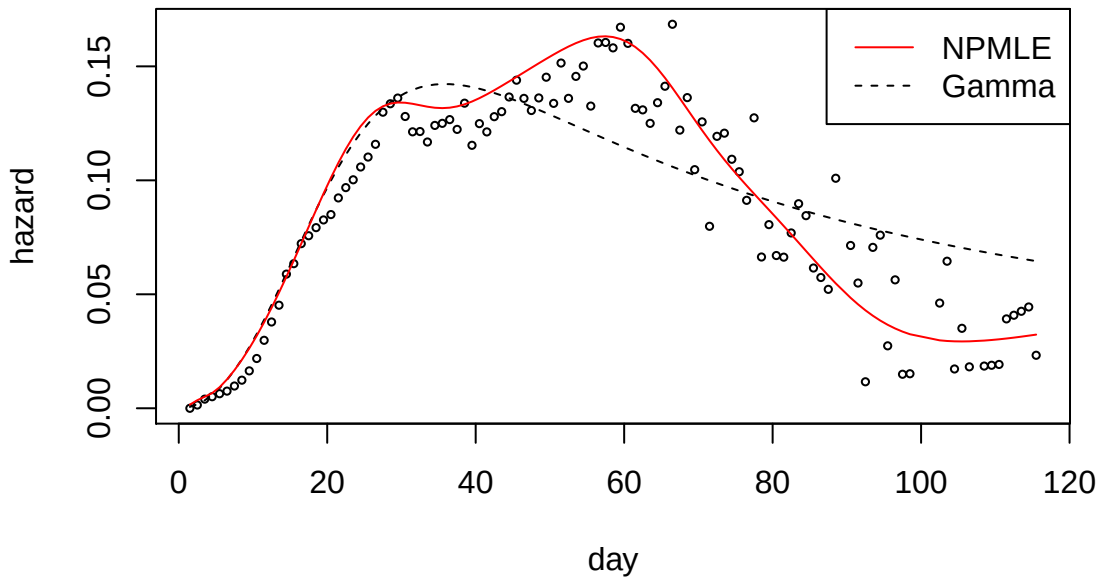

FIGURE 8. Parametric versus Nonparametric Estimates of Medfly Mortality Rates

```
R> plot(day, hazard, cex = 0.5)
R> lines(day, GammaFrailty(z$par,day,num, hazard = TRUE), lty = 2)
R> hW <- hweibull(day, alpha = 2.95, lambda = 1, fitW)
R> lines(day, hW, col = 2)
R> legend("topright", c("NPMLE", "Gamma"), col = 2:1, lty = 1:2)
```

Estimating the parametric gamma frailty model by maximum likelihood is straightforward as indicated in the code above, giving $(\hat{\alpha}, \hat{\beta}, \hat{\delta}) = (3.08, 21.12, 0.41)$, and yielding the hazard function shown in Figure 8. We superimpose the raw mortality rates and the hazard function based on our Kiefer-Wolfowitz NPMLE based on the full sample without distinguishing medfly gender. While the parametric gamma model is capable of capturing the declining portion of the hazard, it is not sufficiently flexible to adapt to the finer features of the observed mortality rates.

Conditional on a draw of θ from the frailty distribution, the premium for a fly of age $T = t$ is,

$$p(t|\theta) = \frac{F(t+1|\theta) - F(t|\theta)}{S(t|\theta)},$$

and integrating with respect to θ we have the unconditional premium,

$$p(t) = \int \frac{F(t+1|\theta) - F(t|\theta)}{S(t|\theta)} h(\theta|t) d\theta,$$

where $h(\theta)$ is the unconditional frailty density, and

$$h(\theta|t) \equiv h(\theta|T > t) = \frac{S(t|\theta)h(\theta)}{S(t)} = \frac{\exp(-\theta A(t))h(\theta)}{\int \exp(-\theta A(t))h(\theta)d\theta}.$$

is the corresponding conditional frailty density. The need to condition the frailty distribution on t may seem odd, but a moment's reflection reveals that mass associated with high frailty values that would imply that subjects would die very quickly, must surely be downweighted once subjects attain an age at which having these values is highly improbable. This is illustrated in Figure 9 where we depict the estimated, conditional frailty based on our NPMLE at four different ages. To exaggerate the magnitude of the smaller mass points of the NPMLE we have plotted the cube root of the density. It is clear that the relatively small mass point at $\log(\theta) = -3.4$ at age 1.5, by age 20 is no longer visible; flies with such a large frailty would almost surely be dead by age 20.

```
R> # Conditional frailty at various ages
R> Gfrailt <- function(age, fit){
+      alpha <- fit[1]
+      beta <- fit[2]
+      delta <- fit[3]
+      A <- (age/beta)^alpha
+      (1 + delta * A)^(-1/delta)
+      }
R> frailt <- function(v, t, alpha, fit){
+          fv = fit$y/sum(fit$y)
+          g = sum(exp(-v * (t^alpha))* fv)
+          exp(-v * (t^alpha)) * fv/g
+          }
R> par(mfrow = c(2,2))
R> v <- exp(fitW$x)
R> for(t in c(1.5, 20, 60, 100)){
+      plot(log(v), frailt(v, t, alpha = 2.95, fitW)^(1/3), type="l",
+          main = paste("age =", t),
+          xlab = expression(log(theta)),
+          ylab = expression(h( theta , t)^{1/3}))
+          }
```

In Figure 10 we plot the ten-day term life insurance premium for medflies at various ages for both the parametric gamma model and the nonparametric model. By varying the parameter k in the `premium` function one can control the term of the life insurance

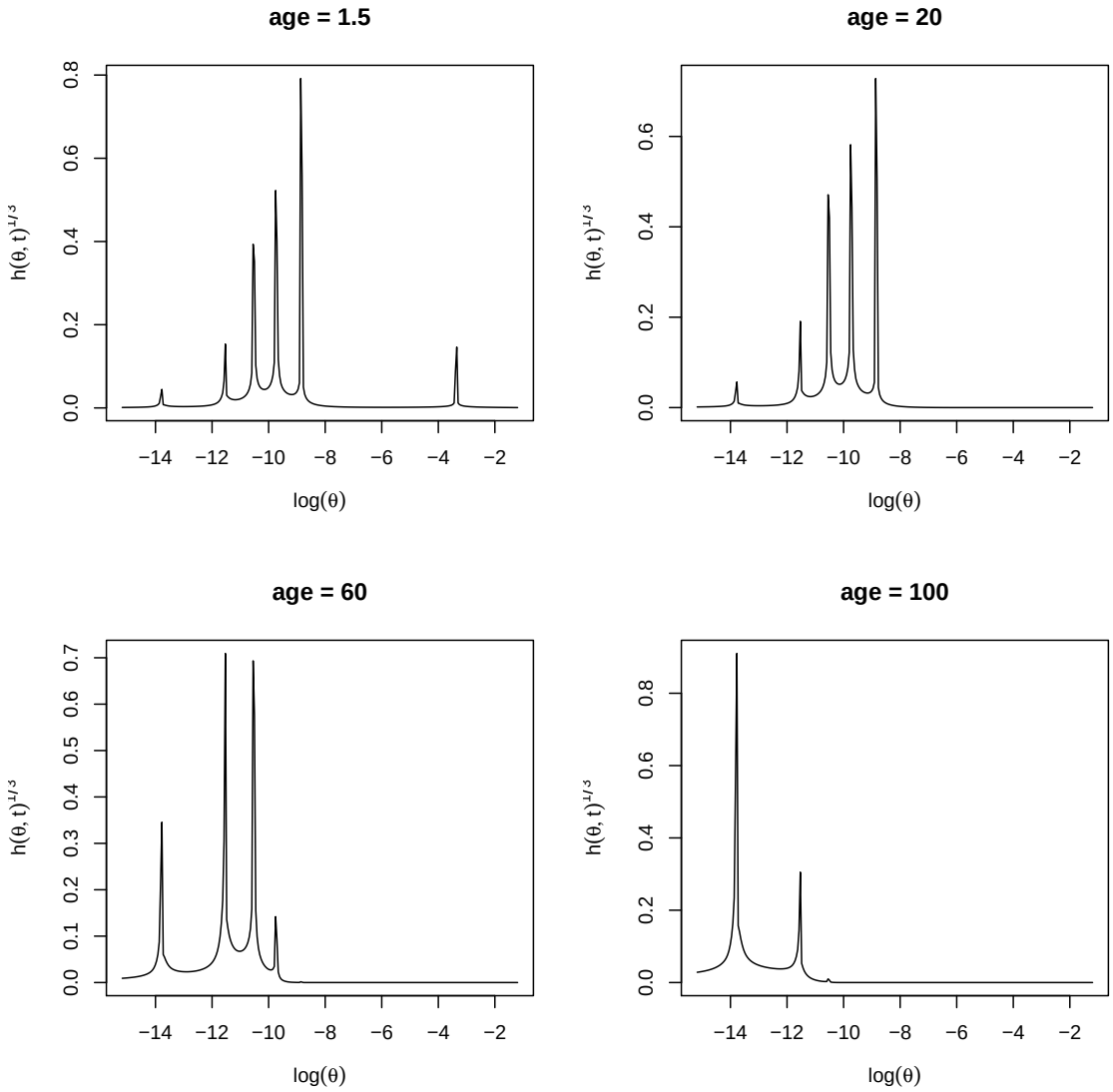

FIGURE 9. Conditional Frailty at Various Ages: Note that the cube root of the frailties have been plotted to accentuate the smaller mass points

policy. In the figure $k = 10$ and the premia profile is somewhat smoother than the instantaneous hazard depicted in Figure 8. Again we see that the gamma model captures the basic shape of the nonparametric rate structure, but misses some of the nuances.

```
R> # Life insurance premia for medflies of various ages
R> premium <- function(v, t, k = 1, alpha, fit){
```

```
+      if("Weibullmix" %in% class(fit)) {
+      R <- t
+      for(i in 1:length(t)){
+          D <- exp(-v * t[i]^alpha) - exp(-v * (t[i] + k)^alpha)
+          D <- D/exp(-v * t[i]^alpha)
+          D[is.nan(D)] <- 1
+          R[i] <- sum(D * frailt(v, t[i], alpha, fit))
+          }
+      }
+      else
+          R <- (Gfrailt(t,fit) - Gfrailt(t+k, fit))/Gfrailt(t, fit)
+      R
+  }
R> v <- exp(fitW$x)
R> R <- premium(v, day, k = 10, alpha = 2.95, fitW)
R> plot(day, R, type = "l", col = 2, ylab = "Premium")
R> R <- premium(v, day, k = 10, alpha = 2.95, fitG)
R> lines(day, R, lty = 2)
R> legend("topright", c("NPMLE", "Gamma"), col = 2:1, lty = 1:2)
```

6. Conclusion

We have described a new approach to computing the nonparametric maximum likelihood estimator of Kiefer and Wolfowitz for a general class of mixture models as implemented in the R package **REBayes**, and illustrated its application in a variety mixture model settings. The approach exploits recent developments in convex optimization as implemented in the Mosek environment of Andersen (2010). Koenker and Mizera (2014a) surveys a broader range of such developments. In addition to the capabilities intended for mixture models the **REBayes** package contains the function `medde` for norm and shape constrained density estimation. Further details on `medde` methods may be found in Koenker and Mizera (2010) and the **REBayes** documentation.

7. Acknowledgements

The authors wish to thank Ivan Mizera for many helpful discussions at the early stages of the development of the **REBayes** package, thanks too for very constructive comments from two referees. This research was partially supported by NSF grant SES-11-53548.

References

Aalen O, Borgan O, Gjessing H. 2008. *Survival and Event History Analysis*. Springer-Verlag.

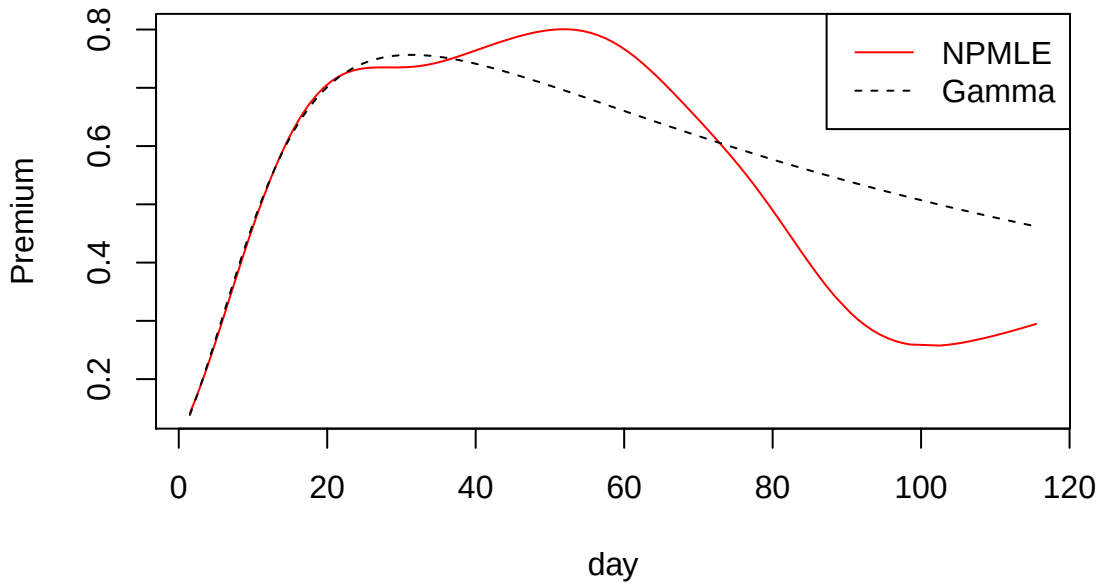

FIGURE 10. Ten-Day Term Life Insurance Premia for Medflies of Various Ages

Andersen ED. 2010. The Mosek optimization tools manual, version 6.0. Available from `http://www.mosek.com`.

Brown L. 2008. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics* **2**: 113–152.

Brown L, Greenshtein E. 2009. Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High Dimensional Vector of Normal Means. *The Annals of Statistics* **37**: 1685–1704.

Bühlmann H, Gisler A. 2005. *A Course in Credibility Theory and its Applications*. Springer-Verlag.

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

Castillo I, van der Vaart A. 2012. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics* **40**: 2069–2101.

Dicker LH, Zhao SD. 2016. High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika* Forthcoming.

Dyson F. 1926. A method for correcting series of parallax observations. *Monthly Notices of the Royal Astronomical Society* **86**: 686–706.

Edgeworth F. 1888. A mathematical theory of banking. *J. Royal Stat Soc* **51**: 113–127.

Efron B. 2008. Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics* **2**: 197–223.

Efron B. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge U. Press: Cambridge.

Efron B. 2011. Tweedie's formula and selection bias. *Journal of the American Statistical Association* **106**: 1602–1614.

Efron B. 2016. Empirical bayes deconvolution estimates. *Biometrika* **103**: 1–20.

Friberg HA. 2012. Users guide to the R-to-Mosek interface. Available from `http://rmosek.r-forge.r-project.org`.

Gu J, Koenker R. 2016a. Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data. *Journal of Applied Econometrics* Forthcoming.

Gu J, Koenker R. 2016b. On a problem of Robbins. *International Statistical Review* **84**: 224–244.

Gu J, Koenker R. 2016c. Unobserved heterogeneity in income dynamics: An empirical Bayes perspective. *J. of Economic and Business Statistics* Forthcoming.

Haastrup S. 2000. Comparison of some Bayesian analyses of heterogeneity in group life insurance. *Scand. Actuarial J.* : 2–16.

Heckman J, Singer B. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**: 63–132.

Jara A, Hanson T, Quintana F, Müller P, Rosner G. 2011. DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software* **40**: 1–30. URL `http://www.jstatsoft.org/v40/i05/`

Jiang W, Zhang CH. 2009. General Maximum Likelihood Empirical Bayes Estimation of Normal Means. *The Annals of Statistics* **37**: 1647–1684.

Jiang W, Zhang CH. 2015. Generalized likelihood ratio test for normal mixtures. *Statistica Sinica* Forthcoming.

Johnstone I, Silverman B. 2004. Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences. *The Annals of Statistics* : 1594–1649.

Johnstone IM, Silverman BW. 2005. Ebayesthresh: R and s-plus programs for empirical bayes thresholding. *Journal of Statistical Software* **12**: 1–38.

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. 2013. Adaptive estimation of regression parameters for the Gaussian scale mixture model. In Beran J, Feng Y, Hebbel H (eds.) *Empirical Economic and Financial Research: A Festschrift for Siegfried Heiler*. Springer-Verlag, 373–378.

Koenker R. 2014. A Gaussian compound decision bakeoff. *Stat* **3**: 12–16.

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

Koenker R, Gu J. 2013. Frailty, profile likelihood and medfly mortality. In Lahiri S, Schick A, Sengupta A, Sriram T (eds.) *Contemporary Developments in Statistical Theory: A Festschrift for Hira Lal Koul*. Springer-Verlag, 227–237.

Koenker R, Mizera I. 2010. Quasi-Concave Density Estimation. *The Annals of Statistics* **38**: 2998–3027.

Koenker R, Mizera I. 2014a. Convex optimization in r. *Journal of Statistical Software* **60**.

Koenker R, Mizera I. 2014b. Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *J. of Am. Stat. Assoc.* **109**: 674–685.

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

Lindsay B. 1983. The geometry of mixture likelihoods: A general theory. *The Annals of Statistics* **11**: 86–94.

Liu J. 1996. Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics* **24(3)**: 911–930.

Martin R, Walker SG. 2014. Asymptotically minimax empirical Bayes estimation of a sparse normal mean. *Electronic Journal of Statistics* **8**: 2188–2206.

Mukherjee G, Brown LD, Rusmevichientong P. 2016. Efficient empirical bayes prediction under check loss using asymptotic risk estimates. Available from: https://arxiv.org/abs/1511.00028.

Norberg R. 1989. Experience rating in group life insurance. *Scand. Actuarial J.* : 194–224.

Pearson K. 1894. Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. London A* **185**: 71–110.

Robbins H. 1950. A generalization of the method of maximum likelihood; estimating a mixing distribution (preliminary report). *The Annals of Mathematical Statistics* **21**: 314–315.

Robbins H. 1951. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, volume I. University of California Press: Berkeley, 131–149.

Robbins H. 1956. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume I. University of California Press: Berkeley, 157–163.

Robbins H. 1982. Estimating many variances. In Gupta S, Berger JO (eds.) *Statistical Decision Theory and Related Topics III*, volume 2. Academic Press: New York, 75–85.

Sun W, Cai TT. 2007. Oracle and adaptive compound decision rules for false discovery rate control. *Journal American Statistical Association* **102**: 901–912.

Sun W, McLain AC. 2012. Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association* **107**: 673–687.

Teicher H. 1967. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics* **38**: 1300–1302.

van der Vaart A. 1996. Efficient maximum likelihood estimation in semiparametric mixture models. *Annals of Statistics* **24(2)**: 862–878.

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