Maximum Approximate Bernstein/Beta Likelihood Estimation in R package 'mable'

Zhong Guan

24/09/2024

Contents

1 Introduction

Any continuous density function f on a known closed interval $[a, b]$ can be approximated by Bernstein polynomial $f_m(x; p) = (b-a)^{-1} \sum_{i=0}^{m} p_i \beta_{mi}[(x-a)/(b-a)]$, where $p_i \geq 0$, $\sum_{i=0}^{m} p_i = 1$ and $\beta_{mi}(u) = (m+1)\binom{m}{i} u^i (1-u)^{m-i}$, $i = 0, 1, \ldots, m$, is the beta density with shapes $(i+1, m-i+1)$. This provides a way to approximately model the unknwon underlying density with a mixture beta model with an appropriate *model degree* m and solve a nonparametric or semiparametric statistical problem "parametrically'' using the maximum likelihood method. For instance, based on one-sample data, $x_1, \ldots, x_n$, one can estimate a nonparametric density f parametrically by maximizing the approximate likelihood $\ell(p) = \sum_{j=0}^{n} \log f_m(x_j; p)$ with respect to p (Guan 2016).

Since the Bernstein polynomial model of degree m is nested in the model of degree $m+1$, the maximum likelihood is increasing in m. The increase is negligible when the model becomes overfitting. Therefore an optimal degree can be chosen as the change-point of the log likelihood ratios over a set of consecutive candidate model degrees.

This approach works surprisingly well for even more complicated models and data. With an estimate $\hat{p} = (\hat{p}_0, \ldots, \hat{p}_m)$ of p one can estimate the cumulative distribution function F by $\hat{F}(x) = F_m(x; \hat{p}) = \sum_{i=0}^{m} \hat{p}_i B_{mi}[(x-a)/(b-a)]$, where $B_{mi}(u) = \int_0^u \beta_{mj}(t)dt$, $i = 0, \ldots, m$, is the beta distribution function with shapes $(i+1, m-i+1)$. This manual will illustrate the use of the R package `mable` for obtaining not only smooth estimates of density, cumulative distribution, and survival functions but also estimates of parameters such as regression coefficients.

2 One-sample Problems

2.1 Raw Data

Let $x_1, \ldots, x_n$ be a sample from a population with cumulative distribution function F and density function f on $[a, b]$. If $[a, b]$ is unknown we choose $[a, b] \supset [x_{(1)}, x_{(n)}]$, where $x_{(1)}$ and $x_{(n)}$ are the minimum and maximum statistics, respectively. A lower bound for unimodal density is $m_b = \mu(1-\mu)/\sigma^2 - 3$, where μ and σ^2 are, respectively, the mean and variance of the distribution after transformation $Y = (X-a)/(b-a)$.

2.1.1 Example: Vaal River Annual Flow Data

For the annual flow data of Vaal River at Standerton as given by Table 1.1 of Linhart and Zucchini (1986) give the flow in millions of cubic metres,

```
data(Vaal.Flow)
head(Vaal.Flow, 3)
```

```
##   Year Flow
## 1 1905  222
## 2 1906 1094
## 3 1907  452
```

we want to estimate the density and the distribution functions of annual flow

```
vaal<-mable(Vaal.Flow$Flow, M = c(2,100), interval = c(0, 3000), IC = "all",
       controls = mable.ctrl(sig.level = 1e-8, maxit = 2000, eps = 1.0e-9))
```

% the following hided data contain object vaal obtained by dput(vaal)

Here we truncate the density by `interval=c(0, 3000)` and choose an optimal degree m among the candidate degrees M[1]:M[2] using the method of change-point. The maximum number of iterations is `maxit` and the convergence criterion is `eps` for each m of M[1]:M[2]. The search of an optimal degree stops when the p-value `pval` of change-point test falls below the specified significance level `sig.level` or the largest degree, M[2], has been reached. If the latter occurs a warning message shows up. In this case we should check the last value of `pval`. In the above example, we got warning message and the last `pval`, 1.3396916×10^{-8}, which is small enough. The selected optimal degree is $m = 19$. One can also look at the Bayesian information criteria, `BIC`, and other information criteria, Akaike information criterion `AIC` and Hannan–Quinn information criterion `QHC`, at each candidate degree. These information criteria are not reliable due to the difficulty of determining the model dimension. The `plot` method for `mable` class object can visualize some of the results.

```
op <- par(mfrow = c(1,2))
layout(rbind(c(1, 2), c(3, 3)))
plot(vaal, which = "likelihood", cex = .5)
plot(vaal, which = "change-point", lgd.x = "topright")
hist(Vaal.Flow$Flow, prob = TRUE, xlim = c(0,3000), ylim =c(0,.0022), breaks =100*(0:30),
    main = "Histogram and Densities of the Annual Flow of Vaal River",
    border = "dark grey",lwd = 1, xlab = "Flow", ylab = "Density", col ="light grey")
lines(density(x<-Vaal.Flow$Flow, bw = "nrd0", adjust = 1), lty = 2, col = 2,lwd = 2)
lines(y<-seq(0, 3000, length=100), dlnorm(y, mean(log(x)), sqrt(var(log(x)))),
    lty = 4, col = 4, lwd = 2)
plot(vaal, which = "density", add = TRUE, lwd = 2)
legend("topright", lty = c(1, 4, 2), col = c(1, 4, 2), bty = "n",lwd = 2,
c(expression(paste("MABLE: ",hat(f)[B])), expression(paste("Log-Normal: ",hat(f)[P])),
    expression(paste("KDE: ",hat(f)[K]))))
```

```
par(op)
```

In Figure 1, the unknown density f is estimated by `MABLE` $\hat{f}_B$ using optimal degrees $m = 19$ selected using the exponential change-point method, the parametric estimate using `Log-Normal` model and the kernel density estimate `KDE`: $\hat{f}_K$.

We can also look at the plots of AIC, BIC, and QHC, and likelihood ration(LR) of gamma change-point.

```
M <- vaal$M[1]:vaal$M[2]
aic <- vaal$ic$AIC
bic <- vaal$ic$BIC
qhc <- vaal$ic$QHC
vaal.gcp <- optim.gcp(vaal) # choose m by gamma change-point model
lr <- vaal.gcp$lr
plot(M, aic, cex = 0.7, xlab = "m", ylab = "", main = "AIC, BIC, QHC, and LR",
      ylim = c(ymin<-min(aic,bic,qhc,lr), ymax<-max(aic,bic,qhc,lr)), col = 1)
points(M, bic, pch = 2, cex = 0.7, col = 2)
points(M, qhc, pch = 3, cex = 0.7, col = 3)
points(M[-1], lr, pch = 4, cex = 0.7, col = 4)
segments(which.max(aic)+M[1]-1->m1, ymin, m1, max(aic), lty = 2)
segments(which.max(bic)+M[1]-1->m2, ymin, m2, max(bic), lty = 2, col = 2)
segments(which.max(qhc)+M[1]-1->m3, ymin, m3, max(qhc), lty = 2, col = 3)
segments(which.max(lr)+M[1]->m4, ymin, m4, max(lr), lty = 2, col = 4)
axis(1, c(m1,m2, m3, m4),  as.character(c(m1,m2,m3,m4)), col.axis = 4)
legend("topright", pch=c(1,2,3,4), c("AIC", "BIC", "QHC", "LR"), bty="n", col=c(1,2,3,4))
```

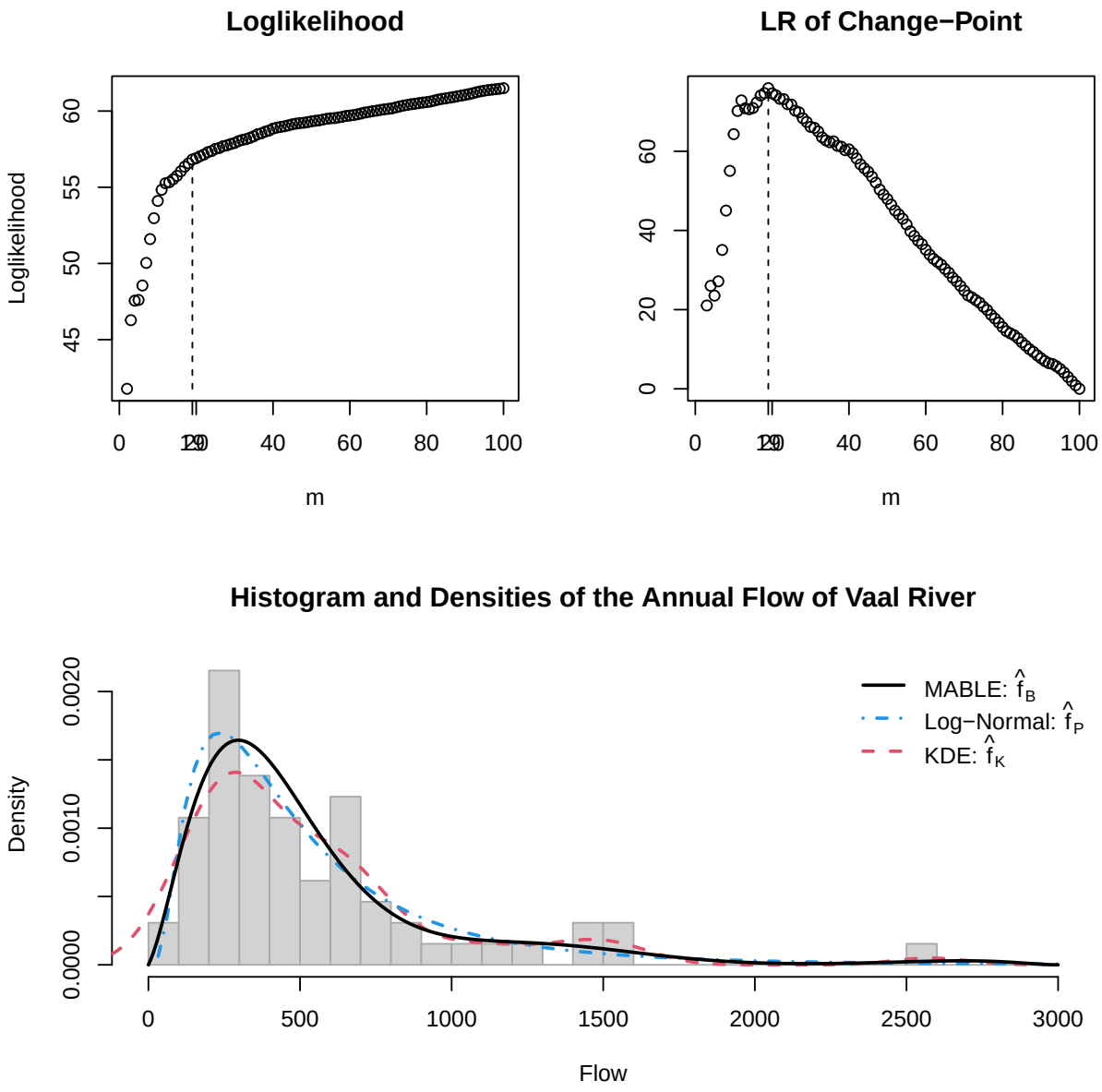

Figure 1: Vaal River Annual Flow Data

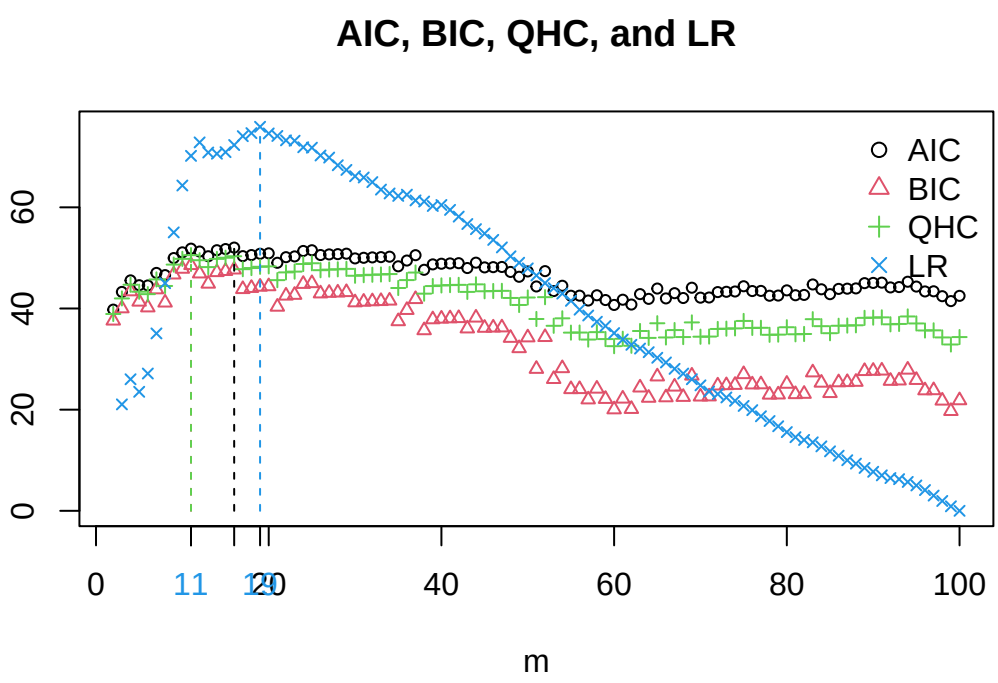

Figure 2: AIC and BIC based on Vaal River Data

From Figure 2 we see that the gamma change-point method gives the same optimal degree $m = 19$ as the exponential method; BIC and QHC give the same degree $m = 11$; the degree $m = 16$ selected by AIC method is closer to the one selected by change-point methods. From this Figure we also see that unlike the LR plot the information criteria do not have clear peak points.

For any given degree m, one can fit the data by specifying `M=m` in `mable()` to obtain an estimate of f.

The `summary` method `summary.mable` prints and returns invisibly the summarized results.

```
summary(vaal)
```

```
## Call: mable() for raw data
## Obj Class: mable
## Input Data: Vaal.Flow$Flow
## Dimension of data: 1
## Optimal degree m: 19
## P-value of Change-point: 1.339692e-08
## Maximum loglikelihood: 56.81939
## MABLE of p: can be retrieved using name 'p'
## Note: the optimal model degrees are selected by the method of change-point.
```

The mixing proportions, the coefficients of the Bernstein polynomials, p can be obtained as either `p<-vaal$p`

```
p<-vaal$p
p
```

```
##  [1] 6.663951e-142  1.049278e-01  7.594313e-01  2.309036e-08  2.508441e-16
##  [6]  7.194500e-21  1.715117e-20  1.049294e-07  1.202354e-01  1.097612e-22
## [11]  3.689746e-96 2.261501e-228  0.000000e+00  0.000000e+00  0.000000e+00
## [16] 5.038909e-234  1.226874e-39  1.540542e-02 4.735457e-197  0.000000e+00
```

or `summary(res)$p`.

The method of change-point for choosing model degree is computer-intensive especially for multimodal density. A better lower bound for model degree is

$$m_b = \frac{\mu(1-\mu) - \sigma^2}{\sigma^2 - \sum_{i=1}^{k} \lambda_i(\mu_i - \mu)^2} - 2,$$

where λ_i and μ_i are, respectively, the mixing proportion and mean value of teh ith component density. Less computer-intensive methods such as the method of moment and the method of mode implemented by `optimal()` can be used (see Figure 3).

2.1.2 Example: The Old Faithful data

```
x<-faithful
x1<-faithful[,1]
x2<-faithful[,2]
a<-c(0, 40); b<-c(7, 110)
mu<-(apply(x,2,mean)-a)/(b-a)
s2<-apply(x,2,var)/(b-a)^2
# mixing proportions
lambda<-c(mean(x1<3),mean(x2<65))                                    # gu
mu1<-(c(mean(x1[x1<3]), mean(x2[x2<65]))-a)/(b-a)
mu2<-(c(mean(x1[x1>=3]), mean(x2[x2>=65]))-a)/(b-a)
# estimate lower bound for m
mb<-ceiling((mu*(1-mu)-s2)/(s2-lambda*(1-lambda)*(mu1-mu2)^2)-2)
```

```
m1<-optimable(x1, interval=c(a[1],b[1]), nmod=2, modes=c(2,4.5))$m
m2<-optimable(x2, interval=c(a[2],b[2]), nmod=2, modes=c(52.5,80))$m
erupt1<-mable(x1, M=mb[1], interval=c(a[1],b[1]))
erupt2<-mable(x1, M=m1, interval=c(a[1],b[1]))
wait1<-mable(x2, M=mb[2],interval=c(a[2],b[2]))
wait2<-mable(x2, M=m2,interval=c(a[2],b[2]))
```

For this data set, we obtain lower bounds for the marginal densities $m_b = (78, 28)$ and $\hat{m} = (122, 34)$.

```
op<-par(mfrow=c(1,2), cex=0.8)
hist(x1, probability = TRUE, col="grey", border="white", main="", xlab="Eruptions",
     ylim=c(0,.65), las=1)
plot(erupt1, add=TRUE,"density")
plot(erupt2, add=TRUE,"density",lty=2,col=2)
legend("topleft", lty=c(1,2),col=1:2, bty="n", cex=.7,
       c(expression(paste("m = ", m[b])),expression(paste("m = ", hat(m)))))
hist(x2, probability = TRUE, col="grey", border="white", main="", xlab="Waiting", las=1)
plot(wait1, add=TRUE,"density")
plot(wait2, add=TRUE,"density",lty=2,col=2)
legend("topleft", lty=c(1,2),col=1:2, bty="n", cex=.7,
       c(expression(paste("m = ", m[b])),expression(paste("m = ", hat(m)))))
```

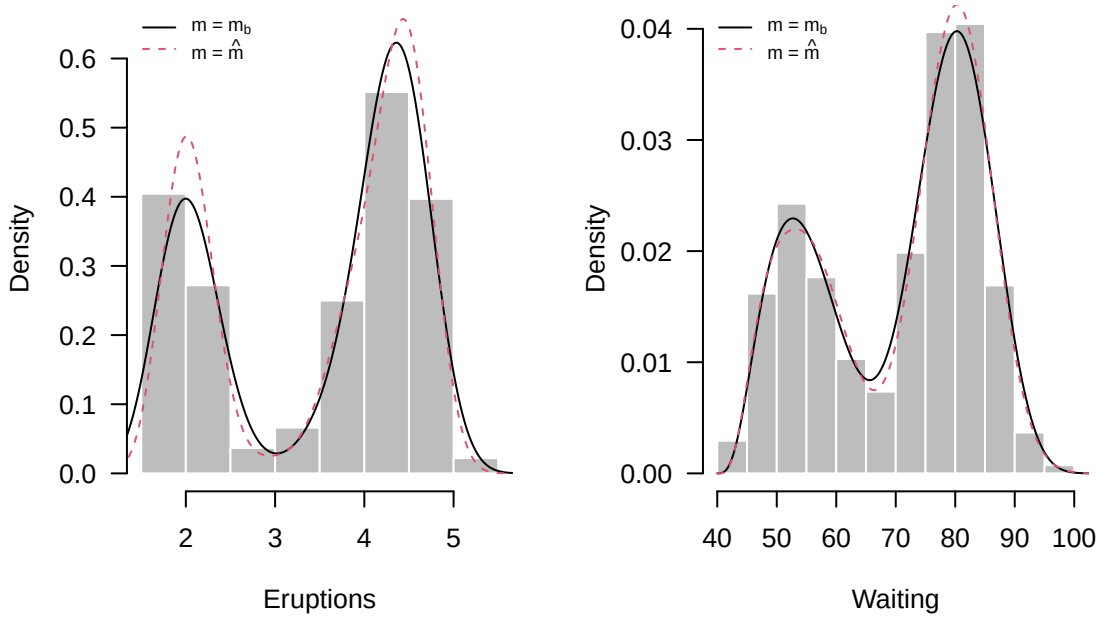

Figure 3: The Old Faithful data

```
par(op)
```

2.2 Grouped Data

With a grouped dataset, a frequency table of data from a continuous population, one can estimate the density from histogram using `mable.group()` with an optimal degree m chosen from `M[1]:M[2]` or with a given degree m using `M=m` (Guan 2017).

2.2.1 Example: The Chicken Embryo Data

Consider the chicken embryo data contain the number of hatched eggs on each day during the 21 days of incubation period. The times of hatching ($n = 43$) are treated as grouped by intervals with equal width of one day. The data were studied first by Jassim et al. (1996). Kuurman et al. (2003) and Lin and He (2006) also analyzed the data using the minimum Hellinger distance estimator, in addition to other methods assuming some parametric mixture models including Weibull model.

```
data(chicken.embryo)
head(chicken.embryo, 2)
```

```
## $day
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##
## $nT
##  [1]  6  5 11  2  2  3  0  0  0  0  1  0  0  0  1  0  1  5  1  4  1
```

```
a <- 0
b <- 21
day <- chicken.embryo$day
nT <- chicken.embryo$nT
```

```
embryo<-mable.group(x = nT, breaks = a:b, M=c(2,100), interval = c(a, b), IC = "aic",
    controls = mable.ctrl(sig.level = 1e-6,  maxit = 2000, eps = 1.0e-7))
```

```
Day <- rep(day,nT)
op <- par(mfrow = c(1,2), lwd = 2)
layout(rbind(c(1, 2), c(3, 3)))
plot(embryo, which = "likelihood")
plot(embryo, which = "change-point")
fk <- density(x = rep((0:20)+.5, nT), bw = "sj", n = 101, from = a, to = b)
hist(Day, breaks = seq(a,b,  length = 12), freq = FALSE, col = "grey", border = "white",
     main = "Histogram and Density Estimates")
plot(embryo, which = "density", cex = 0.7, add = TRUE)
lines(fk, lty = 2, col = 2)
legend("top", lty = c(1:2), c("MABLE", "Kernel"), bty = "n", col = c(1:2))
```

```
par(op)
```

We see in Figure 5 that AIC and gamma change-point method give the same optimal degree as the one, $m = 13$, given by the exponential change-point method. However, BIC fails in choosing a useful model degree.

```
M <- embryo$M[1]:embryo$M[2]
aic <- embryo$ic$AIC
bic <- embryo$ic$BIC
res.gcp <- optim.gcp(embryo) # choose m by gamma change-point model
lr <- res.gcp$lr
plot(M, aic, cex = 0.7, col = 1, xlab = "m", ylab = "", ylim = c(ymin<-min(aic,bic,lr),
    ymax<-max(aic,bic,lr)), main = "AIC, BIC, and LR")
points(M, bic, pch = 2, cex = 0.7, col = 2)
points(M[-1], lr, pch = 3, cex = 0.7, col = 4)
segments(which.max(aic)+M[1]-1->m1, ymin, m1, max(aic), lty = 2)
segments(which.max(bic)+M[1]-1->m2, ymin, m2, max(bic), lty = 2, col = 2)
segments(which.max(lr)+M[1]->m3, ymin, m3, max(lr), lty = 2, col = 4)
axis(1, c(m1,m2, m3),  as.character(c(m1,m2,m3)), col.axis = 4)
legend("right", pch = 1:3, c("AIC", "BIC", "LR"), bty = "n", col = c(1,2,4))
```

The results are summarized as follows.

```
summary(embryo)
```

```
## Call: mable.group() for grouped data
## Obj Class: mable
## Input Data: nT
## Dimension of data: 1
## Optimal degree m: 13
## P-value of Change-point: 9.925959e-07
## Maximum loglikelihood: -107.318
## MABLE of p: can be retrieved using name 'p'
## Note: the optimal model degrees are selected by the method of change-point.
```

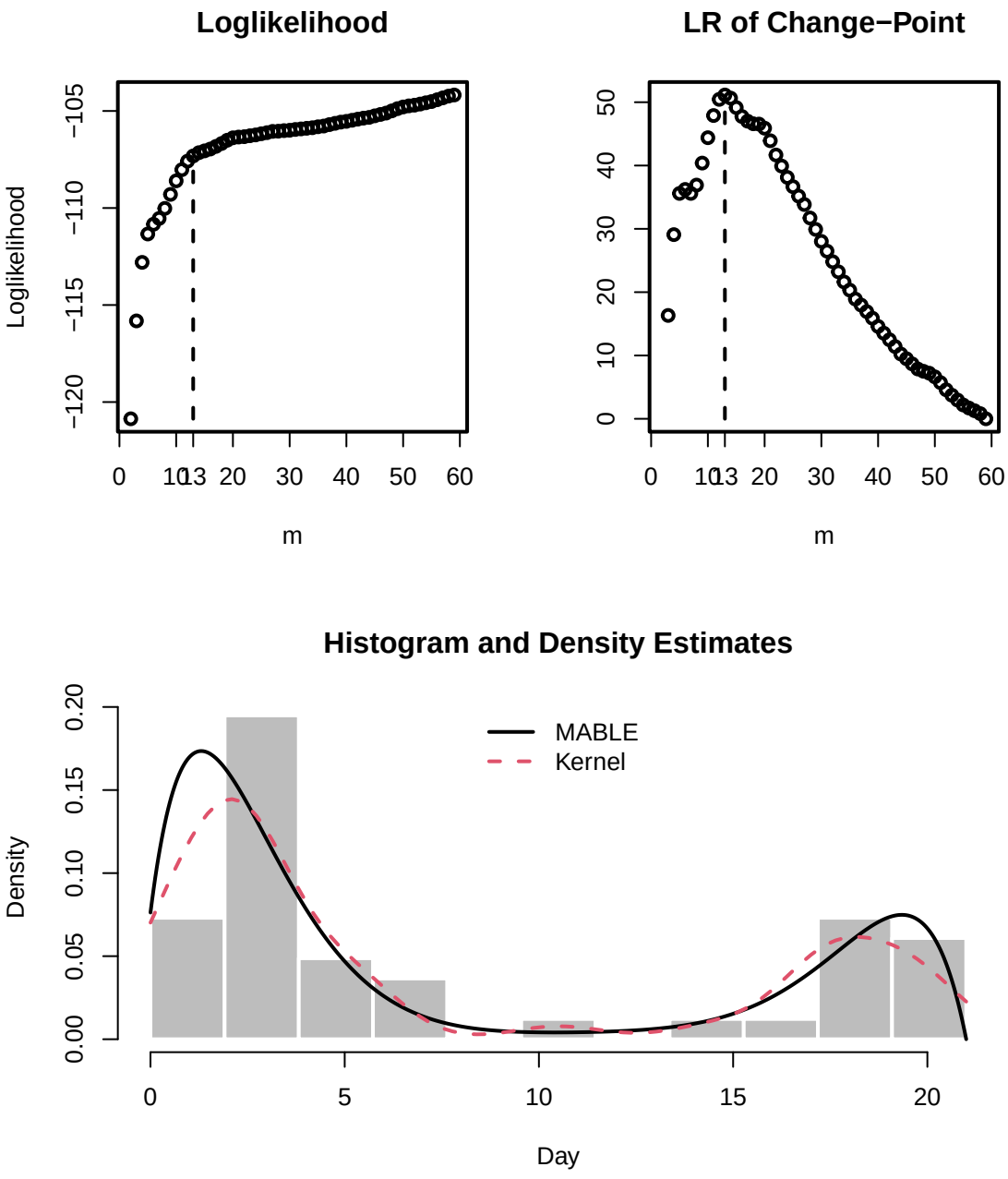

Figure 4: Chicken Embryo Data

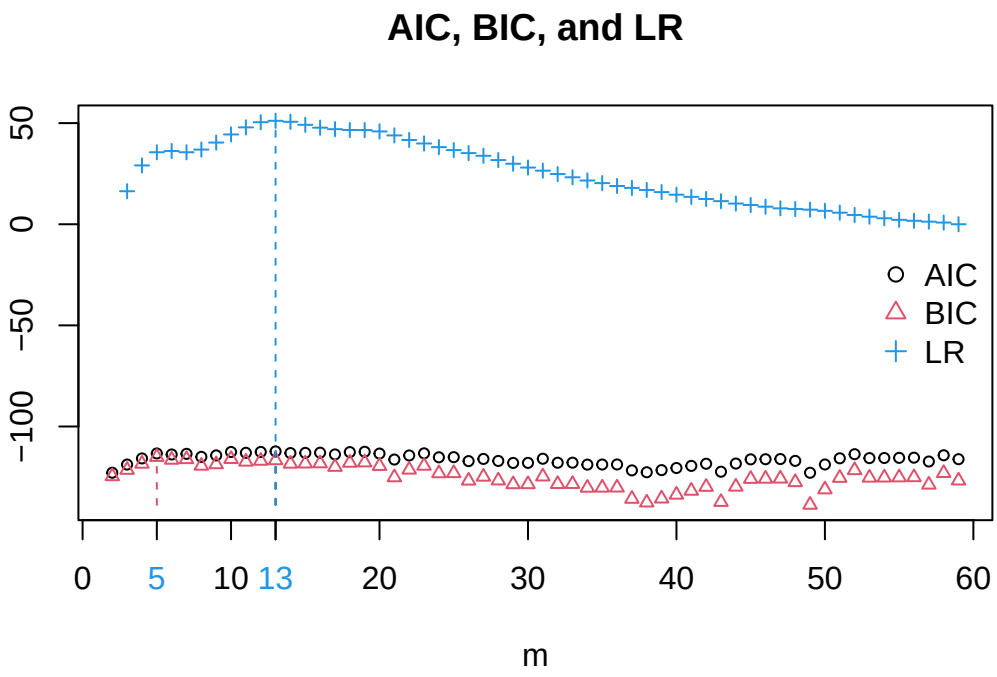

Figure 5: AIC and BIC based on Chicken Embryo Data

2.3 Contaminated Data–Density Deconvolution

Consider the additive measurement error model $Y = X + \epsilon$, where X has an unknown distribution F, ϵ has a known distribution G, and X and ϵ are independent. We want to estimate density $f = F'$ based on independent observations, $y_i = x_i + \epsilon_i$, $i = 1, \ldots, n$, of Y, where x_i's and ϵ_i's are unobservable. `mable.decon()` implements the method of Guan (2021a) and gives an estimate of the density f using the approximate Bernstein polynomial model.

2.3.1 Example: A Simulated Normal Dataset

```
set.seed(123)
mu <- 1; sig <- 2; a <- mu - sig*5; b <- mu + sig*5;
gn <- function(x) dnorm(x, 0, 1)
n <- 50;
x <- rnorm(n, mu, sig); e <- rnorm(n); y <- x + e;
```

```
decn <- mable.decon(y, gn, interval = c(a,b), M = c(5, 50))
```

```
op <- par(mfrow = c(2,2), lwd = 2)
plot(decn, which = "likelihood")
plot(decn, which = "change-point", lgd.x = "right")
plot(xx<-seq(a, b, length = 100), yy<-dnorm(xx, mu, sig), type = "l", xlab = "x",
     ylab = "Density", ylim = c(0, max(yy)*1.1))
plot(decn, which = "density", add = TRUE, lty = 2, col = 2)
```

```r
# kernel density based on pure data
lines(density(x), lty = 5, col = 4)
legend("topright", bty = "n", lty = c(1,2,5), col = c(1,2,4), c(expression(f),
    expression(hat(f)),  expression(tilde(f)[K])))
plot(xx, yy<-pnorm(xx, mu, sig), type = "l", xlab = "x", ylab = "Distribution Function")
plot(decn, which = "cumulative", add = TRUE, lty = 2, col = 2)
legend("bottomright", bty = "n", lty = c(1,2), col = c(1,2), c(expression(F),
    expression(hat(F))))
```

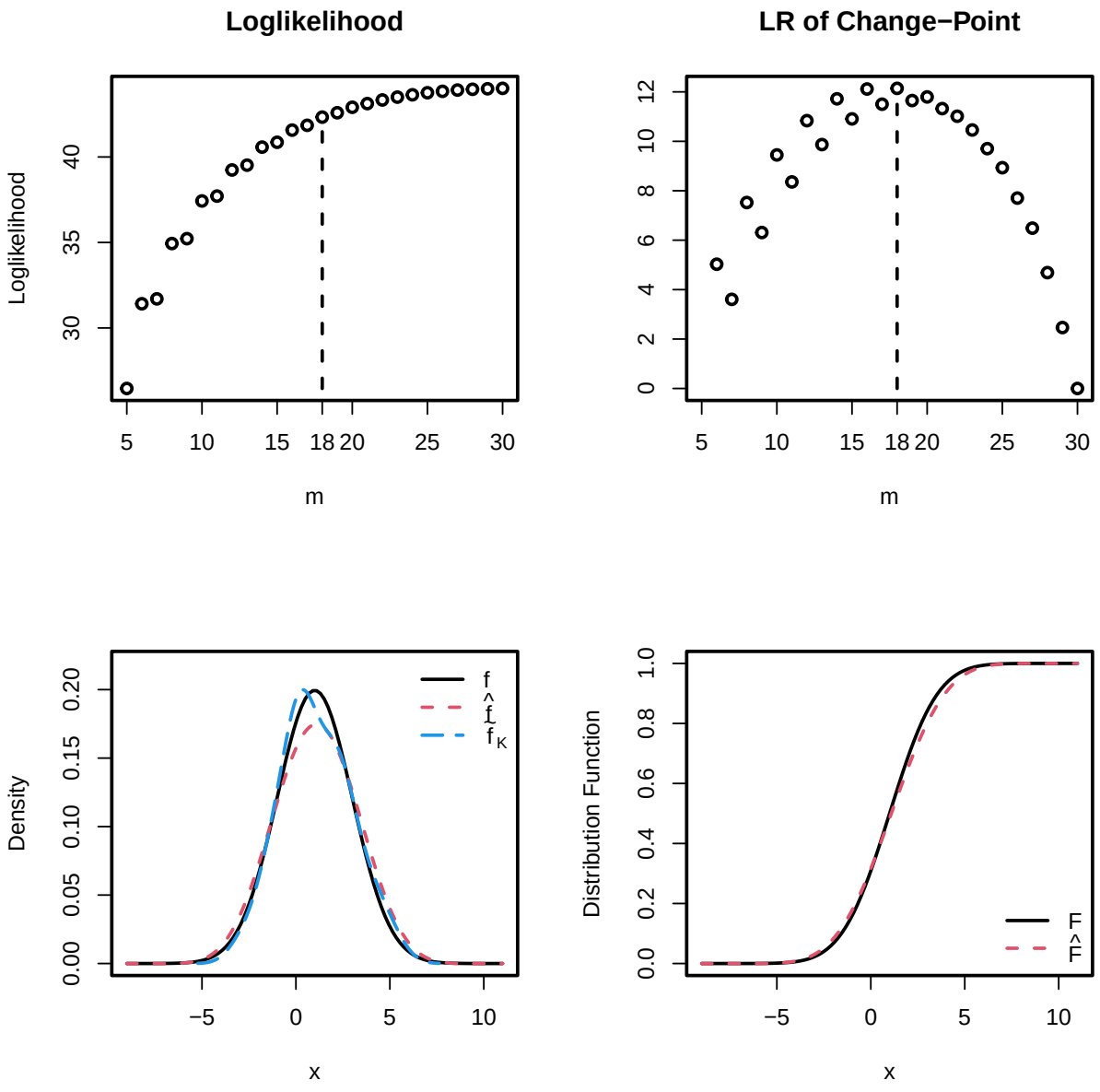

Figure 6: Simulated Normal Data

```
par(op)
```

2.4 Interval Censored Data

When data are interval censored, the "`interval2`'' type observations are of the form (l, u) which is the interval containing the event time. Data is uncensored if $l = u$, right censored if $u =$ `Inf` or $u =$ `NA`, and left censored data if $l = 0$.

Let $f(t)$ and $F(t) = 1 - S(t)$ be the density and cumulative distribution functions of the event time, respectively, on $(0, \tau)$, where $\tau \leq \infty$. If τ is unknown or $\tau = \infty$, then $f(t)$ on $[0, \tau_n]$ can be approximated by $fm(t; p) = \tau_n^{-1} \sum_{i=0}^{m} p_i \beta_{mi}(t/\tau_n)$, where $p_i \geq 0$, $i = 0, \ldots, m$, $\sum_{i=0}^{m} p_i = 1 - p_{m+1}$, and τ_n is the largest observation, either uncensored time, or right endpoint of interval/left censored, or left endpoint of right censored time. So we can approximate $S(t)$ on $[0, \tau]$ by $Sm(t; p) = \sum_{i=0}^{m+1} p_i \bar{B}_{mi}(t/\tau)$, where $\bar{B}_{mi}(u) = 1 - \int_0^u \beta_{mj}(t)dt$, $i = 0, \ldots, m$, is the beta survival function with shapes $(i + 1, m - i + 1)$, $\bar{B}_{m,m+1}(t) = 1$, $p_{m+1} = 1 - \pi$, and $\pi = F(\tau_n)$. For data without right-censored time, $p_{m+1} = 1 - \pi = 0$. The search for optimal degree m among `M=c(m0,m1)` using the method of change-point is stopped if either `m1` is reached or the test for change-point results in a p-value `pval` smaller than `sig.level`. Guan (2021b) proposed a method, as a special case of a semiparametric regression model, for estimating p with an optimal degree m. The method is implemented in R function `mable.ic()`.

2.4.1 Example: The Breast Cosmesis Data

Consider the breast cosmesis data as described in Finkelstein and Wolfe (1985) is used to study the cosmetic effects of cancer therapy. The time-to-breast-retractions in months (T) were subject to interval censoring and were measured for 94 women among them 46 received radiation only ($X = 0$) (25 right-censored, 3 left-censored and 18 interval censored) and 48 received radiation plus chemotherapy ($X = 1$) (13 right-censored, 2 left-censored and 33 interval censored). The right-censored event times were for those women who did not experienced cosmetic deterioration.

We fit the Breast Cosmesis Data as two-sample data separately.

```
head(cosmesis, 3)
```

```
##   left right treat
## 1   45    NA    RT
## 2    6    10    RT
## 3    0     7    RT
```

```
bc.res0 <- mable.ic(cosmesis[cosmesis$treat == "RT",1:2], M = c(1,50), IC = "none")
bc.res1 <- mable.ic(cosmesis[cosmesis$treat == "RCT",1:2], M = c(1,50), IC = "none")
```

As the warning message suggested, we check the `pval`. The `pval` when the search stopped is 0.0452.

```
op <- par(mfrow = c(2,2),lwd = 2)
plot(bc.res0, which = "change-point", lgd.x = "right")
plot(bc.res1, which = "change-point", lgd.x = "right")
plot(bc.res0, which = "survival", xlab = "Months", ylim = c(0,1), main = "Radiation Only")
plot(bc.res1, which = "survival", xlab = "Months", main = "Radiation and Chemotherapy")
```

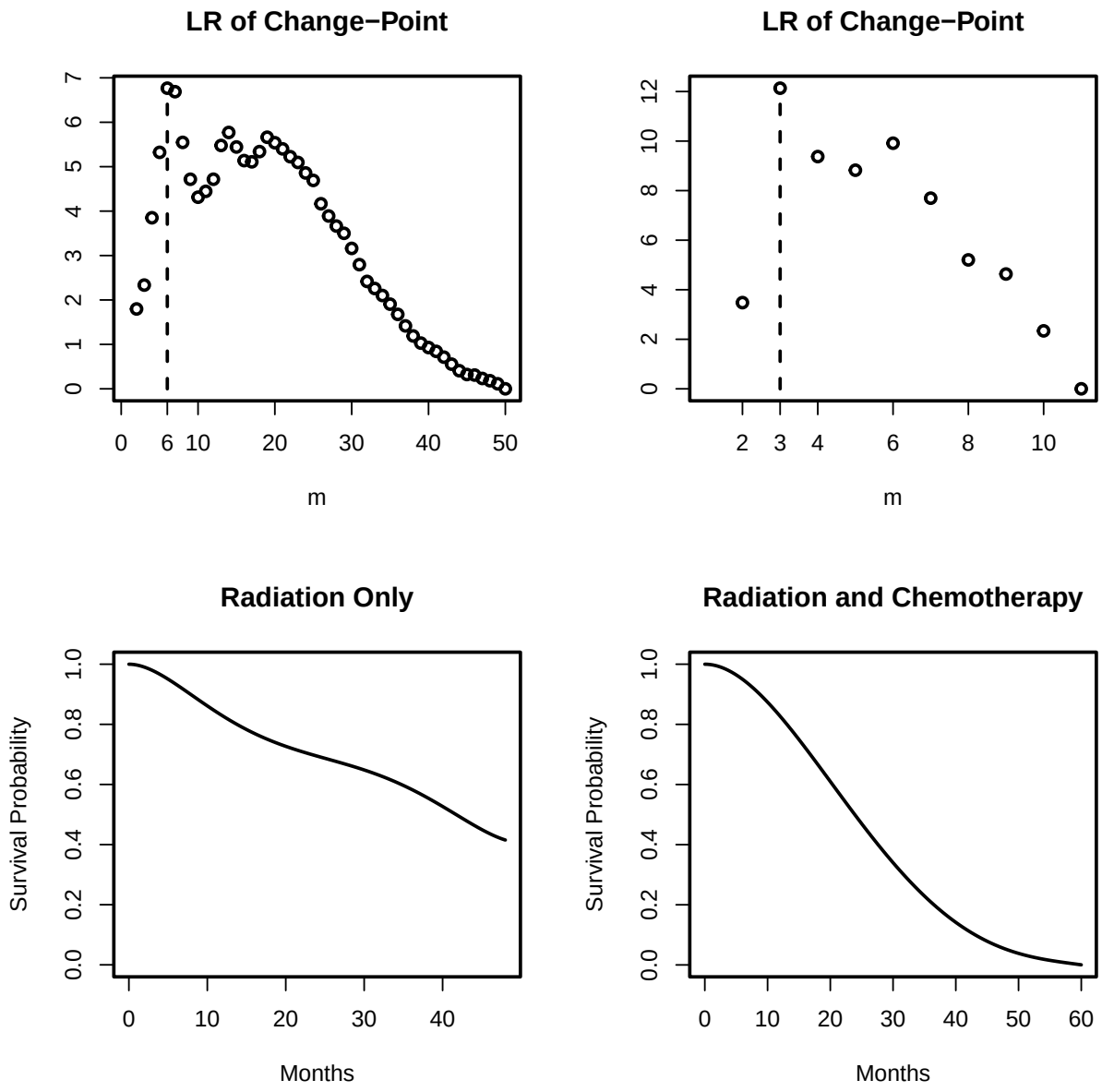

Figure 7: Breast Cosmesis Data

```
par(op)
```

2.5 Multivariate Data

A d-variate density f on a hyperrectangle $[a, b] = [a_1, b_1] \times \cdots \times [a_d, b_d]$ can be approximated by a mixture of d-variate beta densities on $[a, b]$, $\beta_{mj}(x; a, b) = \prod_{i=1}^{d} \beta_{m_i,j_i}[(x_i - a_i)/(b_i - a_i)]/(b_i - a_i)$, with proportions $p(j_1, \ldots, j_d)$, $0 \le j_i \le m_i, i = 1, \ldots, d$. Because all the marginal densities can be approximated by Bernstein polynomials, we can choose optimal degree m_i based on observations of the i-th component of x. For the i-th marginal density, an optimal degree is selected using `mable()`. Then fit the data using EM algorithm with the selected optimal degrees $m = (m_1, \ldots, m_d)$ to obtain a vector p of the mixture proportions $p(j_1, \ldots, j_d)$, arranged in the column-major order of $j = (j_1, \ldots, j_d)$, $(p_0, \ldots, p_{K-1})$, where $K = \prod_{i=1}^{d}(m_i + 1)$. The proposed method of Wang and Guan (2019) is implemented by function `mable.mvar()`.

2.5.1 Example: The Old Faithful Data

```
data(faithful)
head(faithful, 3)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
```

```
a <- c(0, 40); b <- c(7, 110)
```

```
#faith2 <- mable.mvar(faithful, M = c(60,30), interval = cbind(a,b))
faith2 <- mable.mvar(faithful, M = c(46,19), search =FALSE, interval = cbind(a,b))
```

```
plot(faith2, which = "density")
```

The density surface for two-dimensional data can be plot using the `plot` method (see Figure 8). The summarized results are given below.

```
summary(faith2)
```

```
## Call: mable.mvar() for multivariate data
## Obj Class: mable
## Input Data: eruptions waiting
## Dimension of data: 2
## Optimal degrees:
##   eruptions waiting
## m        46      19
## P-value of Change-point:
## Maximum loglikelihood: 507.5615
## MABLE of p: can be retrieved using name 'p'
## Note: the optimal model degrees are selected by the method of change-point.
```

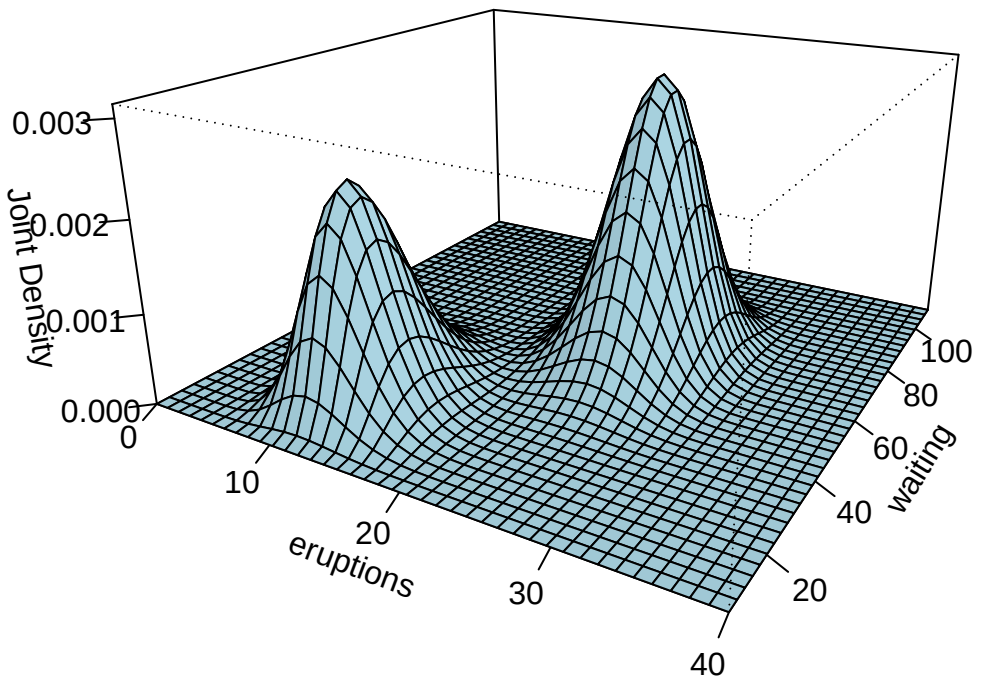

Figure 8: Density Estimate based on the Old Faithful Data

For multivariate density, the model degree can be chosen based on marginal data. As in Section 2.1.2 we obtained $m_b = (78, 28)$ and $\hat{m} = (122, 34)$.

```
x<-faithful
x1<-faithful[,1]
x2<-faithful[,2]
a<-c(0, 40); b<-c(7, 110)
mb<-c(78,28)
m1<-122
m2<-34
```

```
oldfaith1<- mable.mvar(faithful, M = mb, search =FALSE, interval = cbind(a,b))
oldfaith2<- mable.mvar(faithful, M = c(m1,m2), search =FALSE, interval = cbind(a,b))
```

```
op<-par(mfrow=c(1,2), cex=0.7)
plot(oldfaith1, which="density", contour=TRUE)
plot(oldfaith2, which="density", contour=TRUE, add=TRUE, lty=2, col=2)
plot(oldfaith1, which="cumulative", contour=TRUE)
plot(oldfaith2, which="cumulative", contour=TRUE, add=TRUE, lty=2, col=2)
```

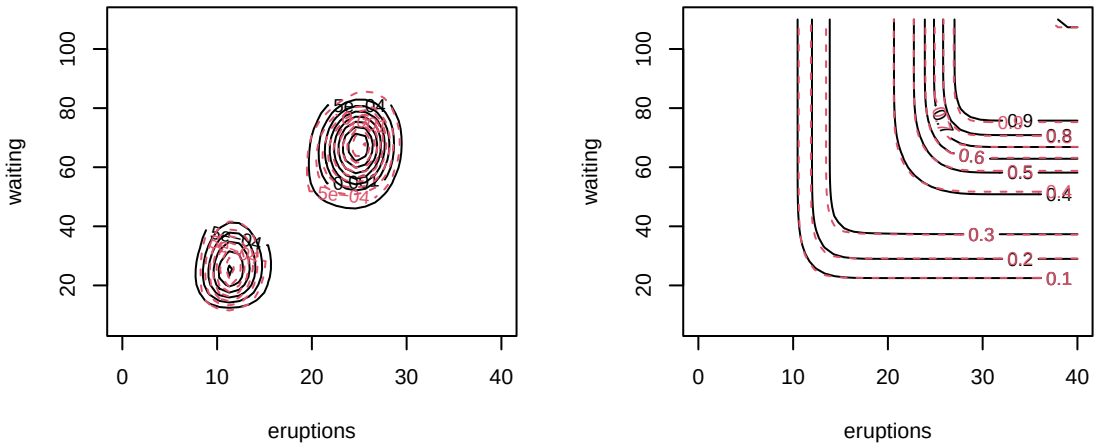

Figure 9: The Old Faithful data

```
par(op)
```

3 Event Time Data with Covariates

Let T be an event time and X be an associated d-dimensional covariate with distribution $H(x)$ on $\mathcal{X}$. We denote the marginal and the conditional survival functions of T, respectively, by $S(t) = \bar{F}(t) = 1 - F(t) = \Pr(T > t)$ and $S(t|x) = \bar{F}(t|x) = 1 - F(t|x) = \Pr(T > t|X = x)$. Let $f(t|x)$ denote the conditional density of a continuous T given $X = x$. The conditional cumulative hazard function and odds ratio, respectively, $\Lambda(t|x) = -\log S(t|x)$ and $O(t|x) = F(t|x)/S(t|x)$. We will consider the general situation where the event time is subject to interval

censoring. The observed data are (X, Y), where $Y = (Y_1, Y_2]$, $0 \le Y_1 \le Y_2 \le \infty$. The reader is referred to Huang and Wellner (1997) for a review and more references about interval censoring. Special cases are right-censoring $Y_2 = \infty$, left-censoring $Y_1 = 0$, and current status data.

Let $f_0(\cdot) = f(\cdot \mid x_0)$, $F_0(\cdot) = F(\cdot \mid x_0)$, and $S_0(\cdot) = S(\cdot \mid x_0)$ be the baseline density, cumulative distribution, and survival functions, respectively. If the baseline density f_0 has support or can be truncated by $[0, \tau]$ then

$$f_0(t) \approx f_m(t; \boldsymbol{p}) = \frac{1}{\tau} \sum_{j=0}^{m} p_j \beta_{mj}\Big(\frac{t}{\tau}\Big), \quad t \in [0, \tau],$$

where $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{x}_0) = (p_0, \ldots, p_m)^\top$, $\sum_{j=0}^{m} p_j = 1$, $p_j \ge 0$, $\beta_{mj}(t) = (m+1)\binom{m}{j} t^j (1-t)^{m-j}$. Thus we have approximations

$$F_0(t) \approx F_m(t; \boldsymbol{p}) = \sum_{j=0}^{m} p_j B_{mj}\Big(\frac{t}{\tau}\Big), \quad t \in [0, \tau],$$

$$S_0(t) \approx S_m(t; \boldsymbol{p}) = \sum_{j=0}^{m} p_j \bar{B}_{mj}\Big(\frac{t}{\tau}\Big), \quad t \in [0, \tau],$$

where $B_{mj}(t) = \int_0^t \beta_{mj}(u)du$, and $\bar{B}_{mj}(t) = 1 - B_{mj}(t)$, $j = 0, \ldots, m$.

3.1 Accelerated Failure Time Model

The accelerated failure time (AFT) model can be specified as

$$f(t \mid x) = f(t \mid x; \gamma) = e^{\gamma^T x} f(te^{\gamma^T x} \mid 0), \quad t \in [0, \infty), \tag{1}$$

where $\gamma \in \mathbb{G} \subset \mathbb{R}^d$. Let $\gamma_0 \in \mathbb{G}$ be the true value of γ. The AFT model (1) is equivalent to

$$S(t \mid x; \gamma) = S(te^{\gamma^T x} \mid 0), \quad t \in [0, \infty).$$

Thus this is actually a scale regression model. The AFT model can also be written as linear regression $\log(T) = -\gamma^T x + \varepsilon$. It is clear that one can choose any x_0 in $\mathcal{X}$ as baseline by transform $\tilde{x} = x - x_0$. If $f_0(t) = f(t \mid x_0)$ has support $[0, \tau)$, $\tau \le \infty$, then $f(t \mid x)$ has support $[0, \tau e^{-\gamma_0^T \tilde{x}})$. The above AFT model can also be written as

$$f(t \mid x; \gamma) = e^{\gamma^T \tilde{x}} f_0(te^{\gamma^T \tilde{x}}), \quad S(t \mid x; \gamma) = S_0(te^{\gamma^T \tilde{x}}),$$

where $f_0(t) = f(t \mid x_0)$ and $S_0(t) = S(t \mid x_0) = \int_t^\infty f_0(u)du$.

We choose $\tau > y_{(n)} = \max\{y_{i1}, y_{j2} : y_{j2} < \infty;\, i, j = 1, \ldots, n\}$ so that $S(\tau)$ and $\max_{x \in \mathcal{X}} S(\tau \mid x)$ are believed very small. Guan (2023) proposed to approximate $f_0(t)$ and $S_0(t)$ on $[0, \tau]$, respectively, by

$$f_m(t; p) = \frac{1}{\tau} \sum_{j=0}^{m} p_j \beta_{mj}\Big(\frac{t}{\tau}\Big), \quad t \in [0, \tau];$$

$$S_m(t; p) = \sum_{j=0}^{m} p_j \bar{B}_{mj}\Big(\frac{t}{\tau}\Big), \quad t \in [0, \tau].$$

Then $f(t \mid x; \gamma)$ and $S(t \mid x; \gamma)$ can be approximated, respectively, by

$$\begin{aligned} f_m(t \mid x; \gamma, p) &= e^{\gamma^T x} f_m\Big(te^{\gamma^T x}; p\Big) \\ &= \frac{e^{\gamma^T x}}{\tau_0} \sum_{j=0}^{m} p_j \beta_{mj}\Big(e^{\gamma^T x} \frac{t}{\tau_0}\Big), \quad t \in [0, \tau_0 e^{-\gamma^T x}]; \end{aligned} \tag{2}$$

$$\begin{aligned} S_m(t \mid x; \gamma, p) &= S_m\Big(te^{\gamma^T x}; p\Big) \\ &= \sum_{j=0}^{m} p_j \bar{B}_{mj}\Big(e^{\gamma^T x} \frac{t}{\tau_0}\Big), \quad t \in [0, \tau_0 e^{-\gamma^T x}]. \end{aligned} \tag{3}$$

Guan (2023)'s proposal is implemented by functions `mable.aft()` and `maple.aft()`.

3.1.1 Example: Breast Cosmesis Data

```
library(mable)
g <- 0.41 # Hanson and Johnson 2004, JCGS
aft.res<-mable.aft(cbind(left, right) ~ treat, data = cosmesis, M =c(1, 30), g=.41,
                   tau =100, x0=data.frame(treat = "RCT"))
```

```
summary(aft.res)
```

```
## Call: mable.aft(cbind(left, right) ~ treat)
## Data: cosmesis
## Obj Class: mable_reg
## Dimension of response: 1
## Dimension of covariate: 1
## Optimal degree m: 6
## P-value: 0.008966527
## Maximum loglikelihood: -143.1529
## MABLE of p: can be retrieved using name 'p'
##
##            Estimate  Std.Error    z value   Pr(>|z|) Baseline x0
## treatRCT 0.57792280 0.13888899 4.16104112 0.00003168           1
##
## Note: the optimal model degree is selected by the method of change-point.
```

```
op <- par(mfrow = c(1,2), lwd = 1.5)
plot(x = aft.res, which = "likelihood")
plot(x = aft.res, y = data.frame(treat = "RT"), which = "survival",
     type = "l", col = 1, main = "Survival Function")
plot(x = aft.res, y = data.frame(treat = "RCT"), which = "survival",
     lty = 2, col = 1, add = TRUE)
legend("topright", bty = "n", lty = 1:2, col = 1, c("Radiation Only",
    "Radiation & Chemotherapy"), cex = .7)
```

```
par(op)
```

Alternatively we can use `mable.reg()`.

```
aft.res1 <- mable.reg(cbind(left, right) ~ treat, data = cosmesis, model='aft', M = c(1, 30),
      g=.41, tau=100, x0=data.frame(treat = "RCT"))
```

3.2 Proportional Hazards Model

Consider the Cox proportional hazard regression model (Cox 1972)

$$S(t|x) = S(t|x;\gamma, f_0) = S_0(t)^{\exp(\gamma^T \tilde{x})}, \tag{4}$$

where $\gamma \in \mathbb{G} \subset \mathbb{R}^d$, $\tilde{x} = x - x_0$, x_0 is any fixed covariate value, $f_0(\cdot) = f(\cdot|x_0)$ is the unknown baseline density and $S_0(t) = \int_t^\infty f_0(s)ds$. Define $\tau = \inf\{t : F(t|x_0) = 1\}$. It is true that τ is independent of x_0,

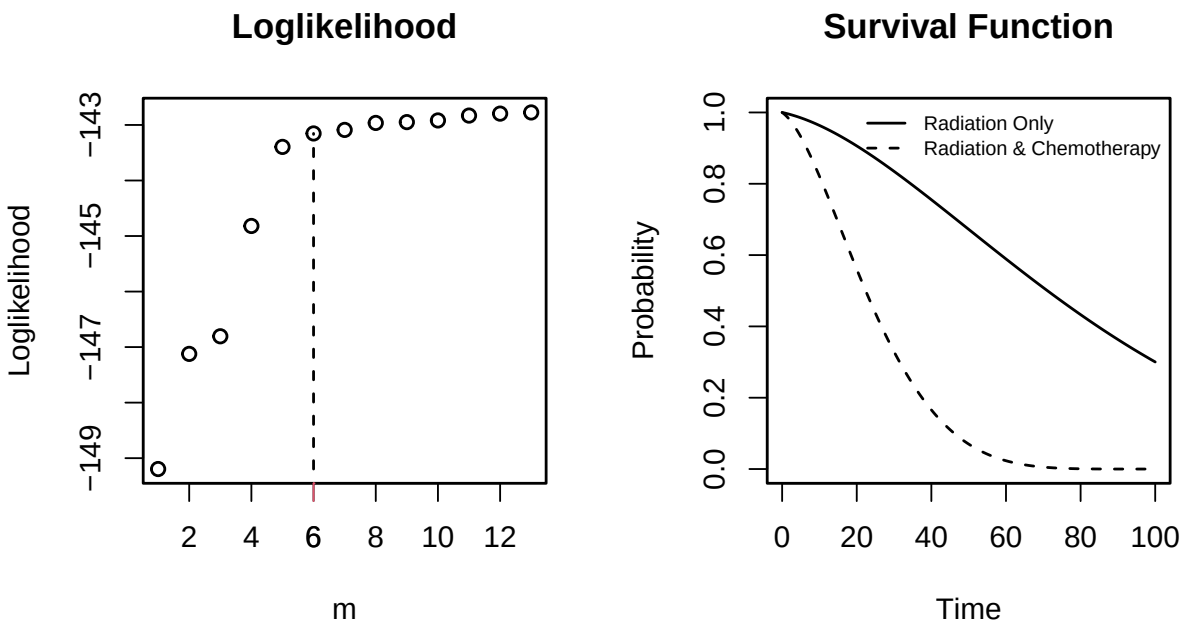

Figure 10: AFT Model Fit for Breast Cosmesis Data

$0 < \tau \leq \infty$, and $f(t|x)$ have the same support $[0, \tau]$ for all $x \in \mathcal{X}$. Let $(x_i, y_i) = (x_i, (y_{i1}, y_{j2}])$, $i = 1, \ldots, n$, be independent observations of (X, Y) and $\tau_n \geq y_{(n)} = \max\{y_{i1}, y_{j2} : y_{j2} < \infty; i, j = 1, \ldots, n\}$. Given any x_0, denote $\pi = \pi(x_0) = 1 - S_0(\tau_n)$. For integer $m \geq 1$ we define $\mathbb{S}_m \equiv \{(u_0, \ldots, u_m)^T \in \mathbb{R}^{m+1} : u_i \geq 0, \sum_{i=0}^m u_i = 1.\}$. Guan (2021b) propose to approximate $f_0(t)$ on $[0, \tau_n]$ by $f_m(t; p) = \tau_n^{-1} \sum_{i=0}^m p_i \beta_{mi}(t/\tau_n)$, where $p = p(x_0) = (p_0, \ldots, p_{m+1})$ are subject to constraints $p \in \mathbb{S}_{m+1}$ and $p_{m+1} = 1 - \pi$. Here the dependence of π and p on x_0 will be suppressed. If $\pi < 1$, although we cannot estimate the values of $f_0(t)$ on (τ_n, ∞), we can put an arbitrary guess on them such as $f_m(t; p) = p_{m+1}\alpha(t - \tau_n)$, $t \in (\tau_n, \infty)$, where $\alpha(\cdot)$ is a density on $[0, \infty)$ such that $(1 - \pi)\alpha(0) = (m + 1)p_m/\tau_n$ so that $f_m(t; p)$ is continuous at $t = \tau_n$, e.g., $\alpha(t) = \alpha(0) \exp[-\alpha(0)t]$. If τ is finite and known we choose $\tau_n = \tau$ and specify $p_{m+1} = 0$. Otherwise, we choose $\tau_n = y_{(n)}$. For data without right-censoring or covariate we also have to specify $p_{m+1} = 0$ due to its unidentifiability.

The above method is implemented in function `mable.ph()` which returns maximum approximate Bernstein likelihood estimates of (γ, p) with an optimal model degree m and a prespecified m, respectively. With a efficient estimate of γ obtained using other method, `maple.ph()` can be used to get an optimal degree m and a mable of p.

The `plot` method `plot.mable_reg()` for class `mable_reg` object returned by all the above functions produces graphs of the loglikelihoods at m in a set `M[1]:M[2]` of consecutive candidate model degrees, the likelihood ratios of change-point at m in `(M[1]+1):M[2]`, estimated density and survival function on the truncated `support`$=[0, \tau_n]$.

3.2.1 Example: Ovarian Cancer Survival Data

The Ovarian Cancer Survival Data is included in package `survival`.

```
library(survival)
futime2 <- ovarian$futime
```

```
futime2[ovarian$fustat==0] <- Inf
ovarian2 <- data.frame(age = ovarian$age, futime1 = ovarian$futime, futime2 = futime2)
head(ovarian2, 3)
```

```
##       age futime1 futime2
## 1 72.3315      59      59
## 2 74.4932     115     115
## 3 66.4658     156     156
```

```
ova<-mable.ph(cbind(futime1, futime2) ~ age, data = ovarian2, M = c(2,35),
              g = .16, x0=data.frame(age=35))
```

```
summary(ova)
```

```
## Call: mable.ph(cbind(futime1, futime2) ~ age)
## Data: ovarian2
## Obj Class: mable_reg
## Dimension of response: 1
## Dimension of covariate: 1
## Optimal degree m: 23
## P-value: 0.008999588
## Maximum loglikelihood: -85.73586
## MABLE of p: can be retrieved using name 'p'
##
##       Estimate  Std.Error    z value   Pr(>|z|) Baseline x0
## age 1.7669e-01 1.0517e-02 1.6800e+01 2.4435e-63          35
##
## Note: the optimal model degree is selected by the method of change-point.
```

```
op <- par(mfrow = c(2,2))
plot(ova, which = "likelihood")
plot(ova, which = "change-point")
plot(ova, y=data.frame(age=60), which="survival", type="l", xlab="Days", main="Age = 60")
plot(ova, y=data.frame(age=65), which="survival", type="l", xlab="Days", main="Age = 65")
```

```
par(op)
```

Alternatively we can use `mable.reg()`.

```
ova1 <- mable.reg(cbind(futime1, futime2) ~ age, data = ovarian2, M = c(2,35))
```

3.3 Proportional Odds Model

As an important alternative of the Cox proportional hazards regression model (Cox 1972) the proportional odds (PO) model (McCullagh 1980; Bennett 1983a, 1983b; Pettitt 1984) is defined by

$$\frac{O(t|\boldsymbol{x})}{O(t|\boldsymbol{x}_0)} = e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} \tag{5}$$

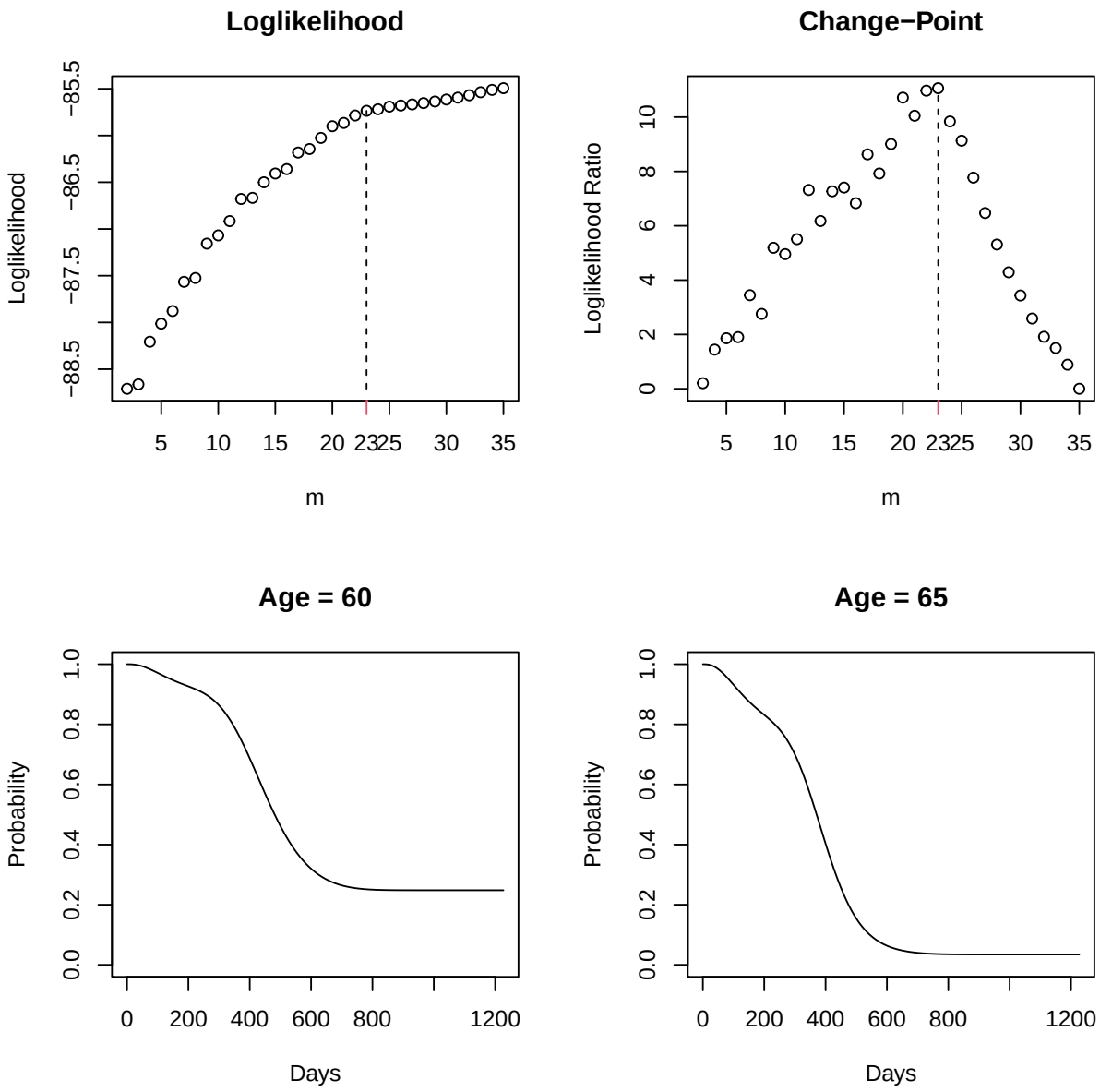

Figure 11: Ovarian Cancer Data

where $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{x}_0$ and $O(t|\boldsymbol{x}_0)$ is an unknown increasing baseline odds function on $(0, \infty)$ and $\boldsymbol{\gamma}$ are unknown regression coefficients. If the baseline density f_0 has support or can be truncated by $[0, \tau]$ then

$$f_0(t) \approx f_m(t; \boldsymbol{p}) = \frac{1}{\tau} \sum_{j=0}^{m} p_j \beta_{mj}\Big(\frac{t}{\tau}\Big), \quad t \in [0, \tau].$$

Therefore $f(t \mid \boldsymbol{x}; \boldsymbol{\gamma}, f_0)$, $F(t \mid \boldsymbol{x}; \boldsymbol{\gamma}, f_0)$, and $S(t \mid \boldsymbol{x}; \boldsymbol{\gamma}, f_0)$ can be approximated, respectively, by

$$\begin{aligned}
f_m(t \mid \boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{p}) &= \frac{e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} f_m(t; \boldsymbol{p})}{[e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} + (1 - e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}}) S_m(t; \boldsymbol{p})]^2} = \frac{e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} f_m(t; \boldsymbol{p})}{[1 + (e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} - 1) F_m(t; \boldsymbol{p})]^2}, \\
F_m(t|\boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{p}) &= \frac{e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} F_m(t; \boldsymbol{p})}{e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} + (1 - e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}}) S_m(t; \boldsymbol{p})} = \frac{e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} F_m(t; \boldsymbol{p})}{1 + (e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} - 1) F_m(t; \boldsymbol{p})}, \\
S_m(t \mid \boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{p}) &= \frac{S_m(t; \boldsymbol{p})}{e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} + (1 - e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}}) S_m(t; \boldsymbol{p})} = \frac{S_m(t; \boldsymbol{p})}{1 + (e^{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{x}}} - 1) F_m(t; \boldsymbol{p})}.
\end{aligned}$$

3.3.1 Example: HIV Infection time Data

The data set is contained in R package `Epi`. This is an interval-censored data. The semiparametric methods (package `icenReg`) are available for PH and PO (but not AFT) models based on interval-censored data. The estimated survival functions are plotted in Fig 12.

```
## Interval-censored data on times to HIV infection. Carstensen (1996)
## see also Lawless and Babineau (2006)
# inspection times(months): started in Dec. 31, 1980,
it<-c(0, 12, 16, 27, 45, 76, 101, Inf)
# observed frequencies d[l,r] in interval (it[l], it[r]], 0<=l<r<=7
x<-c(24, 0, 4, 2,1,4,0,10,0,3,0,3,1,0,5,0,4,0,2,1,1,0,61,8,15,22,34,92)
n<-sum(x)
d<-matrix(c(24,rep(0,6),
            0, 4, rep(0,5),
            2,1,4,0,0,0,0,
            0,10,0,3,0,0,0,
            0,3,1,0,5,0,0,
            0,4,0,2,1,1,0,
            0,61,8,15,22,34,92), nrow=7, ncol=7)
# See Fay's tutorial
#  us: whether visited the US.
# byr: birth year
# age: age at entry
# age=hiv$bth+30,
# pyr: Annual number of sexual partners.
library(Epi)
data(hivDK)
l<-as.numeric(hivDK$well-hivDK$entry)
l.na<-l
l[is.na(l)]<-0
r<- as.numeric(hivDK$ill - hivDK$entry)
r.na<-r
r[is.na(r)]<-Inf
hiv<-data.frame(l,r, us=hivDK$us, byr=hivDK$bth+1950, age=30-hivDK$bth, pyr=hivDK$pyr)
## icenReg
library(icenReg)
```

```r
#library(interval)
phfit<-ic_sp(cbind(l,r)~us+age+pyr, data=hiv, bs_samples =100, model = 'ph')
pofit<-ic_sp(cbind(l,r)~us+age+pyr, data=hiv, bs_samples =100, model = 'po')
```

```r
# MABLE
require(mable)
gama<-phfit$coefficients
tau<-5500 # truncation interval [0,tau]
# Mable PH model
x0<-data.frame(us=0,age=72,pyr=0)
#mblph<-mable.ph(cbind(l,r)~us+age+pyr, data=hiv, M=c(2,40), g=gama, x0=x0, tau=tau)
mblph<-mable.ph(cbind(l,r)~us+age+pyr, data=hiv, M=22, g=gama, x0=x0, tau=tau)
# Mable PO model:
mblpo<-mable.po(cbind(l,r)~us+age+pyr, data=hiv, M=c(3,20), g=gama, x0=x0, tau=tau)
# Mable AFT model: m=12 selected from M=c(5,50)
mblaft<-mable.aft(cbind(l,r)~us+age+pyr, data=hiv, M=12, g=NULL, x0=x0, tau=tau,
                  controls=mable.ctrl(sig.level=0.05, eps.em=1e-7, maxit.em=20000))
```

```r
newdata=data.frame(us=0:1,age=mean(hiv$age),pyr=mean(hiv$pyr))
par(mfrow=c(2,2),lwd=1.5, cex=.7, mar=c(4,4,1,1), oma=c(0,0,3,1))
# upper-left panel
plot(c(0,3057), c(.6,1), type="n", xlab="Days", ylab="Probability")
lines(phfit, y=newdata[1,], lty=3, col="brown")
lines(phfit, y=newdata[2,], lty=4, col=4)
plot(mblpo, y=newdata[1,], which="survival", add=TRUE, col=1, lty=1)
plot(mblpo, y=newdata[2,], which="survival", add=TRUE, col=2, lty=2)
legend("topright", lty=c(1:4), col=c(1:2,"brown",4), bty="n",
       legend=paste0(c("mable-po: us = ", "mable-po: us = ",
                       "semip-ph: us = ", "semip-ph: us = "),  c(rep(newdata$us,2))))
# upper-right panel
plot(c(0,3057), c(.6,1), type="n", xlab="Days", ylab="Probability")
lines(pofit, y=newdata[1,], lty=3, col="brown")
lines(pofit, y=newdata[2,], lty=4, col=4)
plot(mblpo, y=newdata[1,], which="survival", add=TRUE, col=1, lty=1)
plot(mblpo, y=newdata[2,], which="survival", add=TRUE, col=2, lty=2)
legend("topright", lty=c(1:4), col=c(1:2,"brown",4), bty="n",
       legend=paste0(c("mable-po: us = ", "mable-po: us = ",
                       "semip-po: us = ", "semip-po: us = "),  c(rep(newdata$us,2))))
# lower-left panel
#newdata2=data.frame(us=mean(hiv$us),age=c(31,50),pyr=mean(hiv$pyr))
plot(c(0,3057), c(.6,1), type="n", xlab="Days", ylab="Probability")
plot(mblpo, y=newdata[1,], which="survival", add=TRUE, col=1, lty=1)
plot(mblpo, y=newdata[2,], which="survival", add=TRUE, col=2, lty=2)
plot(mblph, y=newdata[1,], which="survival", add=TRUE, col="brown", lty=3)
plot(mblph, y=newdata[2,], which="survival", add=TRUE, col=4, lty=4)
legend("topright", lty=1:4, col=c(1,2,"brown",4), bty="n",
       legend=paste0(c("mable-po: us = ", "mable-po: us = ",
                       "mable-ph: us = ", "mable-ph: us = "),  c(rep(newdata$us,2))))
# lower-right panel
plot(c(0,3057), c(.6,1), type="n", xlab="Days", ylab="Probability")
plot(mblpo, y=newdata[1,], which="survival", add=TRUE, col=1, lty=1)
plot(mblpo, y=newdata[2,], which="survival", add=TRUE, col=2, lty=2)
plot(mblaft, y=newdata[1,], which="survival", add=TRUE, col="brown", lty=3)
```

```
plot(mblaft, y=newdata[2,], which="survival", add=TRUE, col=4, lty=4)
legend("topright", lty=1:4, col=c(1,2,"brown",4), bty="n",
       legend=paste0(c("mable-po: us = ", "mable-po: us = ",
                       "mable-aft: us = ", "mable-aft: us = "),  c(rep(newdata$us,2))))
mtext("Survival Curves: Visited US or Not", side=3, line=1, outer=TRUE, cex=1, font=2)
```

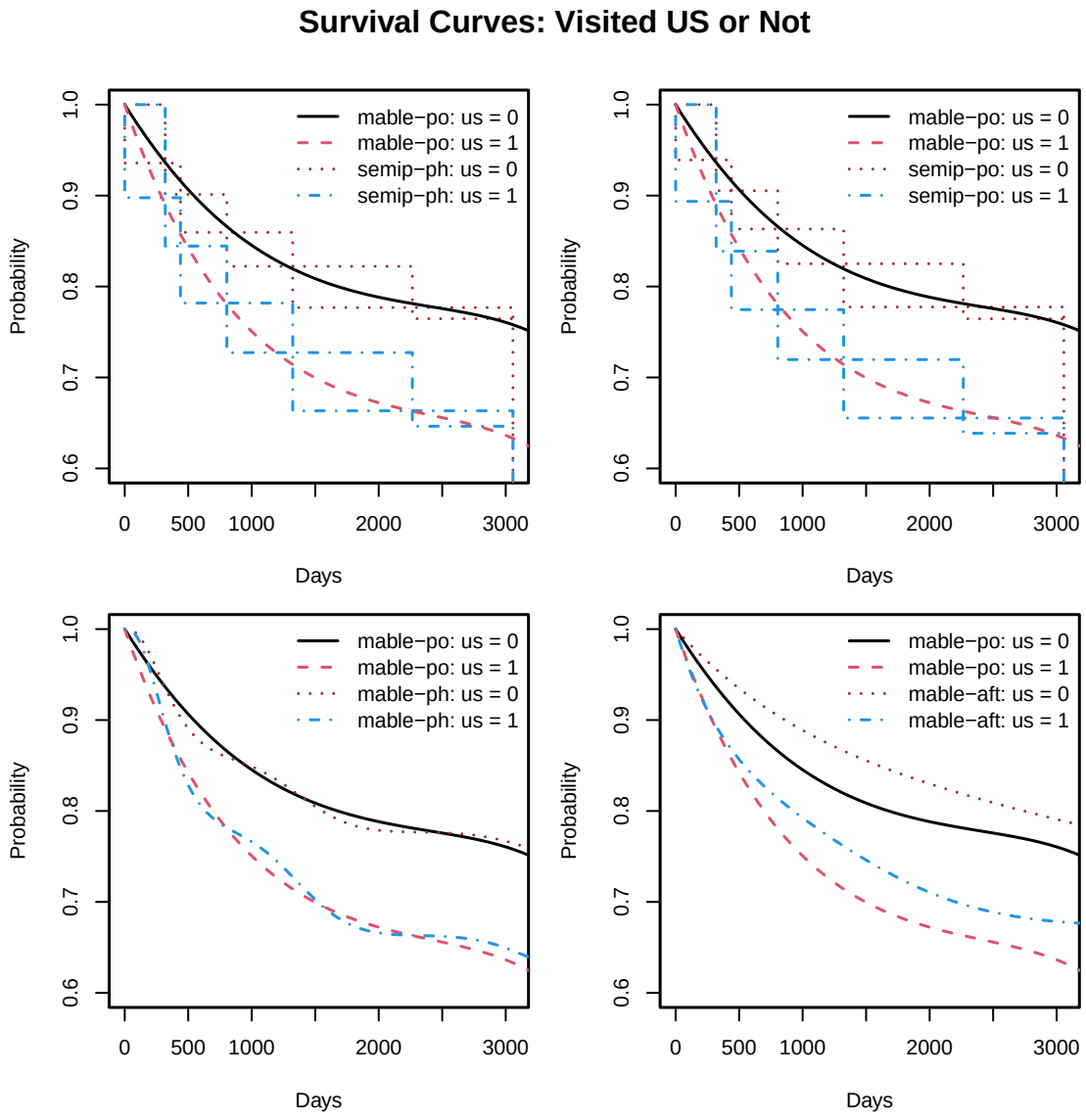

Figure 12: Estimated survival functions for HIV infection time data with `us`=0 and `us`=1, and average values of `age` and `pyr`. MABLE PO model is compared with semiparametric PH model (upper-left), semiparametric PO model (upper-right), and MABLE AFT model (lower-right).

4 Two-sample Data

In addition to the AFT, PH and PO regression models with single binary covariate, other two-sample semiparametric models can be estimated using the maximum approximate Bernstein/Beta likelihood method.

4.1 Density Ratio Model

Two-sample density ratio(DR) model (see for example, Qin and Zhang (1997) and Qin and Zhang (2005)) is useful for fitting case-control data. Suppose that the densities f_0 and f_1 of X_0 and X_1, respectively, satisfy the following density ratio model (see Cheng and Chu (2004) and Qin and Zhang (2005), for example)

$$f_1(x) = f(x;\boldsymbol{\alpha}) = f_0(x)\exp\{\boldsymbol{\alpha}^T\tilde{r}(x)\}, \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_0,\ldots,\alpha_d)^T \in \mathcal{A} \subset R^{d+1}$, and $\tilde{r}(x) = (1, r^T(x))^T$ with linearly independent components. Guan (2021c) proposed to approximate the *baseline* density f_0 by Bernstein polynomial and to estimate f_0 and $\boldsymbol{\alpha}$ using the maximum approximate Bernstein/Beta likelihood estimation. This method is implemented by function `mable.dr()` and `maple.dr()` in which $\boldsymbol{\alpha}$ is a given estimated value such as the one obtained by the logistic regression as described in Qin and Zhang (1997) and Qin and Zhang (2005).

4.1.1 Example: Coronary Heart Disease Data

```
# Hosmer and Lemeshow (1989):
# ages and the status of coronary disease (CHD) of 100 subjects
x<-c(20, 23, 24, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 32,
32, 33, 33, 34, 34, 34, 34, 35, 35, 36, 36, 37, 37, 38, 38, 39,
40, 41, 41, 42, 42, 42, 43, 43, 44, 44, 45, 46, 47, 47, 48, 49,
49, 50, 51, 52, 55, 57, 57, 58, 60, 64)
y<-c(25, 30, 34, 36, 37, 39, 40, 42, 43, 44, 44, 45, 46, 47, 48,
48, 49, 50, 52, 53, 53, 54, 55, 55, 56, 56, 56, 57, 57, 57, 57,
58, 58, 59, 59, 60, 61, 62, 62, 63, 64, 65, 69)
a<-20; b<-70
regr<-function(x) cbind(1,x)
```

```
chd.mable<-mable.dr(x, y, M=c(1, 15), regr, interval = c(a,b))
```

```
z<-seq(a,b,length=512)
f0hat<-dmixbeta(z, p=chd.mable$p, interval=c(a, b))
rf<-function(x) regr((x-a)/(b-a))
f1hat<-dtmixbeta(z, p=chd.mable$p, alpha=chd.mable$alpha,
                 interval=c(a, b), regr=regr)
op<-par(mfrow=c(1,2),lwd=1.2, cex=.7, mar=c(5,4,1,1))
hist(x, freq=F, col = "light grey", border = "white", xlab="Age",
  ylab="Density", xlim=c(a,b), ylim=c(0,.055), main="Control")
lines(z, f0hat, lty=1, col=1)
hist(y, freq=F, col = "light grey", border = "white", xlab="Age",
  ylab="Density", xlim=c(a,b), ylim=c(0,.055), main="Case")
lines(z, f1hat, lty=1, col=1)
```

```
par(op)
```

4.1.2 Example: Pancreatic Cancer Biomarker Data

For the logarithmic levels of CA 19-9 of the Pancreatic Cancer Data (Wieand et al. (1989)), the control group is used as baseline because the optimal model degree is smaller that the one using case as baseline.

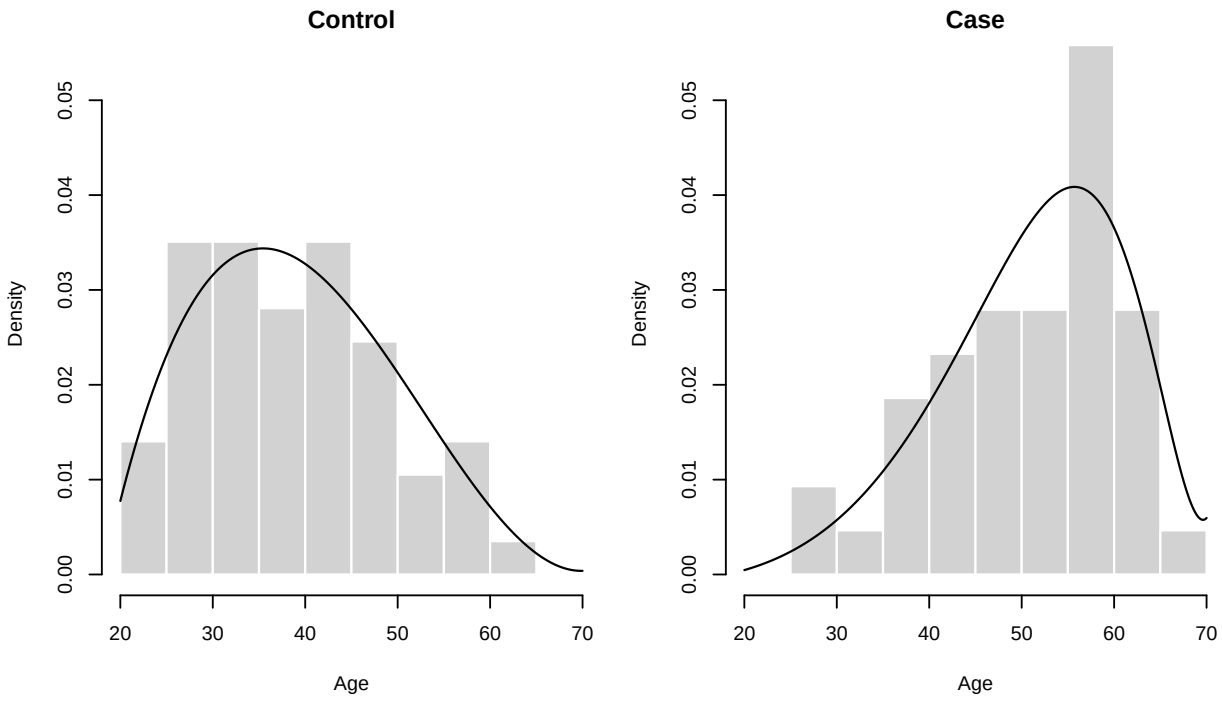

Figure 13: DR Model Fit for Coronary Heart Disease Data

```
data(pancreas)
head(pancreas,3)
```

```
##   ca199 ca125 status
## 1  28.0  13.3      0
## 2  15.5  11.1      0
## 3   8.2  16.7      0
```

```
x<-log(pancreas$ca199[pancreas$status==0])
y<-log(pancreas$ca199[pancreas$status==1])
a<-min(x,y); b<-max(x,y)
M<-c(1,29)
regr<-function(x) cbind(1,x,x^2)
```

```
m=maple.dr(x, y, M, regr=regr, interval=c(a,b), controls=mable.ctrl(sig.level=.001))$m
pc.mable<-mable.dr(x, y, M=m, regr=regr, interval=c(a,b),
                   controls=mable.ctrl(sig.level=1/length(c(x,y))))
#pc.mable
```

```
z<-seq(a,b,length=512)
# baseline is "case"
f1hat<-dmixbeta(z, p=pc.mable$p, interval=c(a, b))
rf<-function(x) regr((x-a)/(b-a))
f0hat<-dtmixbeta(z, p=pc.mable$p, alpha=pc.mable$alpha,
                 interval=c(a, b), regr=regr)
op<-par(mfrow=c(1,2),lwd=1.2, cex=.7, mar=c(5,4,1,1))
```

```r
hist(x, freq=F, col = "light grey", border = "white", xlab="log(CA19-9)",
  ylab="Density", xlim=c(a,b),  main="Control")
lines(z, f0hat, lty=1, col=1)
hist(y, freq=F, col = "light grey", border = "white", xlab="log(CA19-9)",
  ylab="Density", xlim=c(a,b), ylim=c(0,.5), main="Case")
lines(z, f1hat, lty=1, col=1)
```

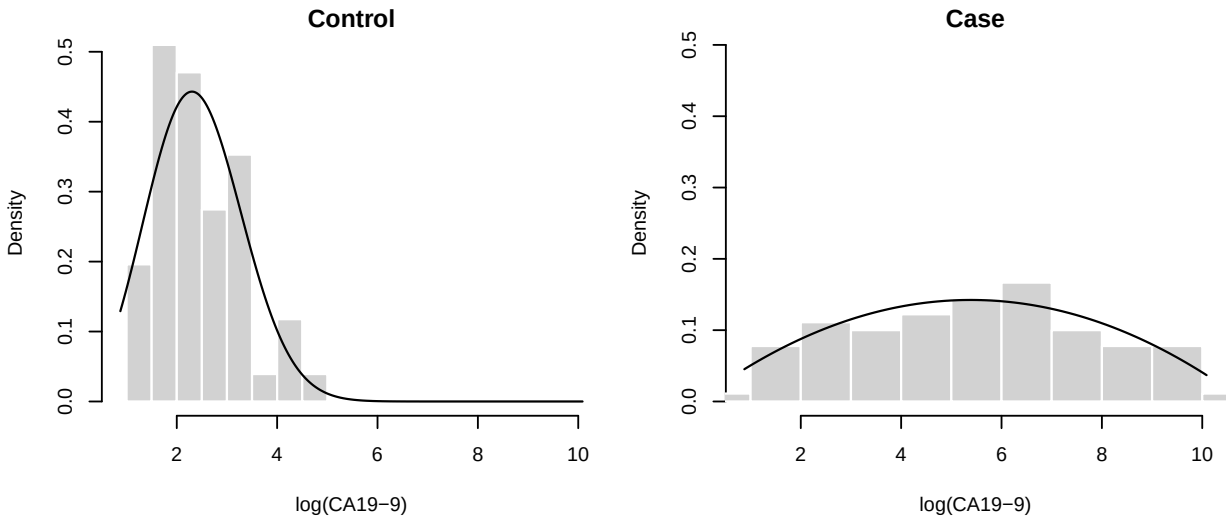

Figure 14: DR Model Fit for Pancreatic Cancer Biomarker Data

```r
par(op)
```

References

Bennett, Steve. 1983a. "Analysis of Survival Data by the Proportional Odds Model." *Statistics in Medicine* 2: 273–77.

———. 1983b. "Log-Logistic Regression Models for Survival Data." *Applied Statistics* 32 (2): 165–71.

Cheng, K. F., and C. K. Chu. 2004. "Semiparametric Density Estimation Under a Two-Sample Density Ratio Model." *Bernoulli* 10 (4): 583–604.

Cox, D. R. 1972. "Regression Models and Life-Tables." *J. Roy. Statist. Soc. Ser. B* 34: 187–220.

Finkelstein, Dianne M., and Robert A. Wolfe. 1985. "A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data." *Biometrics* 41 (4): 933–45.

Guan, Zhong. 2016. "Efficient and Robust Density Estimation Using Bernstein Type Polynomials." *Journal of Nonparametric Statistics* 28 (2): 250–71. https://doi.org/10.1080/10485252.2016.1163349.

———. 2017. "Bernstein Polynomial Model for Grouped Continuous Data." *Journal of Nonparametric Statistics* 29 (4): 831–48. https://doi.org/10.1080/10485252.2017.1374384.

———. 2021a. "Fast Nonparametric Maximum Likelihood Density Deconvolution Using Bernstein Polynomials." *Statistica Sinica* 31 (2): 891–908. https://doi.org/10.5705/ss.202018.0173.

———. 2021b. "Maximum Approximate Bernstein Likelihood Estimation in Proportional Hazard Model for Interval-Censored Data." *Statistics in Medicine* 40 (3): 758–78. https://doi.org/https://doi.org/10.1002/sim.8801.

———. 2021c. "Maximum Approximate Bernstein Likelihood Estimation of Densities in a Two-Sample Semiparametric Model." https://arxiv.org/abs/2103.00648.

———. 2023. "Maximum Approximate Likelihood Estimation in Accelerated Failure Time Model for Interval-Censored Data." *Statistics in Medicine* 42: 4886–96. https://doi.org/https://doi.org/10.1002/sim.9893.

Huang, J., and J. A. Wellner. 1997. "Interval Censored Survival Data: A Review of Recent Progress." In *Proceedings of the First Seattle Symposium in Biostatistics*, edited by Lin D. Y. and Fleming T. R., 123:123–69. Lecture Notes in Statistics. New York, NY: Springer.

Jassim, E. W., M. Grossman, W. J. Koops, and R. A. J. Luykx. 1996. "Multiphasic Analysis of Embryonic Mortality in Chickens." *Poultry Sci* 75: 464–71.

Kuurman, W. W., B. A. Bailey, W. J. Koops, and M. Grossman. 2003. "A Model for Failure of a Chicken Embryo to Survive Incubation." *Poultry Sci.* 82: 214–22.

Lin, Nan, and Xuming He. 2006. "Robust and Efficient Estimation Under Data Grouping." *Biometrika* 93 (1): 99–112.

Linhart, H., and W. Zucchini. 1986. *Model Selection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc.

McCullagh, Peter. 1980. "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society. Series B. Methodological* 42 (2): 109–42.

Pettitt, A. N. 1984. "Proportional Odds Model for Survival Data and Estimates Using Ranks." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 33 (2): 169–75.

Qin, Jing, and Biao Zhang. 1997. "A Goodness-of-Fit Test for Logistic Regression Models Based on Case-Control Data." *Biometrika* 84 (3): 609–18.

———. 2005. "Density Estimation Under a Two-Sample Semiparametric Model." *Journal of Nonparametric Statistics* 17 (6): 665–83. https://doi.org/10.1080/10485250500039346.

Wang, Tao, and Zhong Guan. 2019. "Bernstein Polynomial Model for Nonparametric Multivariate Density." *Statistics* 53 (2): 321–38. https://doi.org/10.1080/02331888.2019.1574299.

Wieand, Sam, Mitchell H. Gail, Barry R. James, and Kang L. James. 1989. "A Family of Nonparametric Statistics for Comparing Diagnostic Markers with Paired or Unpaired Data." *Biometrika* 76 (3): 585–92. https://doi.org/10.1093/biomet/76.3.585.