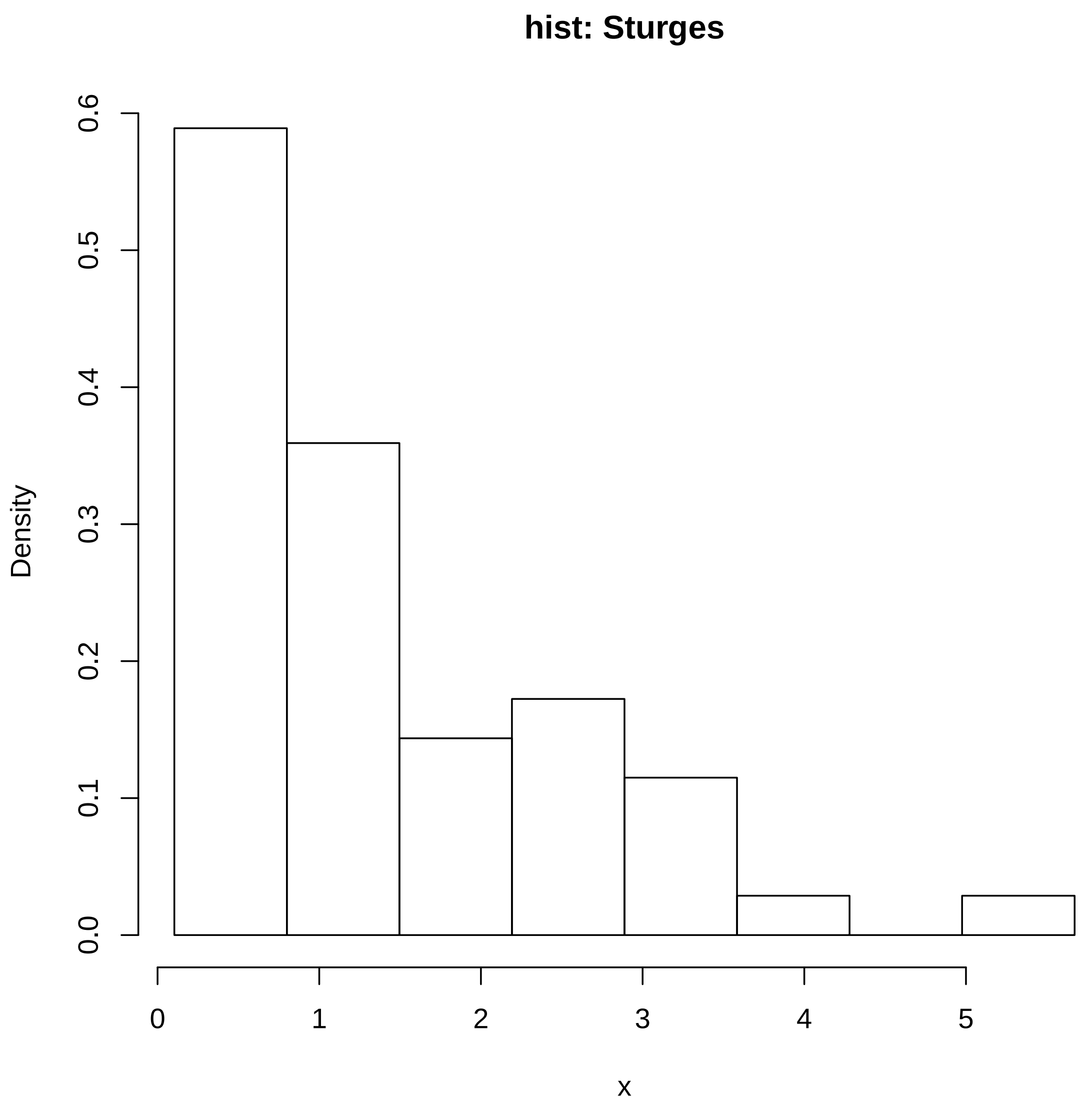
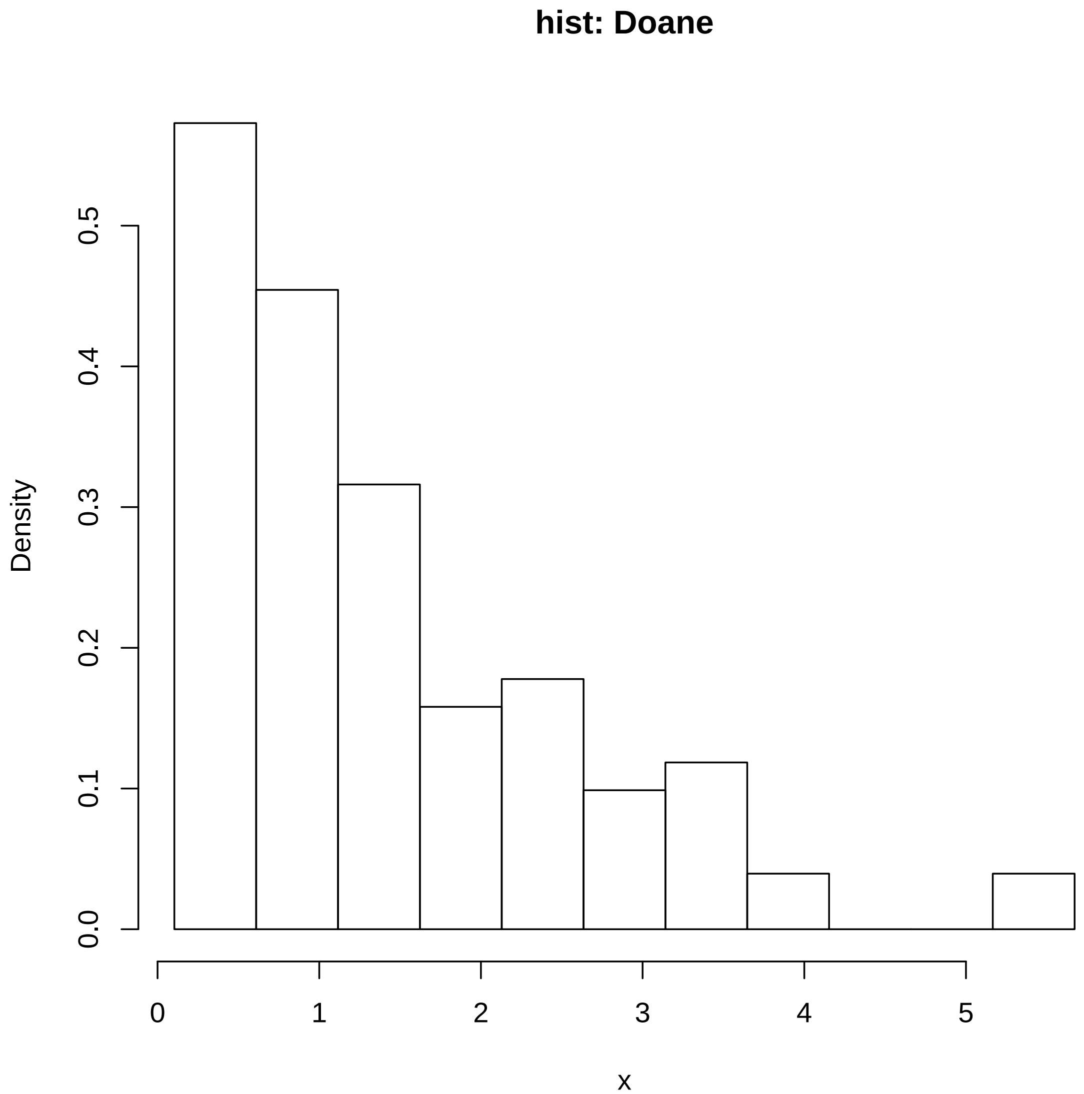
- 10.1 Histogram estimate of the density of a standard lognormal using Sturge's rule and Doane's correction for skewness.

  Using code adapted from the book example 10.1.

```
require(xtable)

## Loading required package:  xtable

set.seed(3)
n <- 100
x <- rlnorm(n)
# calc breaks according to Sturges' Rule
nclass <- ceiling(1 + log2(n))
cwidth <- diff(range(x)/nclass)
breaks <- min(x) + cwidth * 0:nclass

h.sturges <- hist(x, breaks = breaks, freq = FALSE, main = "hist: Sturges")
```




```r
# apply correction for skewness
require(moments)

## Loading required package:  moments

sk <- skewness(x)
sesk <- sqrt((6 * (n - 2))/((n + 1) * (n + 3)))
k <- log2(1 + sk/sesk)
nclass.d <- ceiling(1 + log2(n) + k)
cwidth.d <- diff(range(x)/nclass.d)
breaks.d <- min(x) + cwidth.d * 0:nclass.d
h.doane <- hist(x, breaks = breaks.d, freq = FALSE, main = "hist: Doane")
```




```r
# compare density to the theoretical values
lnorm.dens <- dlnorm(qlnorm(seq(0.1, 0.9, 0.1)))
compare.doane <- rbind(h.doane$breaks[which(h.doane$breaks <= max(qlnorm(seq(0.1,
    0.9, 0.1))))], h.doane$density[which(h.doane$breaks <= max(qlnorm(seq(0.1,
    0.9, 0.1))))], dlnorm(h.doane$breaks[which(h.doane$breaks <= max(qlnorm(seq(0.1,
    0.9, 0.1))))]))
compare.sturges <- rbind(h.sturges$breaks[which(h.sturges$breaks <= max(qlnorm(seq(0.1,
    0.9, 0.1))))], h.sturges$density[which(h.sturges$breaks <= max(qlnorm(seq(0.1,
    0.9, 0.1))))], dlnorm(h.sturges$breaks[which(h.sturges$breaks <= max(qlnorm(seq(0.1,
    0.9, 0.1))))]))
rownames(compare.doane) <- rownames(compare.sturges) <- c("Breaks within the first 90\\% of the the
    "Estimates Density", "Theoretical Density")
xtable(compare.sturges, caption = "Comparison of the Sturges density estimate with theoretical valu
```

	1	2	3	4	5	6
Breaks within the first 90\% of the theoretical density	0.10	0.80	1.50	2.19	2.89	3.58
Estimates Density	0.59	0.36	0.14	0.17	0.11	0.03
Theoretical Density	0.30	0.49	0.25	0.13	0.08	0.05

Table 1: Comparison of the Sturges density estimate with theoretical values

```
xtable(compare.doane, caption = "Comparison of the Doane corrected density estimates with the theor
```

	1	2	3	4	5	6	7
Breaks within the first 90\% of the theoretical density	0.10	0.61	1.12	1.62	2.13	2.63	3.14
Estimates Density	0.57	0.45	0.32	0.16	0.18	0.10	0.12
Theoretical Density	0.30	0.58	0.36	0.22	0.14	0.09	0.07

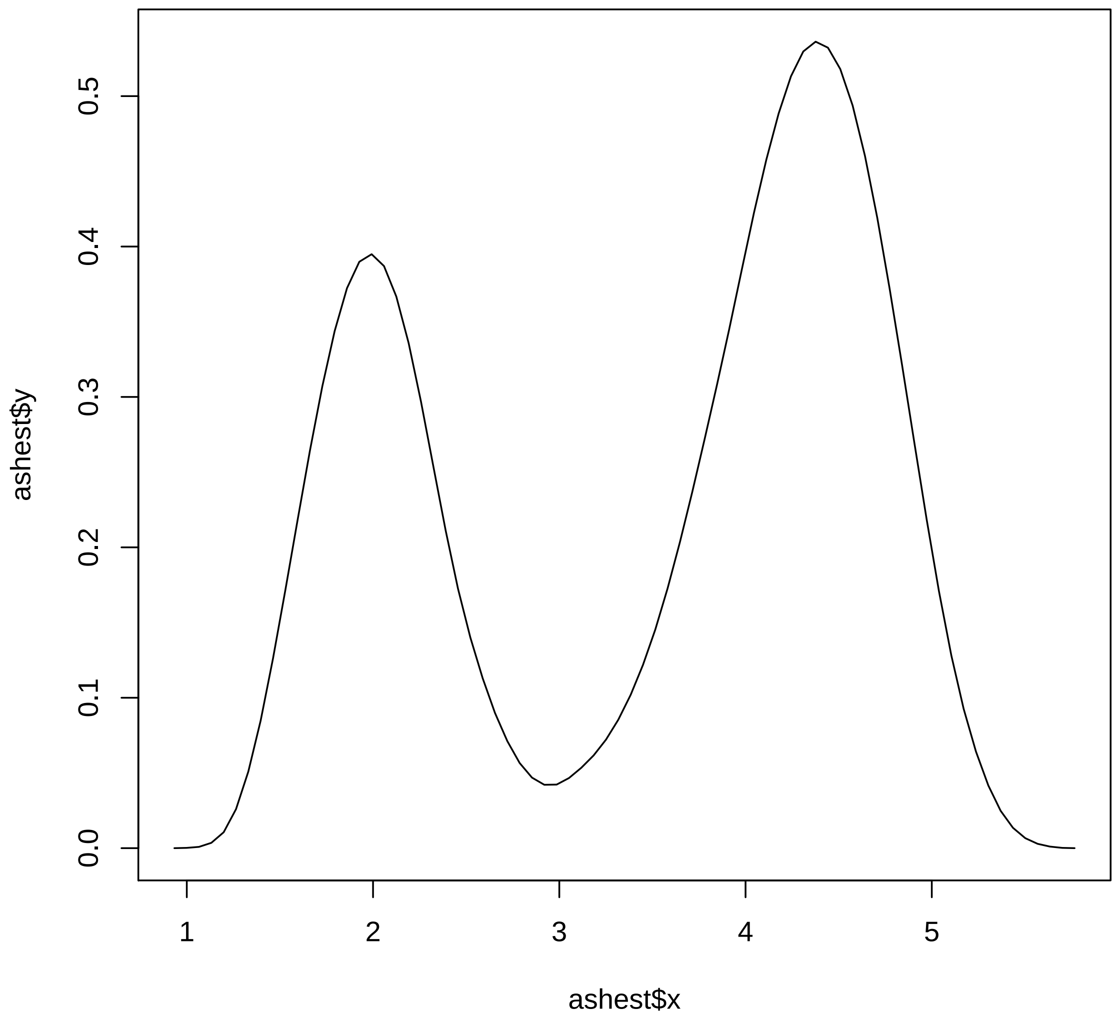
Table 2: Comparison of the Doane corrected density estimates with the theoretical values

The Doane correction adds 3 more bars (and breaks) than the Sturges histogram but only one of these was within the first 9 deciles. In general both methods did a poor job of estimating the density near 0 but got better further from zero. The Doane estimate did slightly better in genreral.

10.6 Normal reference rule ASH estimate of the eruptions density.

I will use the ash function since it seems well equiped for this sort of thing.

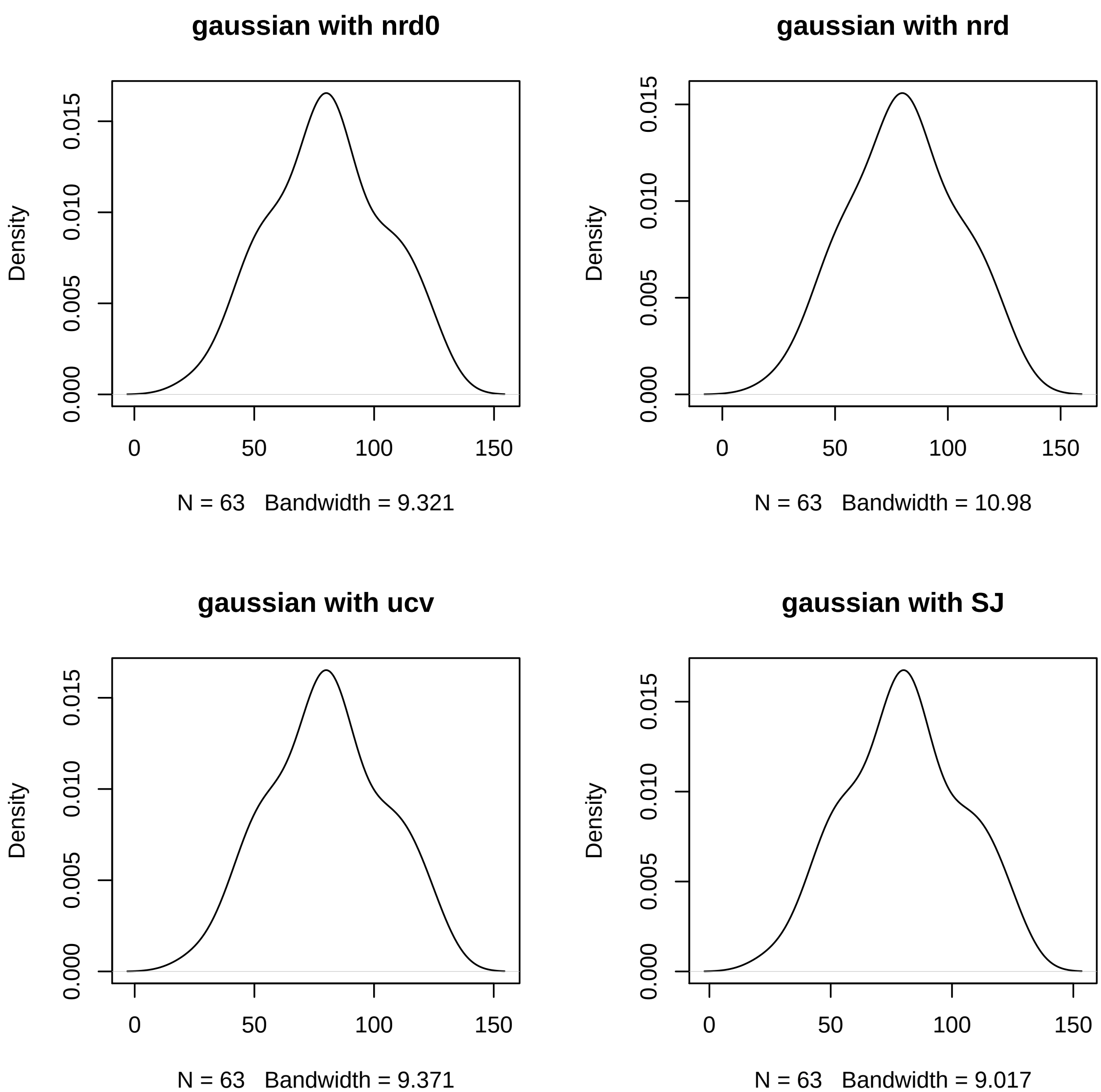
```
require(MASS)

## Loading required package:  MASS

require(ash)

## Loading required package:  ash

erupt <- faithful$eruptions
n <- length(erupt)
# scott's normal reference rule
h <- 2.576 * sd(erupt) * n^(-0.2)
R <- diff(range(erupt))
numbin <- ceiling(R * 20/h)
# set the interval by adding 20% of the range to either end of the
# interval
interval <- c(range(erupt)[1] - R * 0.2, range(erupt)[2] + R * 0.2)
bins <- bin1(erupt, nbin = numbin, ab = interval)
ashest <- ash1(bins = bins, m = 10, kopt = c(2, 2))
plot(ashest, type = "l")
```

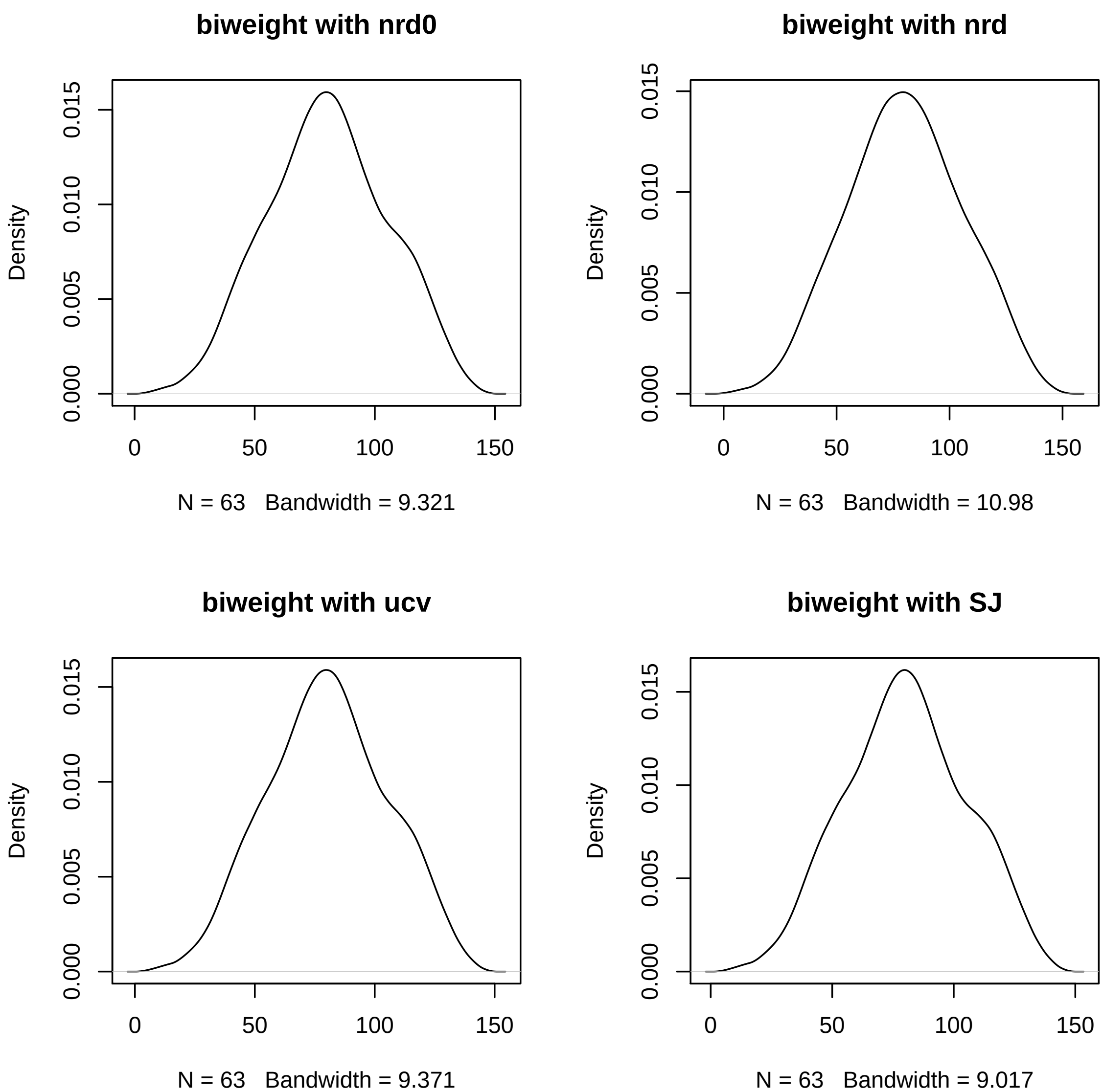


10.8 Kernel density estimates of the buffalo data.

```
require(gss)

## Loading required package:  gss

data(buffalo)
bws <- c("nrd0", "nrd", "ucv", "SJ")
kerns <- c("gaussian", "biweight")
for (i in 1:length(kerns)) {
    par(mfrow = c(2, 2))
    kern <- kerns[i]
    for (j in 1:length(bws)) {
```

```
        plot(density(buffalo, bw = paste0(bws[j]), kernel = paste0(kern)), main = paste0(kern,
            " with ", bws[j]))
    }
    # plot(truehist(buffalo))
    par(mfrow = c(1, 1))
}
```



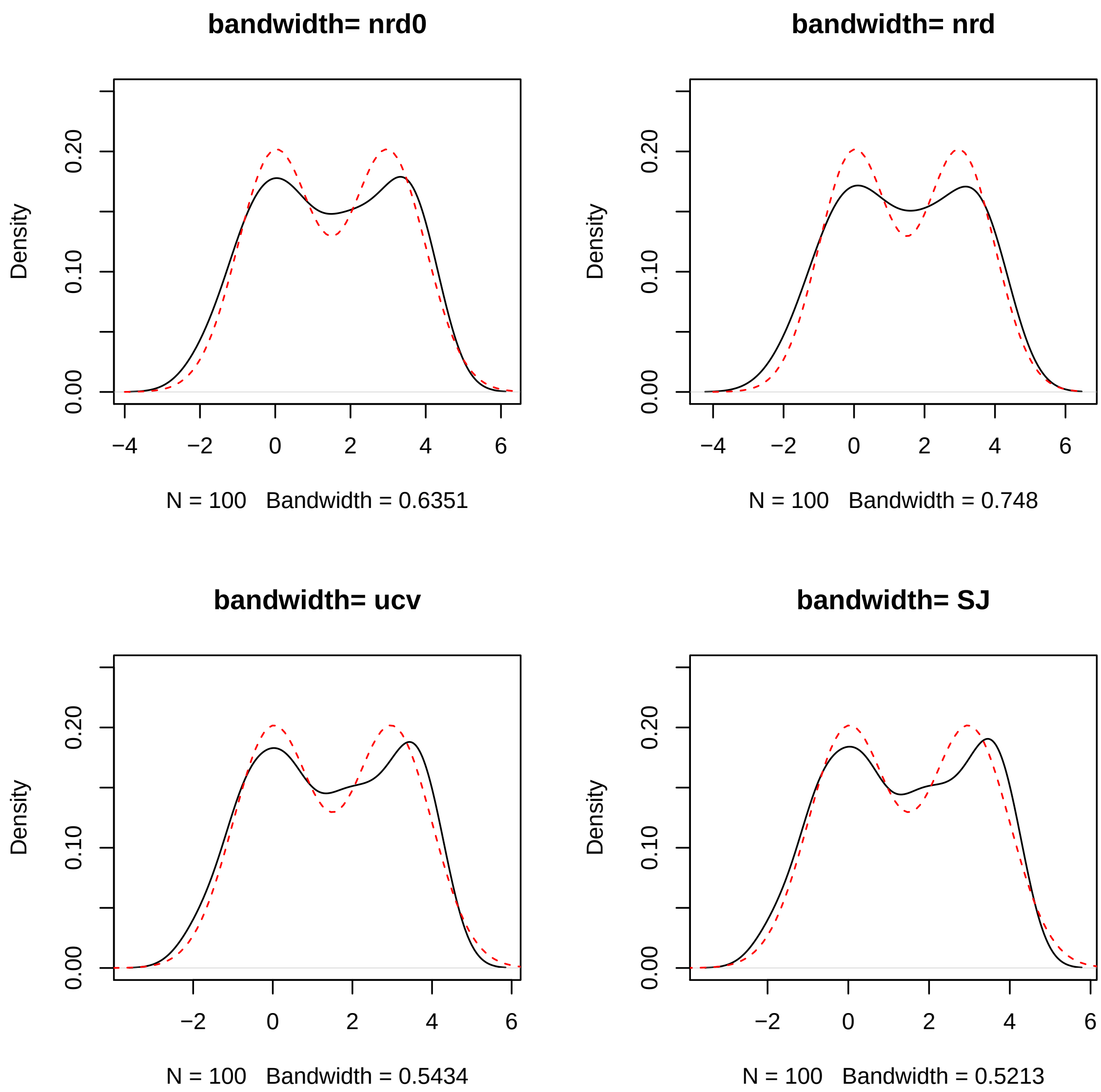


It appears that the bandwidth choice has more of an impact than the kernel choice.

10.9 KDE of a normal mixture.

```
set.seed(3)
n <- 100
mu <- sample(c(0, 3), size = n, replace = T)
x <- rnorm(n, mu)
par(mfrow = c(2, 2))
for (i in 1:length(bws)) {
    plot(density(x, bw = paste0(bws[i]), kernel = "gaussian"), main = paste0("bandwidth= ",
        bws[i]), ylim = c(0, 0.25))
```

```
    curve(0.5 * dnorm(x, 0, 1) + 0.5 * dnorm(x, 3, 1), xlim = c(-4, 6.5), col = "red",
        lty = 2, add = T)
}
```

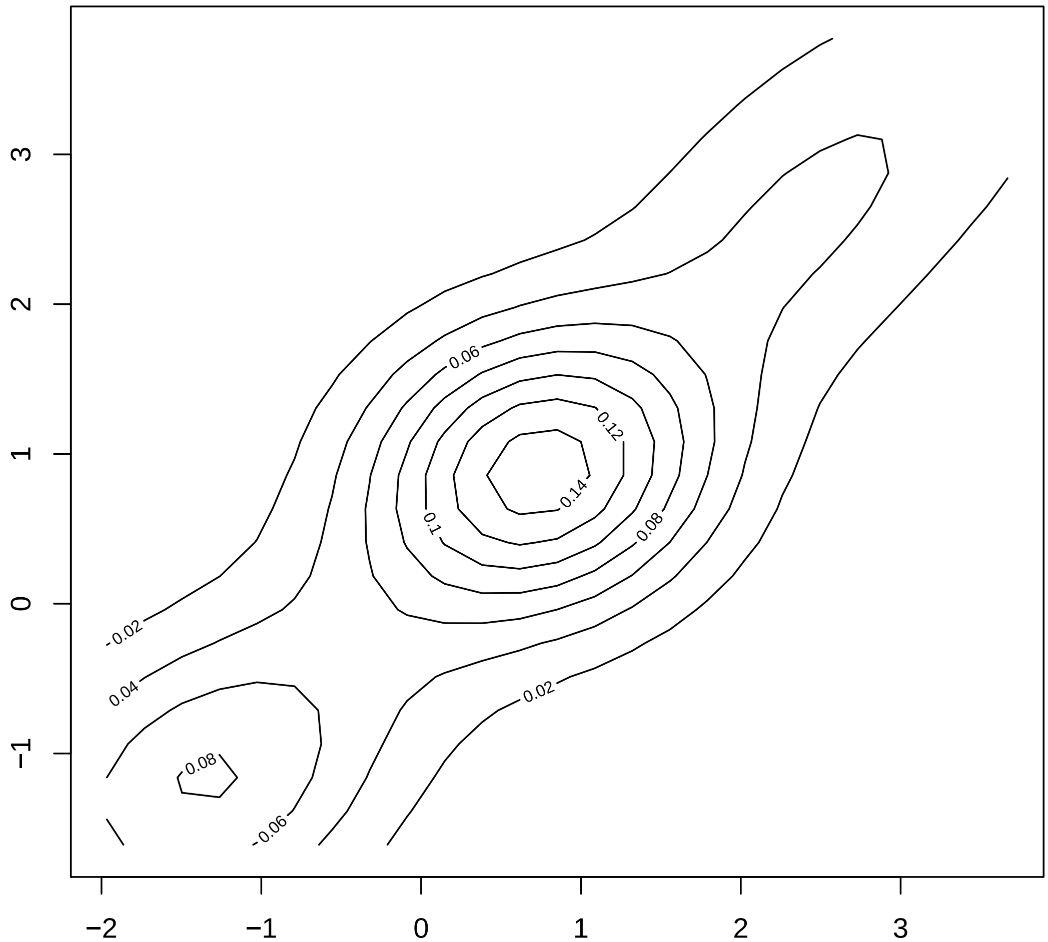



```
par(mfrow = c(1, 1))
```

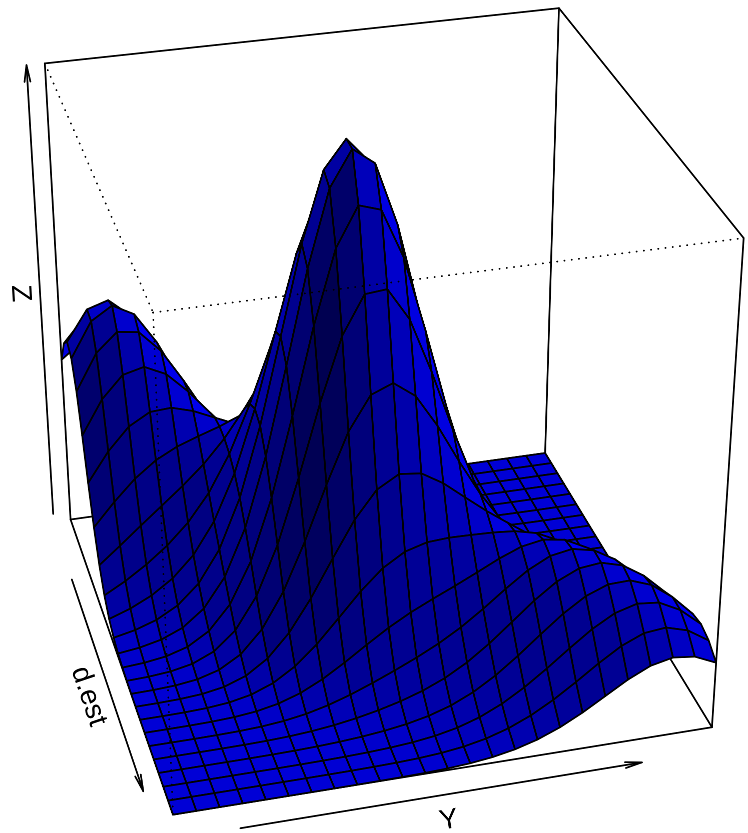
The "ucv" bandwidth appears to have picked up the two modes well but it also picked up some of the noise. The "nrd0" appears to have the best approximation of the density without the noise.

10.a Bivariate kernel density estimate.

```
dat <- read.csv("C:/Users/dominic/Documents/StatsGIDP/Statistics papers and courses/STAT675-Statist
d.est <- kde2d(dat$x, dat$y)
contour(d.est)
```




```
persp(d.est, phi = 30, theta = 75, d = 5, col = "blue", shade = 0.25)
```


Z
d.est
Y