HowTo Use the Bioconductor `edd` package

Vince Carey `stvjc@channing.harvard.edu`

February 22, 2005

Contents

1 Introduction

edd is a package that assists with one aspect of exploratory data analysis for microarrays. The basic question addressed in *edd* is the variety of shapes of gene-specific distributions of expression in collections of microarrays. Use of the package is most sensible when there are numerous arrays obtained under the same experimental condition or for a given clinical condition. The key idea is that marginal gene-specific distributions may have a relatively number of different qualitative shapes, some of which may be of considerable substantive interest (e.g., multimodal shapes), and some of which may be of methodologic importance (e.g., when one group of subjects has a skewed distribution for a gene, and another has a symmetric distribution for the same gene, use of a log transform is counterindicated).

In this brief HOWTO, we illustrate directly the use of the *edd* package. We will investigate the diversity of distributions in the two main groups of Golub's leukemia dataset.

2 Important caveat

The `edd` function will transform all gene-specific expression distributions to have common location and scale. This process can make noise have the appearance of signal. Before using edd, remove all genes that have small variability. See the next section for an example of this filtering process.

3 Distributional shapes in Golub's data

First we attach the necessary libraries and data frames. *edd* will require the *golubEsets* library.

```
> library(edd)

Loading required package: golubEsets
Loading required package: Biobase
Welcome to Bioconductor
         Vignettes contain introductory material.  To view,
         simply type: openVignette()
         For details on reading vignettes, see
         the openVignette help page.
Loading required package: Biobase
Loading required package: class
Loading required package: nnet
Loading required package: xtable

> library(golubEsets)
> library(xtable)
> data(golubMerge)
```

3.1 Filtering out genes with low variation

Next we filter the Golub data to require reasonable dispersion (confine attention to upper half sample defined by size of MAD) and reasonable expression (confine attention to genes with minimum expression level 300).

```
> madvec <- apply(exprs(golubMerge), 1, mad)
> minvec <- apply(exprs(golubMerge), 1, min)
> keep <- (madvec > median(madvec)) & (minvec > 300)
> gmfilt <- golubMerge[keep == TRUE, ]
```

3.2 Forming stratum-specific exprSets

Finally we split the dataset into the ALL and AML samples:

```
> ALL <- gmfilt$ALL.AML == "ALL"
> gall <- gmfilt[, ALL == TRUE]
> gaml <- gmfilt[, ALL == FALSE]
> show(gall)

Expression Set (exprSet) with
        540 genes
        47 samples
                 phenoData object with 11 variables and 47 cases
         varLabels
                Samples: Sample index
                ALL.AML: Factor, indicating ALL or AML
                BM.PB: Factor, sample from marrow or peripheral blood
                T.B.cell: Factor, T cell or B cell leuk.
                FAB: Factor, FAB classification
                Date: Date sample obtained
                Gender: Factor, gender of patient
                pctBlasts: pct of cells that are blasts
                Treatment: response to treatment
                PS: Prediction strength
                Source: Source of sample
```

3.3 Running edd

We will apply edd using an nnet classifier with the default reference catalog. See the edd-Details vignette for information about the reference catalog.

```
> set.seed(12345)
> alldists <- edd(gall, meth = "nnet", size = 10, decay = 0.2)

# weights:  579
initial  value 2005.605756
iter  10 value 1158.770201
iter  20 value 797.570948
iter  30 value 634.060511
iter  40 value 469.480315
iter  50 value 389.861685
iter  60 value 364.560886
iter  70 value 347.635653
iter  80 value 339.191762
```

```
iter  90 value 330.752808
iter 100 value 327.646425
final  value 327.646425
stopped after 100 iterations
```

```
> amldists <- edd(gaml, meth = "nnet", size = 10, decay = 0.2)
```

```
# weights:  359
initial  value 2294.213103
iter  10 value 1165.933356
iter  20 value 894.339722
iter  30 value 780.245051
iter  40 value 705.307718
iter  50 value 663.866225
iter  60 value 640.941637
iter  70 value 630.171545
iter  80 value 625.410486
iter  90 value 623.011611
iter 100 value 621.293653
final  value 621.293653
stopped after 100 iterations
```

An example of the results is given by the classification calls for the first 5 genes in the filtered exprSet:

```
              hum_alu_at AFFX-HUMGAPDH/M33197_3_at   AFFX-HSAC07/X00351_5_at
    ".75N(0,1)+.25N(4,1)"                    "t(3)"                    "t(3)"
  AFFX-HSAC07/X00351_3_at          AFFX-M27830_M_at
                   "t(3)"                  "X^2(1)"
```

We can use edd with other classification methods.

```
> alldistsKNN <- edd(gall, meth = "knn", k = 1, l = 0)
> alldistsTEST <- edd(gall, meth = "test", thresh = 0.3)
```

The agreement between nnet and knn procedures is not exact. See table 1. Choice between these methods and selection of tuning parameters is context-dependent.

```
> cap <- "Comparison of distribution shape classification by nnet (rows) and by knn (
> print(xtable(latEDtable(table(alldists, alldistsKNN), reorder = greo),
+     digits = rep(0, length(table(alldists)) + 1), caption = cap,
+     label = "conc1"))
```

The test procedure is the only one at present that allows an outcome of 'doubt'.

```
> print(table(alldistsTEST))
```

	Φ	t_3	$LN_{0,1}$	χ^2_1	$\beta_{8,2}$	$U_{0,1}$	$\beta_{2,8}$	$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$
Φ	55	5	0	0	5	0	4	2	1
t_3	19	67	5	1	0	0	45	17	0
$LN_{0,1}$	0	3	47	22	0	0	24	3	0
χ^2_1	0	0	1	2	0	0	0	0	0
$\beta_{8,2}$	0	1	0	0	7	0	0	0	0
$U_{0,1}$	1	0	0	0	1	3	0	0	0
$\beta_{2,8}$	10	1	0	0	0	3	119	15	0
$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	0	0	5	1	0	0	9	33	0
$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$	0	0	0	0	2	0	0	0	1

Table 1: Comparison of distribution shape classification by nnet (rows) and by knn (columns) methods in edd.

```
alldistsTEST
.25N(0,1)+.75N(4,1) .75N(0,1)+.25N(4,1)              B(2,8)              B(8,2)
                  9                  93                 169                  26
             N(0,1)              U(0,1)              X^2(1)           logN(0,1)
                 68                  26                   3                  40
            outlier                t(3)
                  2                 104
```

3.4 Assessing the results

We can assess the relative frequencies of the different shapes in the ALL samples with a table, see Table 2.

```
> cap <- "Frequencies of distributional shapes in filtered ALL data."
> print(xtable(latEDtable(table(alldists), reorder = greo), digits = rep(0,
+     length(table(alldists)) + 1), caption = cap, label = "marg1"))
```

Φ	t_3	$LN_{0,1}$	χ^2_1	$\beta_{8,2}$	$U_{0,1}$	$\beta_{2,8}$	$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$
72	154	99	3	8	5	148	48	3

Table 2: Frequencies of distributional shapes in filtered ALL data.

We can use barplots also; see Figure 1.

Discordance between distributional shapes in gene expression for the AML and ALL groups can be assessed using the cross-classification, see Table 3.

```
> cap <- "Rows are gene-specific distribution shapes for ALL, columns for AML, and ce
> print(xtable(latEDtable(table(alldists, amldists), reord = greo),
+     cap = cap, label = "disco1"))
```

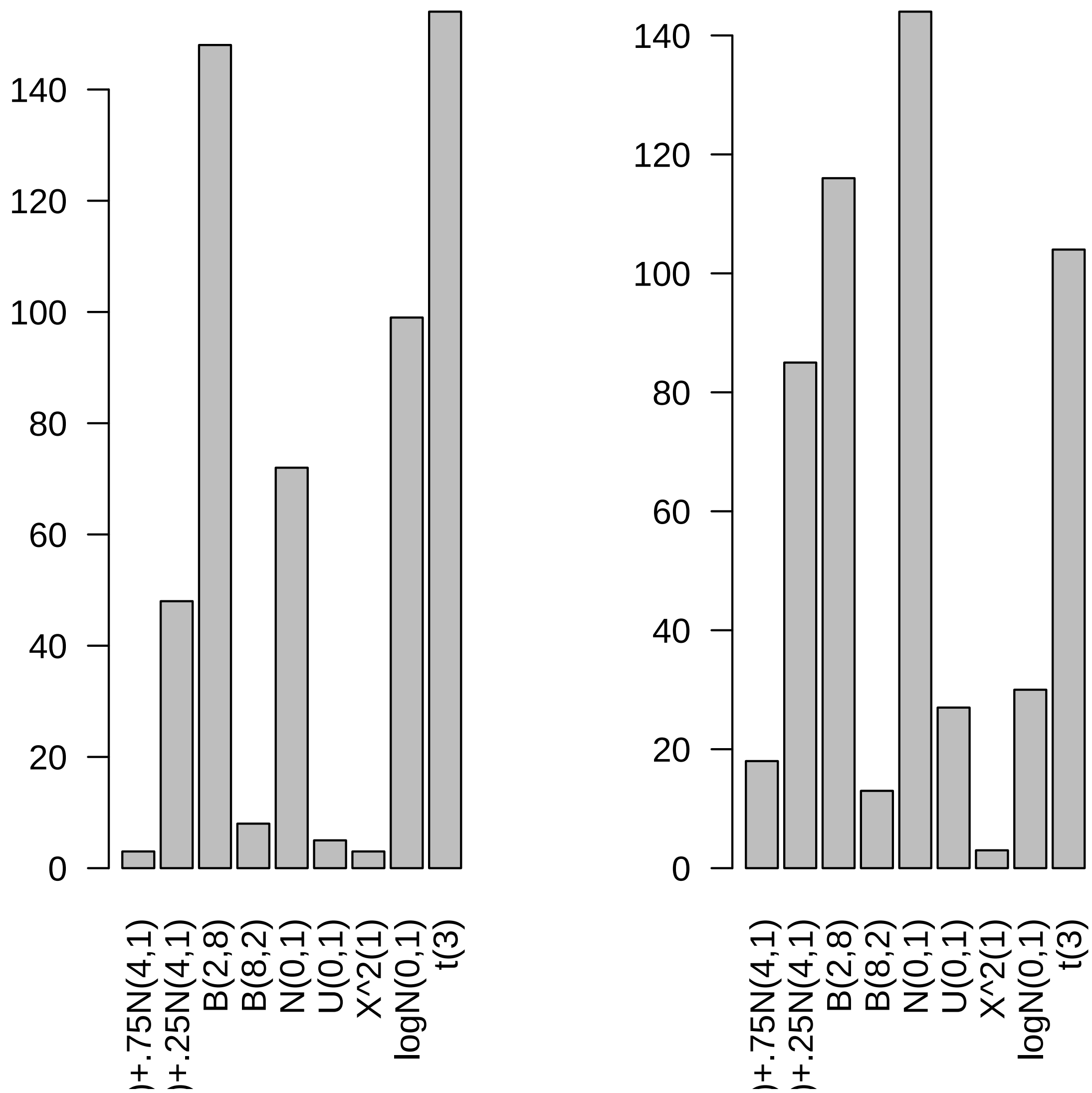

Figure 1: Compositions of distributional shapes within strata.

	Φ	t_3	$LN_{0,1}$	χ_1^2	$\beta_{8,2}$	$U_{0,1}$	$\beta_{2,8}$	$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$
Φ	29.00	12.00	1.00	0.00	6.00	2.00	10.00	8.00	4.00
t_3	40.00	41.00	6.00	0.00	2.00	7.00	21.00	30.00	7.00
$LN_{0,1}$	21.00	20.00	10.00	2.00	0.00	6.00	25.00	15.00	0.00
χ_1^2	0.00	1.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00
$\beta_{8,2}$	1.00	1.00	0.00	0.00	2.00	1.00	0.00	1.00	2.00
$U_{0,1}$	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
$\beta_{2,8}$	35.00	18.00	11.00	1.00	2.00	9.00	45.00	24.00	3.00
$\frac{3}{4}\Phi + \frac{1}{4}\Phi_{4,1}$	17.00	10.00	2.00	0.00	0.00	1.00	12.00	6.00	0.00
$\frac{1}{4}\Phi + \frac{3}{4}\Phi_{4,1}$	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 3: Rows are gene-specific distribution shapes for ALL, columns for AML, and cell entries are counts of genes.

Let's see what these discordances mean. To begin, let's get some indices for genes with bimodally shaped expression distribution for ALL, but approximately gaussian expression distribution for AML:

```
> print((1:540)[alldists == ".75N(0,1)+.25N(4,1)" & amldists ==
+     "N(0,1)"][1:5])
```

```
[1]   7  65  78 135 141
```

We consider the gene with probe D87953_at. The top left panel gives the model (solid density trace) and a kernel density estimate applied to the expression levels among ALL patients, and the top right is the corresponding histogram.

While the specific mixture model used as reference is not a perfect fit to the ALL data, the neural net classifier was sensitive to the bimodality. The Gaussian model does not seem particularly appropriate for the AML data, but was the closest match in the reference catalog.

4 Extending the reference catalog

The reference catalog supplied with edd has components

```
> names(eddDistList)
```

```
[1] "N01"   "T3"    "LN01"  "CS1"   "B82"   "U01"   "B28"   "MIXN1" "MIXN2"
```

There is nothing sacred about this set. Let's consider its scope (we'll look at 8 of nine reference distributions):

From the example above we see that it might be useful to have a mixture of Gaussians with modes separated by 6SD. To add such a model we construct an instance of the `eddDist` class:

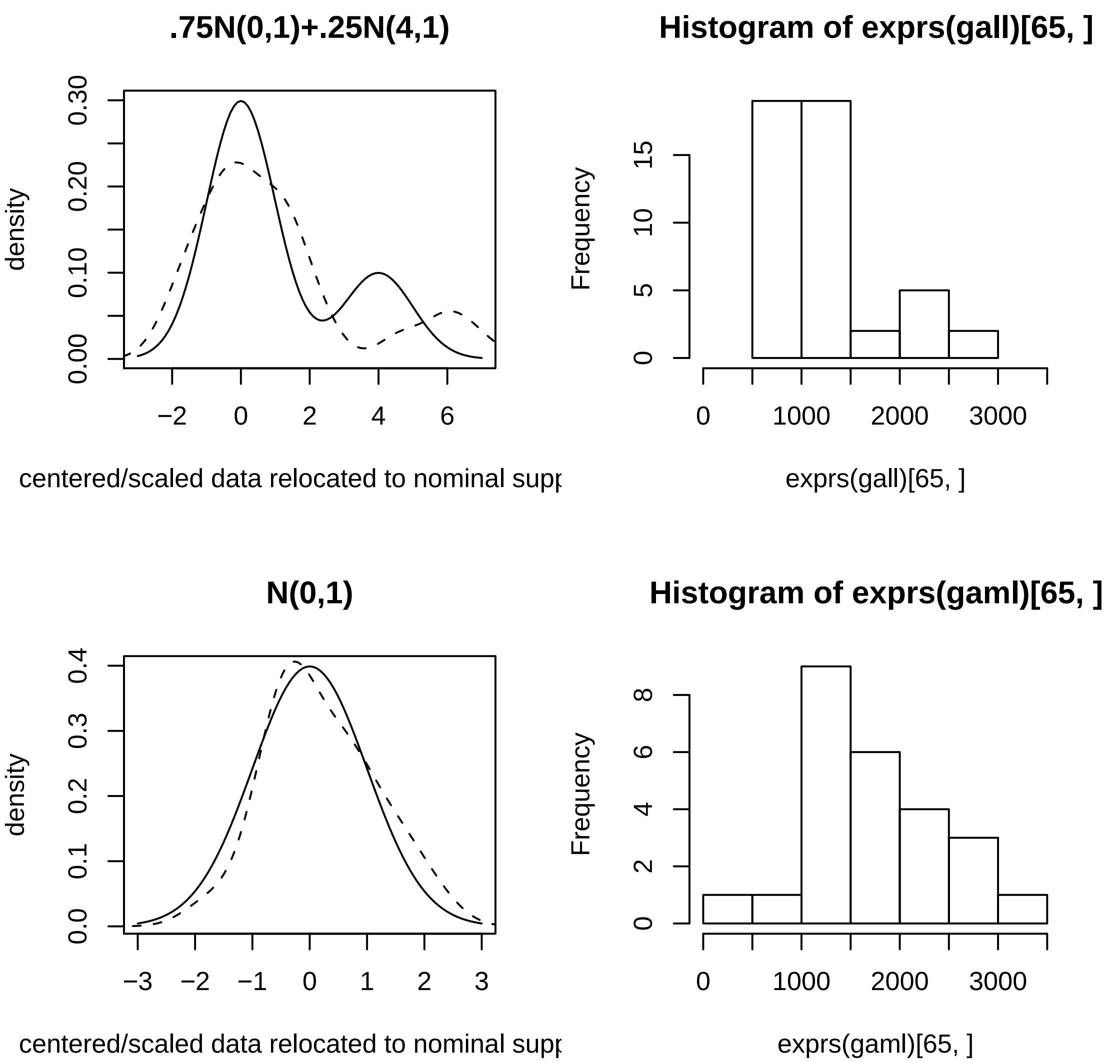

Figure 2: Two models for D87953_at in ALL and AML patients.

```
> par(mfrow = c(4, 2))
> for (i in 1:8) plotED(eddDistList[[i]])
```

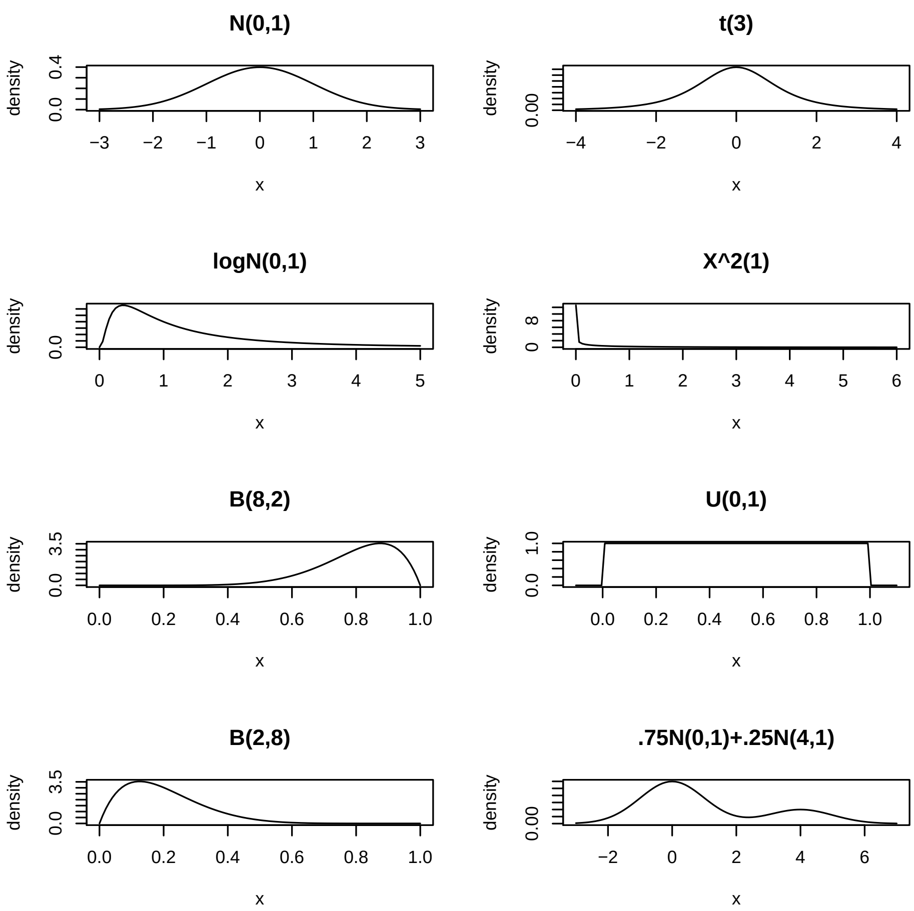

Figure 3: Eight of the reference distributions in the eddDistList supplied with *edd*.

```
> MIXN3 <- new("eddDist", stub = "mixnorm", parms = c(p1 = 0.75,
+     m1 = 0, s1 = 1, m2 = 6, s2 = 1), median = 0.43, mad = 1.55,
+     tag = ".75N(0,1)+.25N(6,1)", plotlim = c(-3, 11), latexTag = "$\\frac{3}{4}\\Ph
> eddDistList[["MIXN3"]] <- MIXN3
> set.seed(12345)
> alldists2 <- edd(gall, meth = "nnet", size = 10, decay = 0.2)
```

```
# weights:  590
initial  value 2774.830327
iter  10 value 1331.561921
iter  20 value 954.245585
iter  30 value 748.740961
iter  40 value 605.394017
iter  50 value 540.382629
iter  60 value 489.062645
iter  70 value 436.862377
iter  80 value 407.641990
iter  90 value 389.290441
iter 100 value 380.181412
final  value 380.181412
stopped after 100 iterations
```

```
> print(alldists2[65])
```

```
[1] ".75N(0,1)+.25N(6,1)"
```

The symbol MIXN3 used to name the list element is arbitrary, as are the values of the tag and latexTag slots. But the user should choose meaningful values for those items. The new reference distribution is used for classification of probe D87953_at. The two fits for the different mixtures are shown in Figures 4, 5.

```
> plotED(MIXN3, data = exprs(gall)[65, ])
```

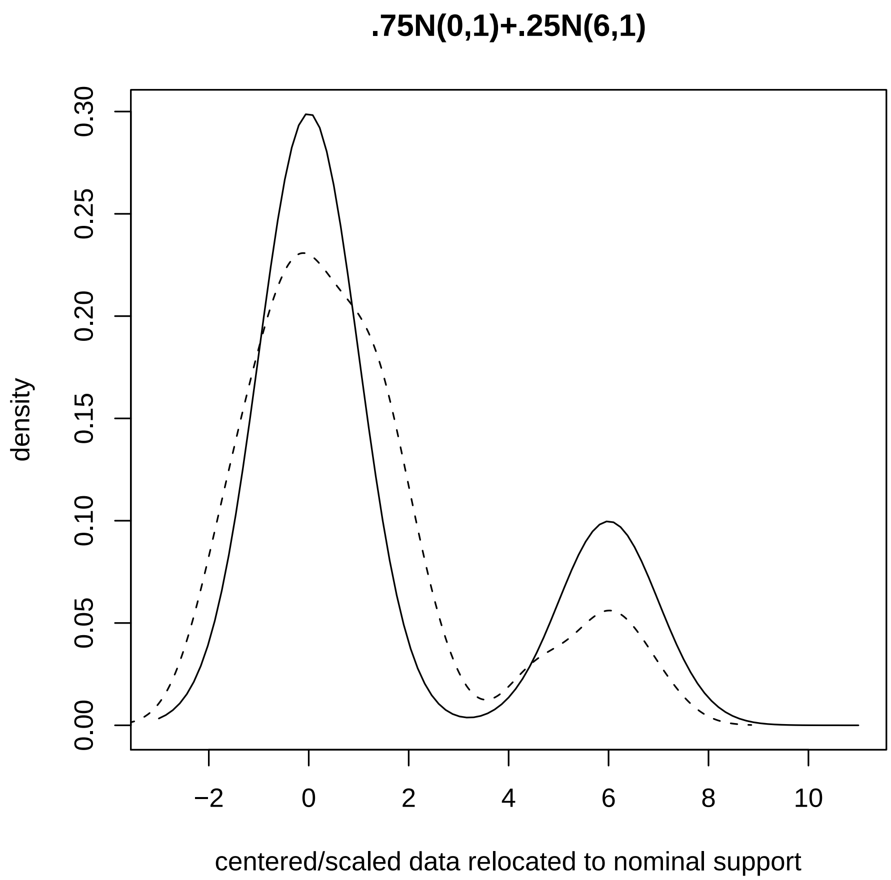

Figure 4: Reference catalog element: mixture with modes separated by 6SD. Superimposed is the kernel smooth of centered/scaled and then translated data for D87953_at.

```
> plotED(MIXN1, data = exprs(gall)[65, ])
```

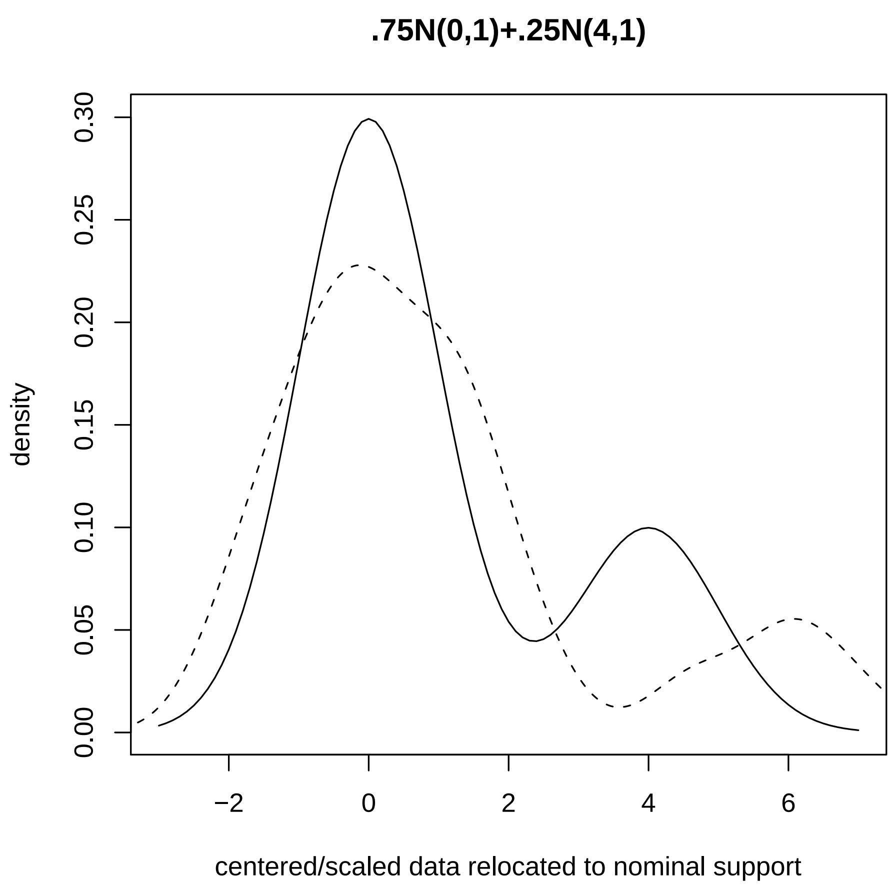

Figure 5: Reference catalog element: mixture with modes separated by 3SD. Superimposed is the kernel smooth of centered/scaled and then translated data for D87953_at.