prim: an R package for Patient Rule Induction Method (PRIM) estimation of highest density difference regions

Tarn Duong
Department of Statistics, University of New South Wales
Sydney Australia

29 June 2007

1 Introduction

The Patient Rule Induction Method (PRIM) was introduced by Friedman and Fisher (1999). It is a technique from data mining for finding 'interesting' regions in high-dimensional data. We start with regression-type data $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ where $\boldsymbol{X}_i$ is d-dimensional and Y_i is a scalar response variable. We are interested in the conditional expectation function

$$m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{x}).$$

In the case where we have a single sample then PRIM finds regions which correspond to the modal regions of $m(\boldsymbol{x})$. These regions are closely related to the highest density regions (HDR) of Hyndman (1996), defined in the form, for some threshold τ,

$$\{m(\boldsymbol{x}) \geq \tau\}.$$

In the case where we have 2 samples, we can label the response as

$$Y_i = \begin{cases} 1 & \text{if } \boldsymbol{X}_i \text{ is from sample 1} \\ -1 & \text{if } \boldsymbol{X}_i \text{ is from sample 2.} \end{cases}$$

Then PRIM finds the regions where the samples are most different. Here we have a positive HDR (where sample 1 points dominate) and a negative HDR (where sample 2 points dominate).

2 Bivariate 1-sample PRIM

We use a subset of the `Boston` data set in the `MASS` library. It contains housing data measurements for 506 towns in the Boston, USA area. For the explanatory variables, we take the nitrogen oxides concentration in parts per 10 million (`nox`) and the average number of room per dwelling (`rm`). The response is the per capita crime rate (`crim`). We are interested in characterising those areas with higher crime rates in order to provide better support infrastructure.

```
> library(prim)
> library(MASS)
> data(Boston)
> x <- Boston[, 5:6]
> y <- Boston[, 1]
> boston.prim <- prim.box(x = x, y = y, threshold.type = 1)
```

The default settings for `prim.box` are

- peeling quantile: `peel.alpha=0.05`
- pasting is carried out: `pasting=TRUE`
- pasting quantile: `paste.alpha=0.01`
- minimum box mass (proportion of points inside a box): `mass.min=0.05`
- `threshold` is the overall mean of the response variable `y`
- search for positive and negative HDR: `threshold.type=0`

We use the default settings except we wish to only find high crime areas $\{m(\boldsymbol{x}) \geq$ `threshold`$\}$ so we set `threshold.type=1`.

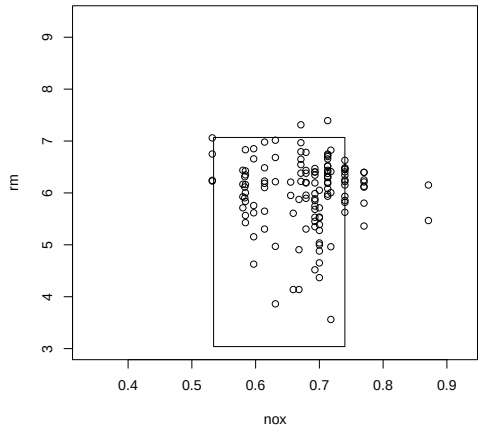
We view the output using a `summary` command. This displays three columns: the box mean, the box mass, and the HDR indicator (1 = positive HDR, –1 = negative HDR). Each line is a summary for each box, as well as an overall summary. Any box which is asterisked indicates that it does not form part of the HDR estimate. There is one box which contains 42.89% of the towns and where the average crime rate is 7.622. This is our HDR estimate. This regions comprises the bulk of the high crime areas, and is described in terms of nitrogen oxides levels in $[0.5341, 0.7400]$ and average number of rooms in $[3.0391, 7.0691]$. The other 57.11% of the towns have an average crime rate of 0.6036.

```
> summary(boston.prim)

         box-mean  box-mass box-ind
box1    7.6222290 0.4288538       1
box2*   0.6035267 0.5711462      NA
overall 3.6135236 1.0000000      NA

* - box not in highest density region at level = 3.613524

Box limits for box1
       nox     rm
min 0.5341 3.0391
max 0.7400 7.0691

Box limits for box2
        nox      rm
```

```
min 0.33640 3.03910
max 0.92446 9.35409
```

We plot the PRIM boxes, including all those towns whose crime rate exceeds 3.5. Thus verifying that the majority of high crime towns fall inside the HDR.

```
> plot(boston.prim, col = "transparent")
> points(x[y > 3.5, ])
```



There are many options for the graphical display. See the help guide for more details `?plot.prim`.

3 Quinti-variate highest density regions

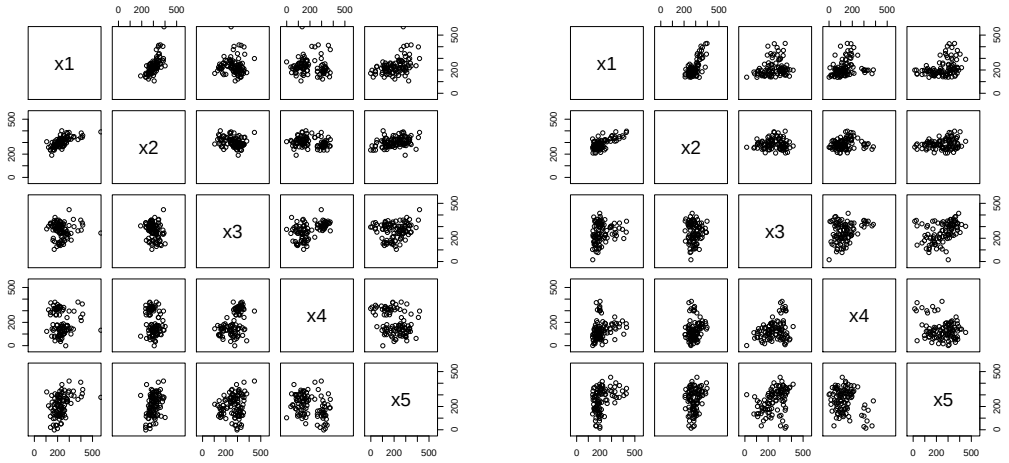
We started with a bivariate example since it is easy to visualise the results. However, PRIM was developed with much higher dimensional data in mind. So we look at a 5-dimensional data set (`quasiflow`) included in the `prim` library. It is a randomly generated data set from two normal mixture distributions whose structure mimics some light scattering data, taken from a machine known as a flow cytometer.

```
> library(prim)
> data(quasiflow)
> yflow <- quasiflow[, 6]
> xflowp <- quasiflow[yflow == 1, 1:5]
> xflowm <- quasiflow[yflow == -1, 1:5]
> xflow <- rbind(xflowp, xflowm)
```

We can think of `xflowp` as flow cytometric measurements from an HIV+ patient, and `xflowm` from an HIV– patient.

```
> xlim <- c(-30, 550)
> ylim <- c(-30, 550)
> pairs(xflowp[1:100, ], xlim = xlim, ylim = ylim)
> pairs(xflowm[1:100, ], xlim = xlim, ylim = ylim)
```

HIV+ HIV−



There are two ways of using `prim.box` to estimate where the two samples are most different (or equivalently to estimate the HDRs of the difference of the density functions). In the first way, we assume that we have suitable values for the thresholds. Then we can use

```
> qflow.thr <- c(0.85, -0.6)
> qflow.prim <- prim.box(x = xflow, y = yflow, threshold = qflow.thr,
+     threshold.type = 0)
```

An alternative is compute PRIM box sequences which cover the entire data range, and then use `prim.hdr` to experiment with different threshold values. This two-step process is more efficient and faster than calling `prim.box` for each different threshold. We're happy with the positive HDR threshold so we can compute the positive HDR directly:

```
> qflow.hdr.plus <- prim.box(x = xflow, y = yflow, threshold = 0.85,
+     threshold.type = 1)
```

On the other hand, we're not sure about the negative HDR thresholds.

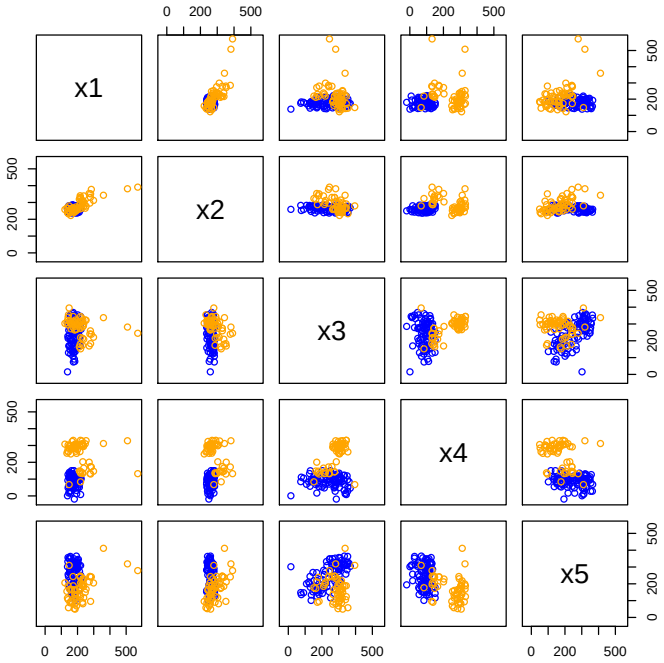
```
> qflow.minus <- prim.box(x = xflow, y = yflow, threshold.type = -1)
> qflow.hdr.minus1 <- prim.hdr(qflow.minus, threshold = -0.3, threshold.type = -1)
> qflow.hdr.minus2 <- prim.hdr(qflow.minus, threshold = -0.4, threshold.type = -1)
> qflow.hdr.minus3 <- prim.hdr(qflow.minus, threshold = -0.6, threshold.type = -1)
```

After examining the summaries and plots, we choose `qflow.hdr.minus3` to combine with `qflow.hdr.plus`.

```
> qflow.prim2 <- prim.combine(qflow.hdr.plus, qflow.hdr.minus3)
```

In the plot below, the positive HDR is coloured orange, and the negative HDR is coloured blue. These 5-dimensional HDRs can be more or less distinct some 2-dimensional projections e.g. (x_3, x_4) whereas in others they can overlap considerably e.g. (x_1, x_2). We conclude that there are more HIV+ patients within the orange regions and more HIV– patients within the blue regions. (The plot is not exactly what is produced by the `plot` command below – the number of data points has been thinned for purposes of clarity).

```
> xmin <- rep(-30, 5)
> xmax <- rep(550, 5)
> col <- qflow.prim2$ind
> col[col == 1] <- "orange"
> col[col == -1] <- "blue"
> plot(qflow.prim2, col = col, xmin = xmin, xmax = xmax)
```



The next step is to study the statistical properties of these HDR estimates. For some preliminary work, see Duong, Koch and Wand (2007).

References

Duong, T., Koch, I., and Wand, M. P. (2007). A generalised chi-squared test for comparing samples from data rich sources: an example from flow cytometry. In preparation.

Friedman, J. H. and Fisher, N. I. (1999). Bump-hunting for high dimensional data. *Statistics and Computing*, **9**, 123–143.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, **50**, 120–126.