Using prim for bump hunting

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

1 August 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 the bumps of $m(\boldsymbol{x})$.

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`

- `threshold.type=0`

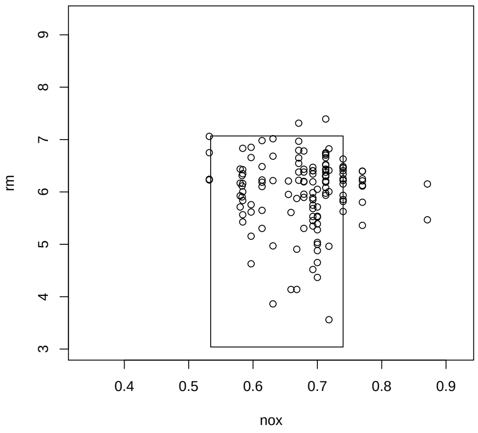
We use the default settings except we wish to only find high crime areas $\{m(\boldsymbol{x}) \geq \texttt{threshold}\}$ so we set `threshold.type=1`.

We view the output using a `summary` command. This displays three columns: the box mean, the box mass, and the threshold type. Each line is a summary for each box, as well as an overall summary. The box which is asterisked indicates that it is the box which contains the rest of the data not processed by PRIM. 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, print.box = TRUE)

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

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

Box limits for box2
       nox     rm
min 0.3364 3.0391
max 0.9196 9.3019
```

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 bump.

```
> 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`.

References

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