

# Classifying record pairs by means of Extreme Value Theory

Andreas Borg, Murat Sariyar

July 27, 2016

This document is a (practical) description of a procedure for Record Linkage by means of Extreme Value Theory (EVT). No labeled training data are needed, but user decisions are necessary for the selection of thresholds in a mean residual life plot (also known as mean excess plot).

In the following, the data set `RLdata500` will be used. As classification with EVT is weight-based, weights have to be calculated for the record pairs to classify. In this case an EM algorithm is applied.

```
> data(RLdata500)
> rpairs=compare.dedup(RLdata500,identity=identity.RLdata500,
+   blockfld=list(1,3,5,6,7),strcmp=1:4)
> rpairs=emWeights(rpairs)
```

```
>
```

Calling `getParetoThreshold` opens a mean residual life (MRL) plot for the computed weights, as shown in figure 1. From this graph, an interval has to be selected where the graph has a relatively long and approximately linear descent. Usually this can be found in the range between 0 and 20 for weights computed with `emWeights` or between 0.5 and 0.9 for weights computed with `epiWeights`. Figure 2 shows the same MRL plot with the appropriate segment marked.

The interval is selected by clicking on the endpoints of the desired segment of the graph. In some cases the right endpoint is identical to the edge of the graph, in this case only selection of the left endpoint is necessary. See the documentation of `identify` for more information on selecting points on a plot.

```
> ## Not run: getParetoThreshold(rpairs)
```

As an alternative to interactive selection, the interval can be given as argument to `getParetoThreshold`. The return value is in every case a threshold which can be used directly for classification.

```
> threshold=getParetoThreshold(rpairs,interval=c(1.2,12.8))
> result=emClassify(rpairs,threshold)
> summary(result)
```

Deduplication Data Set

500 records  
18643 record pairs

50 matches  
18593 non-matches  
0 pairs with unknown status

Weight distribution:

|           |           |           |           |          |
|-----------|-----------|-----------|-----------|----------|
| [-30,-25] | (-25,-20] | (-20,-15] | (-15,-10] | (-10,-5] |
| 13320     | 2505      | 1492      | 1079      | 175      |
| (-5,0]    | (0,5]     | (5,10]    | (10,15]   | (15,20]  |
| 22        | 8         | 18        | 21        | 0        |
| (20,25]   | (25,30]   |           |           |          |
| 0         | 3         |           |           |          |

42 links detected  
0 possible links detected  
18601 non-links detected

alpha error: 0.160000  
beta error: 0.000000  
accuracy: 0.999571

Classification table:

|             | classification |   |    |
|-------------|----------------|---|----|
| true status | N              | P | L  |
| FALSE       | 18593          | 0 | 0  |
| TRUE        | 8              | 0 | 42 |

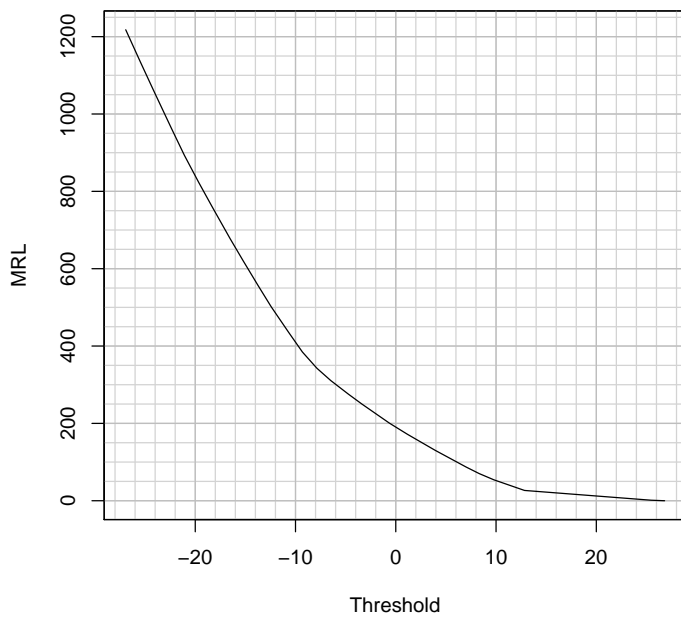


Figure 1: MRL plot

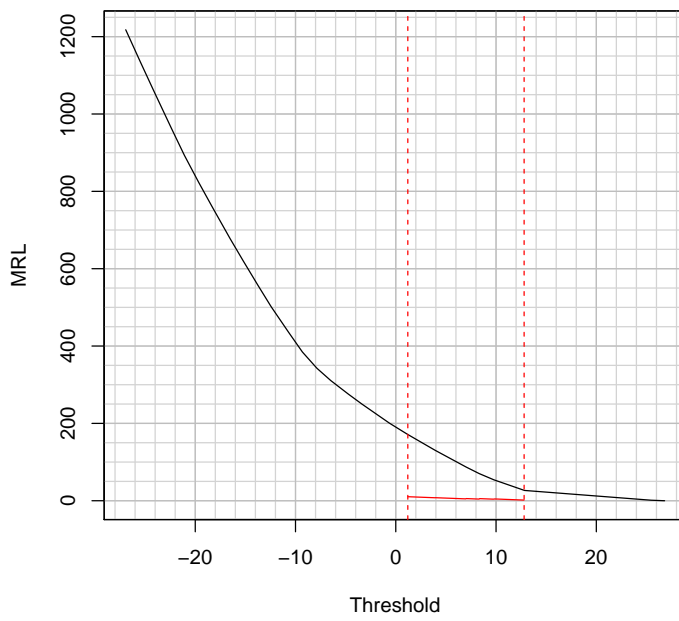


Figure 2: MRL plot with appropriate graph segment marked