Maximally Selected Rank Statistics in R

by Torsten Hothorn and Berthold Lausen

This document gives some examples on how to use the `maxstat` package and is basically an extention to Hothorn and Lausen (2002).

1 Introduction

The determination of two groups of observations with respect to a simple cutpoint of a predictor is a common problem in medical statistics. For example, the distinction of a low and high risk group of patients is of special interest. The selection of a cutpoint in the predictor leads to a multiple testing problem, cf. Figure 1. This has to be taken into account when the effect of the selected cutpoint is evaluated. Maximally selected rank statistics can be used for estimation as well as evaluation of a simple cutpoint model. We show how this problems can be treated with the `maxstat` package and illustrate the usage of the package by gene expression profiling data.

2 Maximally Selected Rank Statistics

The functional relationship between a quantitative or ordered predictor X and a quantitative, ordered or censored response Y is unknown. As a simple model one can assume that an unknown cutpoint μ in X determines two groups of observations regarding the response Y: the first group with X-values less or equal μ and the second group with X-values greater μ. A measure of the difference between two groups with respect to μ is the absolute value of an appropriate

standardized two-sample linear rank statistic of the responses. We give a short overview and follow the notation in Lausen and Schumacher (1992).

The hypothesis of independence of X and Y can be formulated as

$$H_0 : P(Y \leq y | X \leq \mu) = P(Y \leq y | X > \mu)$$

for all y and $\mu \in \mathbb{R}$. This hypothesis can be tested as follows. For every reasonable cutpoint μ in X (e.g. cutpoints that provide a reasonable sample size in both groups), the absolute value of the standardized two-sample linear rank statistic $|S_\mu|$ is computed. The maximum of the standardized statistics

$$M = \max_\mu |S_\mu|$$

of all possible cutpoints is used as a test statistic for the hypothesis of independence above. The cutpoint in X that provides the best separation of the responses into two groups, i.e. where the standardized statistics take their maximum, is used as an estimate of the unknown cutpoint.

Several approximations for the distribution of the maximum of the standardized statistics S_μ have been suggested. Lausen and Schumacher (1992) show that the limiting distribution is the distribution of the supremum of the absolute value of a standardized Brownian bridge and consequently the approximation of Miller and Siegmund (1982) can be used. An approximation based on an improved Bonferroni inequality is given by Lausen et al. (1994). For small sample sizes, Hothorn and Lausen (2003) derive an lower bound of the distribution function based on the exact distribution of simple linear rank statistics. The algorithm by Streitberg and Röhmel (1986) is used for the computations. The exact distribution of a maximally selected Gauß statistic can be computed using the algorithms by Genz (1992). Because simple linear rank statistics are asymptotically normal, the results can be applied to approximate the distribution of maximally selected rank statistics (see Hothorn and Lausen, 2003).

3 The maxstat Package

The package `maxstat` implements both cutpoint estimation and the test procedure above with several P-value approximations as well as plotting of the empirical process of the standardized statistics. It depends on the packages `exactRankTests` for the computation of the distribution of linear rank statistics (Hothorn, 2001) and `mvtnorm` for the computation of the multivariate normal distribution (Hothorn et al., 2001). The package is available at CRAN. The generic method `maxstat.test` provides a formula interface for the specification of predictor and response. An object of class `maxtest` is returned. The methods `print.maxtest` and `plot.maxtest` are available for inspectation of the results.

4 Gene Expression Profiling

The distinction of two types of diffuse large B-cell lymphoma by gene expression profiling is studied by Alizadeh et al. (2000). Hothorn and Lausen (2003) suggest the mean gene expression (MGE) as quantitative factor for the discrimination between two groups of patients with respect to overall survival time. The dataset `DLBCL` is included in the package. The maximally selected log-rank statistic for cutpoints between the 10% and 90% quantile of MGE using the upper bound of the P-value by Hothorn and Lausen (2003) can be computed by

```
>library("maxstat")
>library("survival")
>data("DLBCL", package="maxstat")
>mtHL <- maxstat.test(Surv(time, cens) ~ MGE,
+         data=DLBCL, smethod="LogRank", pmethod="HL")
>mtHL

        Maximally selected LogRank statistics using HL

data:  Surv(time, cens) by MGE
M = 3.171, p-value = 0.02218
sample estimates:
estimated cutpoint
         0.1860526
```

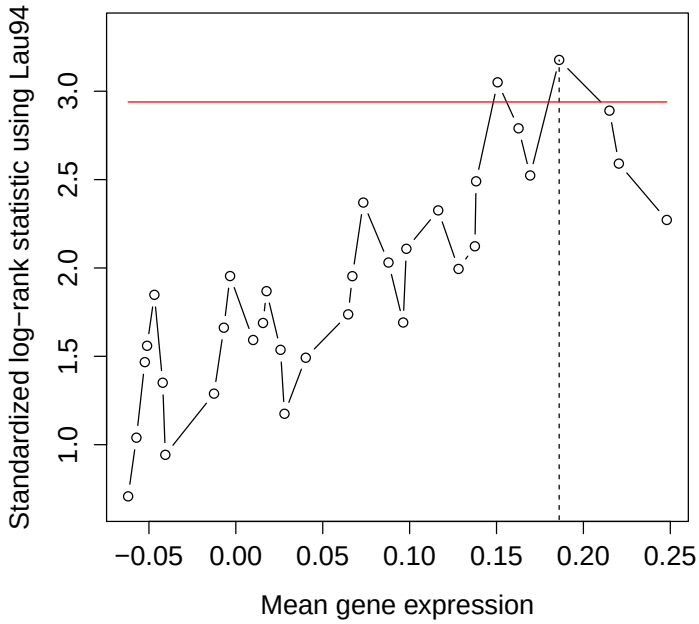

Figure 1: Absolute standardized log-rank statistics and significance bound based on the improved Bonferroni inequality.

For censored responses, the formula interface is similar to the one used in package `survival`: `time` specifies the time until an event and `cens` is the status indicator (`dead=1`). For quantitative responses y, the formula is of the form `y` $\sim$ `x`. Currently it is not possible to specify more than one predictor `x`. `smethod` allows the selection of the statistics to be used: `Gauss, Wilcoxon, Median, NormalQuantil` and `LogRank` are available. `pmethod` defines which kind of P-value approximation is computed: `Lau92` means the limiting distribution, `Lau94` the approximation based on the improved Bonferroni inequality, `exactGauss` the distribution of a maximally selected Gauß statistic and `HL` is the upper bound of the P-value by Hothorn and Lausen (2003). All implemented approximations are known to be conservative and therefore their minimum P-value is available

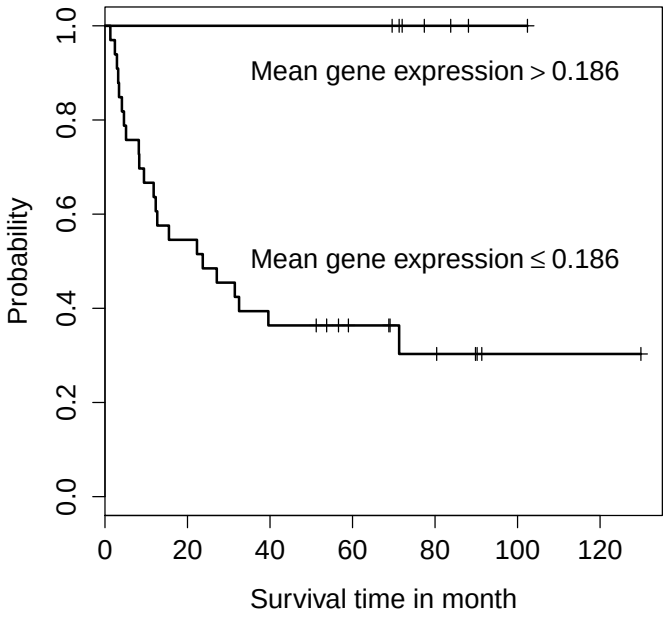

Figure 2: Kaplan-Meier curves of two groups of DLBCL patients separated by the cutpoint 0.186 mean gene expression.

by choosing `pmethod="min"`.

For the overall survival time, the estimated cutpoint is 0.186 mean gene expression, the maximum of the log-rank statistics is $M = 3.171$. The probability that, under the null hypothesis, the maximally selected log-rank statistic is greater $M = 3.171$ is less then than 0.022. The empirical process of the standardized statistics together with the α-quantile of the null distribution can be plotted using `plot.maxtest`. If the significance level `alpha` is specified, the corresponding quantile is computed and drawn as a horizonal red line. The estimated cutpoint is plotted as vertical dashed line, see Figure 1. The difference in overall survival time between the two groups determined by a cutpoint of 0.186 mean gene expression is plotted in Figure 2. No event was observed for patients

with mean gene expression greater 0.186.

The exact conditional p-value can be simulated via conditional Monte-Carlo. This may be time consuming but is an easy way to check the goodness of the p-value approximations. The argument `B` specifies the number of Monte-Carlo replications to be performed (and defaults to 10000).

```
>maxstat.test(Surv(time, cens) ~ MGE,
+               data=DLBCL, smethod="LogRank",
+               pmethod="condMC", B = 9999)

        Maximally selected LogRank statistics using condMC


data:  Surv(time, cens) by MGE
M = 3.1772, p-value = 0.0105
sample estimates:
estimated cutpoint
         0.1860526
```

5 More than One Predictor

If the cutpoints in more than one predictor are evalutated, the problem is to test the null hypothesis of independence of the response and any of the predictors under consideration. Furthermore, the "best" split in the "most significant" predictor needs to be selected. For example, we evaluate both mean gene expression and the International Prognostic Index (IPI) simultaneously:

```
>mmax <- maxstat.test(Surv(time, cens) ~ MGE + IPI,
+                data=DLBCL, smethod="LogRank",
+                pmethod="exactGauss", abseps=0.01)
>mmax

         Optimally Selected Prognostic Factors

Call: maxstat.test.data.frame(formula = Surv(time, cens) ~ MGE + IPI,
    data = DLBCL, smethod = "LogRank", pmethod = "exactGauss",
    abseps = 0.01)
```

```
 Selected:

        Maximally selected LogRank statistics using
        exactGauss

data:  Surv(time, cens) by IPI
M = 2.9603, p-value = 0.01092
sample estimates:
estimated cutpoint
                 1

Adjusted p.value:
0.03452566 , error:  0.001266494
```

The p-value of the global test for the null hypothesis "survival is independent from both IPI and MGE" is 0.035 and IPI provides a better distinction into two groups than MGE does. We can display the standardized statistics using `plot` (Figure 3). The methodology used here is described in Lausen et al. (2004).

6 Summary

The package `maxstat` provides a user-friendly interface and implements standard methods as well as recent suggestions for the approximation of the null distribution of maximally selected rank statistics.

References

Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types

of diffuse large B-cell lymphoma identified by expression profiling. *Nature*, 403:503–511, 2000.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149, 1992.

Torsten Hothorn. On exact rank tests in R. *R News*, 1(1):11–12, 2001.

Torsten Hothorn and Berthold Lausen. On maximally selected rank statistics. *R News*, 2(1):3–5, 2002.

Torsten Hothorn and Berthold Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2): 121–137, June 2003.

Torsten Hothorn, Frank Bretz, and Alan Genz. On multivariate t and Gauss probabilities in R. *R News*, 1(2):27–29, 2001.

Berthold Lausen and Martin Schumacher. Maximally selected rank statistics. *Biometrics*, 48:73–85, 1992.

Berthold Lausen, Wilhelm Sauerbrei, and Martin Schumacher. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In P. Dirschedl and R. Ostermann, editors, *Computational Statistics*, pages 483–496, Heidelberg, 1994. Physica-Verlag.

Berthold Lausen, Torsten Hothorn, Frank Bretz, and Martin Schumacher. Assessment of optimal selected prognostic factors. *Biometrical Journal*, 46(3): 364–374, 2004.

Rupert Miller and David Siegmund. Maximally selected chi square statistics. *Biometrics*, 38:1011–1016, 1982.

Bernd Streitberg and Joachim Röhmel. Exact distributions for permutations and rank tests: An introduction to some recently published algorithms. *Statistical Software Newsletter*, 12(1):10–17, 1986.

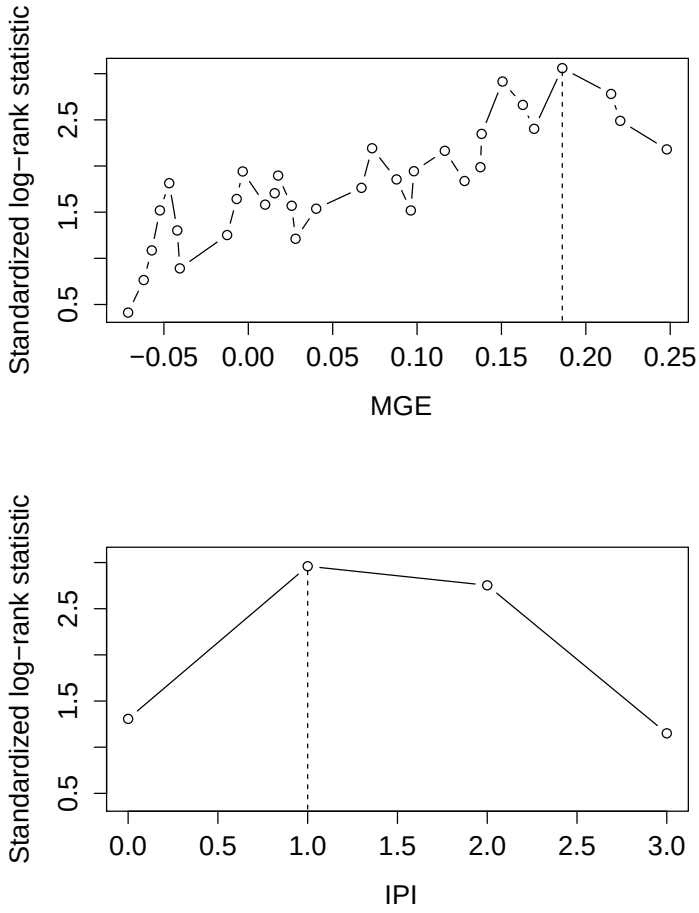

Figure 3: The standardized logrank statistics for the two predictors MGE and IPI.