Generalized Maximally Selected Statistics

Torsten Hothorn
Ludwig-Maximilians-Universität München

Achim Zeileis
Wirtschaftsuniversität Wien

Abstract

Maximally selected statistics for the estimation of simple cutpoint models are embedded into a generalized conceptual framework based on conditional inference procedures. This powerful framework contains most of the published procedures in this area as special cases, such as maximally selected χ^2 and rank statistics, but also allows for direct construction of new test procedures for less standard test problems. As an application, a novel maximally selected rank statistic is derived from this framework for a censored response partitioned with respect to two ordered categorical covariates and potential interactions. This new test is employed to search for a high-risk group of rectal cancer patients treated with a neo-adjuvant chemoradiotherapy. Moreover, a new efficient algorithm for the evaluation of the asymptotic distribution for a large class of maximally selected statistics is given enabling the fast evaluation of a large number of cutpoints.

Keywords: asymptotic distribution, changepoint, conditional inference, maximally selected statistics.

1. Introduction

Dichotomization of variables measured at higher scale levels prior to model building is bad practice (Royston *et al.* 2006, among many others). It will result in loss of power and sophisticated regression models that adapt themselves to the complexity of the regression problem at hand are widely available. However, simple regression models capturing step-shaped relationships between two variables (such as a single jump in the mean function) are valuable for the implementation of scientific results into the real world: a one-parameter 'good–poor' or 'high–low' decision rule is attractive to practitioners because of its simplicity.

Such rules of thumb are frequently used to investigate new predictor variables for patient survival in oncology. Galon *et al.* (2006) estimate cutpoints for various characteristics of immune cells within colorectal tumor samples, such as type, density or location, with respect to their ability to differentiate between patients with good and poor prognosis. Buccisano *et al.* (2006) obtain a threshold for residual leukemic cells in acute myeloid leukemia patients from maximally selected log-rank statistics. Beyond applications in oncology, the identification of ecological thresholds is of increasing interest (see Huggett 2005), e.g., the estimation of cutpoints for habitat factors discriminating between ecosystems with low and high abundance of certain indicator species (Müller and Hothorn 2004).

Two questions arise from a statistical point of view. In a first step, we have to make sure that there is some relevant association between response and covariate and in a second step we want to estimate the 'best' cutpoint in order to approximate this relationship by a simple model. It is convenient to deal with both problems separately. The first problem needs to be addressed by a formal hypothesis test for the null hypothesis of independence between covariate (to be dichotomized) and response variable. A test with power against shift alternatives, i.e., departures from the null hypothesis where the distribution of the response variable varies between two groups of observations, is of special interest. Once we are able to reject the null hypothesis, we are interested in the alternative which led to the rejection, i.e., want to estimate a cutpoint or partition.

This is a preprint of an article published in *Biometrics*, **64**, 1263–1269.

doi:10.1111/j.1541-0420.2008.00995.x

The first procedure of this kind, utilizing the maximum over multiple χ^2 statistics for 2×2 tables, was described by Miller and Siegmund (1982). Lausen and Schumacher (1992) derived an approximation for the asymptotical distribution of maximally selected rank statistics, extending the area of application to continuous and censored response variables. Betensky and Rabinowitz (1999) propose a maximally selected χ^2 test for nominal response variables measured at $k > 2$ levels and ordered categorical data.

Based on the ideas underlying these established techniques, we suggest a new generalized class of maximally selected statistics that contains the statistics sketched above as special cases but also allows for direct construction of new test procedures for less standard test problems. For evaluating the distribution of the test statistics, a conditional inference approach is adopted by embedding the tests into the theory of permutation tests of Strasser and Weber (1999). This permits efficient computation of the complete correlation structure of the statistics to be maximized. For statistics derived from cutpoints, the correlations have a special product form which we exploit for evaluation of the conditional asymptotic distribution: A linear-time algorithm is described which enables the fast assessment of a large number of cutpoints and improves upon approximations currently in use.

For illustrating the flexibility of the new framework for generalized maximally selected statistics we exemplify how the methodology can be extended to new areas of application by constructing a maximally selected log-rank statistic for a censored response partitioned with respect to two ordered categorical covariates and potential interactions. This new test is employed to search for a high-risk group determined by the T and N-category of rectal cancer patients. Further novel procedures include maximally selected statistics for multivariate responses or maximally selected permutation tests.

2. Binary partitions and two-sample statistics

We are provided with independent and identically distributed observations $(\mathbf{Y}_i, \mathbf{X}_i)$ for $i = 1, \ldots, n$ and are interested in testing the null hypothesis of independence of the response variable $\mathbf{Y} \in \mathcal{Y}$ and covariate(s) $\mathbf{X} \in \mathcal{X}$

$$H_0 : D(\mathbf{Y}|\mathbf{X}) = D(\mathbf{Y})$$

against shift alternatives. That is, departures from the null hypothesis where the distribution $D(\cdot)$ of the response variable varies between two groups of observations (with respect to $\mathbf{X}$) are of special interest.

Such binary partitions are defined in advance by p candidate sets $A_1, \ldots, A_p$. Each set A_j partitions the observations into two groups based on the covariate(s) only. For an ordered univariate covariate $\mathbf{X}$, these sets are typically constructed via cutpoints, i.e., $A_j = \{\mathbf{X}|\mathbf{X} \leq \xi_j\}$. When $\mathbf{X}$ is a factor at k levels, there are $p = 2^{k-1}$ possible partitions of the observations into two samples. For multivariate covariates, the A_j can code splits in interactions of the components of $\mathbf{X}$. A simple zero–one dummy coding for the jth partition is $g_j(\mathbf{X}) = I(\mathbf{X} \in A_j)$ where I denotes the indicator function. Only partitions satisfying a sample size constraint $\sum_i g_j(\mathbf{X}_i) \in (n\varepsilon, n - n\varepsilon)$ for some fixed $\varepsilon \in (0, 0.5)$ are taken into account (typically $\varepsilon = 0.1$).

The two-sample problem associated with the jth binary partition can be tested using a linear statistic

$$\mathbf{T}_j \quad = \quad \text{vec}\left\{\sum_{i=1}^{n} g_j(\mathbf{X}_i) h(\mathbf{Y}_i)^\top\right\} \in \mathbb{R}^{q \times 1}$$

where $h : \mathcal{Y} \to \mathbb{R}^{q \times 1}$ is an *influence function* applied to the responses. The function $h(\mathbf{Y}_i) = h\{\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)\}$ may depend on the full vector of responses $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, however only in a permutation symmetric way, i.e., the value of the function must not depend on the order in which $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ appear. For example, with h being a rank transformation for a continuous

response $\mathbf{Y}$, the linear statistic $\mathbf{T}_j$ is the sum of the ranks for observations from A_j, i.e., equals the Wilcoxon-Mann-Whitney statistic. When $\mathbf{Y}$ is a factor measured at k levels, $h \in \mathbb{R}^{(k-1)\times 1}$ is the corresponding dummy coding and $\mathbf{T}_j$ corresponds to the $2 \times k$ contingency table of the transformation $g_j(\mathbf{X})$ and response $\mathbf{Y}$.

A joint linear statistic for all binary partitions is

$$\mathbf{T} \quad = \quad (\mathbf{T}_1, \ldots, \mathbf{T}_p) = \text{vec}\left\{\sum_{i=1}^{n} g(\mathbf{X}_i) h(\mathbf{Y}_i)^\top\right\} \in \mathbb{R}^{pq \times 1}$$

including all p two-sample partitions, as defined by $g(\mathbf{X}) = \{g_1(\mathbf{X}), \ldots, g_p(\mathbf{X})\}$, simultaneously for testing H_0.

3. Standardization and estimation

To assess the partitions/cutpoints on a common scale, the corresponding statistics $\mathbf{T}_j$ are typically standardized using some location and scale measure. Consequently, inference can be based on the maximally selected absolute standardized statistics and the best separating partition is the one for which the maximum is attained.

For obtaining valid estimates of the mean and covariance of $\mathbf{T}$, either a parametric model needs to be specified or non-parametric techniques can be employed, such as permutation or re-sampling approaches. Here, we adopt the latter and utilize the permutation test framework established by Strasser and Weber (1999). Thus, $\mathbf{T}$ is standardized via its conditional expectation $\mu = \mathbb{E}(\mathbf{T}|S) \in \mathbb{R}^{pq\times 1}$ and covariance $\Sigma = \mathbb{V}(\mathbf{T}|S) \in \mathbb{R}^{pq\times pq}$, derived under H_0 by conditioning on all possible permutations S of the responses $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Closed-form expressions are as given by Strasser and Weber (1999):

$$\begin{aligned}
\mu = \mathbb{E}(\mathbf{T}|S) &= \text{vec}\left[\left\{\sum_{i=1}^{n} g(\mathbf{X}_i)\right\} \mathbb{E}(h|S)^\top\right] \\
\Sigma = \mathbb{V}(\mathbf{T}|S) &= \frac{n}{n-1}\mathbb{V}(h|S) \otimes \left\{\sum_i g(\mathbf{X}_i) \otimes g(\mathbf{X}_i)^\top\right\} \\
&- \frac{1}{n-1}\mathbb{V}(h|S) \otimes \left\{\sum_i g(\mathbf{X}_i)\right\} \otimes \left\{\sum_i g(\mathbf{X}_i)\right\}^\top
\end{aligned}$$

where $\otimes$ denotes the Kronecker product, and the conditional expectation of the influence function is $\mathbb{E}(h|S) = n^{-1}\sum_i h(\mathbf{Y}_i)$ with corresponding $q \times q$ covariance matrix $\mathbb{V}(h|S) = n^{-1}\sum_i \{h(\mathbf{Y}_i) - \mathbb{E}(h|S)\}\{h(\mathbf{Y}_i) - \mathbb{E}(h|S)\}^\top$.

When the observations are organized in independent blocks (such as centers in a multicenter trial), only permutations within blocks are admissible and thus expectations and covariance matrices have to be computed separately within each block. The expectation μ and covariance matrix Σ of $\mathbf{T}$ are then obtained as the sum over all expectations and covariance matrices. Therefore, it is easily possible to take a block randomization scheme in a randomized clinical trial or dependent sample designs into account.

The key step for constructing a maximally selected statistic is the standardization of $\mathbf{T}$ by its conditional expectation μ and covariance matrix Σ: the test statistic is the absolute maximum of the standardized linear statistic

$$T_{\max} \quad = \quad \max \frac{|\mathbf{T} - \mu|}{\sqrt{\text{diag}(\Sigma)}}.$$

When the test statistic is large enough to indicate a deviation from the null hypothesis we are interested in determining the partition with largest standardized statistic: the best separating

partition $A_{j^\star}$ is the one for which the maximum is attained, i.e., for which the absolute value of the standardized statistic $\mathbf{T}_{j^\star}$ equals $T_{\max}$.

4. Inference

For testing H_0, the conditional distribution of $T_{\max}$ given all permutations of the responses is used as reference distribution. Ideally, we want to compute the exact conditional distribution but this is only possible in special small sample situations (Boulesteix 2006a,b; Boulesteix and Strobl 2007). Conditional Monte-Carlo methods can be used to approximate the exact conditional distribution rather easily: evaluate the test statistic $T_{\max}$ for a large number of randomly shuffled responses $\mathbf{Y}$ and compute the p-value as proportion of permuted statistics that exceed the observed statistic.

Moreover, the exact conditional distribution can be approximated by its limiting distribution. For $n \to \infty$ the distribution of the multivariate linear statistic $\mathbf{T}$ tends to a multivariate normal distribution with mean μ and covariance matrix Σ (Strasser and Weber 1999, Theorem 3). Thus, in order to approximate $\mathsf{P}(T_{\max} > c)$ we have to evaluate the probability $\mathsf{P}\{\max(|Z_1|, ..., |Z_{pq}|) > c\}$ for standard normal random variables $Z_1, \ldots, Z_{pq}$ with correlation matrix $\mathbf{R}$ corresponding to the covariance matrix Σ and some $c > 0$. The computation of this probability is possible using Quasi-Monte-Carlo methods (Genz 1992) for moderate dimensions ($pq < 100$, say) but remains infeasible for higher dimensions. However, for the most important case of statistics maximally selected over cutpoints induced by an ordered covariate $\mathbf{X}$ and an ordered, censored or binary response $\mathbf{Y}$, the distribution can be evaluated numerically by an algorithm with computing time being linear in the number of cutpoints p as will be shown in the following.

5. A new and fast approximation

Let $A_j = (-\infty, \xi_j]$ with $\xi_j < \xi_k$ for $1 \leq j < k \leq p$ denote the partitioning sets and let $q = 1$ (i.e., ordered, censored or binary response variable). Then, the correlation between $\mathbf{T}_j$ and $\mathbf{T}_k$ is given by

$$\rho_{j,k} = \frac{\Sigma_{j,k}}{\sqrt{\Sigma_{j,j}\Sigma_{k,k}}} = \sqrt{\frac{\{n - \sum_i g_k(\mathbf{X}_i)\}\sum_i g_j(\mathbf{X}_i)}{\{n - \sum_i g_j(\mathbf{X}_i)\}\sum_i g_k(\mathbf{X}_i)}}.$$

It follows that the correlation matrix $\mathbf{R} = (\rho_{j,k})_{j,k=1,\ldots,p}$ is completely determined by the subdiagonal elements $\rho_{j,j-1}, j = 2, \ldots, p$ and it holds that

$$\rho_{1,k} = \prod_{j=2}^{k} \rho_{j,j-1}.$$

With $\mathbf{v} = (\rho_{1,1}, \ldots, \rho_{1,p})$ the lower triangular part of $\mathbf{R}$ can be written as $\mathbf{v}(1/\mathbf{v})^\top$ and it follows from Meurant (1992, Section 2.1) that the inverse $\mathbf{R}^{-1}$ of the correlation matrix is a tridiagonal symmetric band matrix:

$$\mathbf{R}^{-1} = \begin{pmatrix} r_{1,1} & r_{1,2} & 0 & 0 & \ldots & 0 \\ r_{1,2} & r_{2,2} & r_{2,3} & 0 & \ldots & 0 \\ 0 & r_{2,3} & r_{3,3} & r_{3,4} & \ldots & 0 \\ 0 & 0 & r_{3,4} & r_{4,4} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & r_{p-1,p-1} & r_{p-1,p} \\ 0 & 0 & 0 & 0 & r_{p-1,p} & r_{p,p} \end{pmatrix}.$$

The probability that any of $|Z_1|, \ldots, |Z_p|$ exceeds $c > 0$ is

$$\mathsf{P}(T_{\max} > c) \quad = \quad 1 - \frac{1}{\sqrt{|\mathbf{R}|(2\pi)^p}} \int_{-c}^{c} \exp\left(-\frac{1}{2}\mathbf{z}^\top \mathbf{R}^{-1}\mathbf{z}\right) d\mathbf{z}.$$

Due to the band structure of $\mathbf{R}^{-1}$ the quadratic form $\mathbf{z}^\top \mathbf{R}^{-1}\mathbf{z}$ simplifies to

$$\mathbf{z}^\top \mathbf{R}^{-1}\mathbf{z} = r_{1,1}z_1^2 + 2r_{2,1}z_1z_2 + r_{2,2}z_2^2 + \cdots + 2r_{p,p-1}z_pz_{p-1} + r_{p,p}z_p^2$$

which is employed for evaluating the multivariate normal distribution numerically (Genz and Kahaner 1986). With $\phi(z) = \exp(-z/2)$ we have

$$\int_{-c}^{c} \phi\left(\mathbf{z}^\top \mathbf{R}^{-1}\mathbf{z}\right) d\mathbf{z} =$$
$$\int_{-c}^{c} \phi(r_{1,1}z_1^2) \int_{-c}^{c} \phi(2r_{2,1}z_1z_2 + r_{2,2}z_2^2) \int_{-c}^{c} \ldots \int_{-c}^{c} \phi(2r_{p,p-1}z_pz_{p-1} + r_{p,p}z_p^2) d\mathbf{z}$$

and with recursively defined functions f_j $(j = 2, \ldots, p+1)$

$$f_j(z) = \int_{-c}^{c} \phi\left(2r_{j,j-1}z\tilde{z} + r_{j,j}\tilde{z}^2\right) f_{j+1}(\tilde{z}) d\tilde{z} \quad \forall j = 2, \ldots, p; \quad f_{p+1}(z) \equiv 1$$

the above integral can be re-formulated recursively:

$$\begin{aligned} \mathsf{P}(T_{\max} > c) \quad &= \quad 1 - \frac{1}{\sqrt{|\mathbf{R}|(2\pi)^p}} \int_{-c}^{c} \phi\left(\mathbf{z}^\top \mathbf{R}^{-1}\mathbf{z}\right) d\mathbf{z} \\ &= \quad 1 - \frac{1}{\sqrt{|\mathbf{R}|(2\pi)^p}} \int_{-c}^{c} \phi(r_{1,1}z^2) f_2(z) dz. \end{aligned}$$

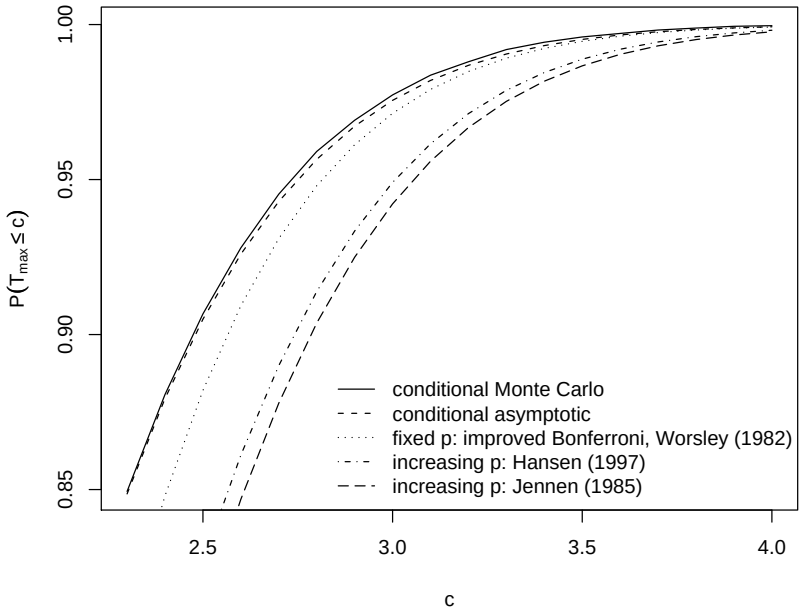
This integral can be evaluated numerically in $O(p)$ starting with f_p utilizing the techniques described by Miwa *et al.* (2000): For a two-dimensional grid of $z \in [-c, c]$ and $\tilde{z} \in [-c, c]$ values, the function f_j is evaluated and aggregated over $\tilde{z}$ only, yielding values of $f_j(z)$ for a grid of z values. These values are then re-used when computing f_{j-1}.

Comparing this new approximation of the asymptotic distribution with previously suggested approximations (see Figure 1), it should be pointed out that these approximations differ with respect to the asymptotics for p, the number of cutpoints. In a conditional framework, it is most natural to treat p as fixed (given the observed data). Taking an unconditional view, it depends on the partition-generating mechanism whether p is fixed as $n \to \infty$ or increases. The former holds for splits at sample quantiles for numeric covariates or for splits in categorical variables where $p = k-1$ splits are possible for ordinal factors or $p = 2^{k-1}$ for unordered factors. However, if all possible splits in a continuous covariate $\mathbf{X}$ are considered, then $p \to \infty$ as $n \to \infty$ and the sequence of test statistics $Z_1, \ldots, Z_p$ is known to converge to a stochastic Gaussian process with continuous paths:

$$Z^0(t) \quad = \quad \frac{B^0(t)}{\sqrt{t(1-t)}}, \quad t \in [0, 1]$$

where $B^0(t)$ is a Brownian bridge that is scaled to zero mean and unit variance (Miller and Siegmund 1982). The correlation of $Z^0(s)$ and $Z^0(t)$ for $s \leq t$ is $\sqrt{s(1-t)}/\sqrt{t(1-s)}$ and is exactly the same as above. More formally, with $t_j = \lim_{n\to\infty} n^{-1} \sum_i g_j(\mathbf{X}_i)$, Z_j and $Z^0(t_j)$ are

Figure 1: Approximations of the distribution of a maximally selected Wilcoxon-statistic for $n = 100$ observations and $p = 20$ cutpoints. The exact distribution as approximated by $30,000$ random permutations of the data is shown as a solid line, most closely approximated by the conditional asymptotic distribution suggested here.



asymptotically identical in distribution. Therefore, the difference between the two approaches is that for increasing p the asymptotical distribution is given by $\sup_{t\in[\varepsilon,1-\varepsilon]} Z^0(t)$ whereas for fixed p it is $\max_{t\in\{t_1,\ldots,t_p\}} Z^0(t)$. Thus, the supremum over the full interval will always be larger than the maximum over a subset of times/partitions because of the additional variation in the intervals (t_j, t_{j+1}). The difference between the two approaches decreases with $t_{j+1} - t_j$. Figure 1 shows that the exact conditional distribution (approximated by $30,000$ permutations) is most closely captured by the conditional asymptotic distribution suggested above. Less accurate is the improved Bonferroni correction (Worsley 1982) which also uses a fixed p. The two approximations for increasing p (Jennen 1985; Hansen 1997) are (not surprisingly) clearly below.

These considerations about the asymptotic behavior of p also raise the question about the quality of the asymptotic approximation for finite samples when p is large compared to n. The joint approximation is appropriate if the normal approximation for each two-sample statistic is. Consider a split in a categorical variable with large number of categories, e.g., 10 categories with 10 observations each leads to $n = 100$ but $p = 2^{10-1} = 512$. But since each two-sample statistic is based on at least 10 and 90 observations, respectively, the normal approximation should work well enough.

6. Applications and special cases

Maximally selected statistics as described in Section 2 can be applied to covariates $\mathbf{X}$ and responses $\mathbf{Y}$ measured at arbitrary scales; appropriate influence functions h for nominal, ordered, numeric, censored and multivariate response variables are given in the sequel (see Hothorn *et al.* 2006, for further details), followed by a description of how to partition the covariate space for nominal,

ordered and multivariate covariates and the derivation of a novel maximally selected statistic.

For categorical responses, h is typically a simple dummy coding for nominal $\mathbf{Y}$ and a vector of numeric scores (corresponding to the k levels) for ordinal $\mathbf{Y}$. Many possible influence functions are available for discrete or continuous covariates, e.g., identity, square root, log, or rank transformations; and for censored responses log-rank or Savage scores can be applied.

The most important situation of a univariate and (at least) ordinally measured covariate $\mathbf{X}$ leads to partitions, and thus functions g_j, induced by cutpoints defined by the realizations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. More specifically, $A_j = (-\infty, \xi_j]$, where ξ_j is the jth element of the increasingly sorted unique realizations of $\mathbf{X}$. Thus, having identified the best separating partition $A_{j^\star}$, the estimated cutpoint is $\xi_{j^\star}$. For nominal covariates, all 2^{k-1} binary partitions of the k levels are considered. For multiple covariates, we simply look at all binary partitions induced by interactions of all covariates simultaneously.

This flexible framework can now be utilized to implement a wide variety of already published as well as novel maximally selected statistics. One should bear in mind that we always utilize the conditional null distribution which might differ from the unconditional distribution as pointed out above.

Maximally selected χ^2 statistics. The response variable is a factor at two levels a and b, say, and $h(\mathbf{Y}_i) = I(\mathbf{Y}_i = a)$ is a dummy coding. The ordered univariate covariate $\mathbf{X}$ offers $p \leq n - 1$ cutpoints $\xi_1, \ldots, \xi_p$ leading to $g_j(\mathbf{X}_i) = I(\mathbf{X}_i \leq \xi_j)$ for $j = 1, \ldots, p$. Thus,

$$\mathbf{T}_j = \sum_{i=1}^{n} g_j(\mathbf{X}_i) h(\mathbf{Y}_i)^\top = \sum_{i=1}^{n} I(\mathbf{X}_i \leq \xi_j) I(\mathbf{Y}_i = a) \in \mathbb{R}$$

is the number of observations i with $\mathbf{X}_i \leq \xi_j$ and $\mathbf{Y}_i = a$. $\mathbf{T}_j$ is one element of the 2×2 contingency table for $I(\mathbf{X} \leq \xi_j)$ and $I(\mathbf{Y} = a)$ and determines the complete table because the margins are fixed. The statistic $(\mathbf{T} - \mu)^2 / \mathrm{diag}(\Sigma) \in \mathbb{R}^{p \times 1}$ is equivalent to the p-vector of χ^2 statistics for all p tables and our maximally selected statistic $T_{\max}$ is a monotone transformation of the maximally selected χ^2 statistic proposed by Miller and Siegmund (1982). For nominal responses $\mathbf{Y}$ with $k > 2$ levels, the statistic $\mathbf{T}_j$ corresponds to the first $k - 1$ columns of the first row of the $2 \times k$ contingency table of $\mathbf{X}_i \leq \xi_j$ and $\mathbf{Y}$. The $T_{\max}$ statistic is the maximum over the maximum of p standardized contingency tables, an alternative to maximally selected χ^2 statistics for larger tables (Betensky and Rabinowitz 1999).

Maximally selected Cochran-Armitage statistics. The ordered response $\mathbf{Y}$ is measured at k ordered levels. The influence function h assigns a score γ_j to each level $j = 1, \ldots, k$. For the special case $\gamma_j = j$ this corresponds to the Cochran-Armitage test and consequently the statistic $T_{\max}$ is equivalent to a maximally selected Cochran-Armitage statistic (Betensky and Rabinowitz 1999). For arbitrary scores, the resulting test is a maximally selected test based on linear-by-linear association statistics.

Maximally selected rank statistics. Let h denote the rank transformation of a univariate response $\mathbf{Y}$. Then, the statistic

$$\mathbf{T}_j = \sum_{i=1}^{n} I(\mathbf{X}_i \leq \xi_j) h(\mathbf{Y}_i)^\top = \sum_{i: \mathbf{X}_i \leq \xi_j} \mathrm{rank}(\mathbf{Y}_i)$$

is the sum of the ranks for all observations with $\mathbf{X}_i \leq \xi_j$, i.e., the Wilcoxon rank sum statistic. More generally, h can be any rank transformation (normal scores, median scores, log-rank scores etc.) and the linear statistic $\mathbf{T}_j$ is equivalent to a linear rank statistic (Hájek *et al.* 1999). Consequently, $T_{\max}$ is equivalent to a maximally selected rank statistic in the sense of Lausen and Schumacher (1992, 1996) and Hothorn and Lausen (2003).

Maximally selected statistics for multiple covariates. When multiple covariates are under test simultaneously, we consider all unique partitions induced by all possible cutpoints in each covariate. For an ordered response, this special case of maximally selected rank statistics for multiple covariates has first been studied by Lausen *et al.* (2004).

Three novel maximally selected statistics. Due to the flexibility of the generalized framework we can easily construct tests adapted to specific problems by choosing a suitable set of candidate partitions g and transformation of the response h. Here, we exemplify three applications: maximally selected permutation tests, maximally selected statistics for multivariate responses and maximally selected statistics for interactions. Instead of using a rank-based transformation, it is often more natural to use the original observations, i.e., employ the identity transformation $h(\mathbf{Y}) = \mathbf{Y}$. Thus, each linear statistic $\mathbf{T}_j$ corresponds to a two-sample permutation test for location alternatives. For a multivariate response, such as abundances of multiple species under investigation (De'ath 2002), the influence function h is a combination of influence functions appropriate for any of the univariate response variables as suggested above. Finally, for multiple covariates, we can not only combine the partitions g for each individual covariate—corresponding to splits in one covariate at a time—but also employ splits in interactions of the covariates. Contrary to previously suggested maximally selected procedures (Lausen *et al.* 2004) or recursive splitting algorithms such as CART (Breiman *et al.* 1984), we can simultaneously search for splits in more than one variable and thus capture interactions like the well-known XOR problem. Below, such a strategy is employed for splitting in two ordinal covariates (an approach to splitting in SNP-SNP interactions is given in Boulesteix *et al.* 2007). To reflect the ordering, it is natural to include only those interactions that correspond to a single cutpoint in each covariate given the level of the other (and vice versa). Thus, interactions that would correspond to multiple cutpoints in one covariate given the level of the other are excluded from the set of all potential interactions.

7. Illustration

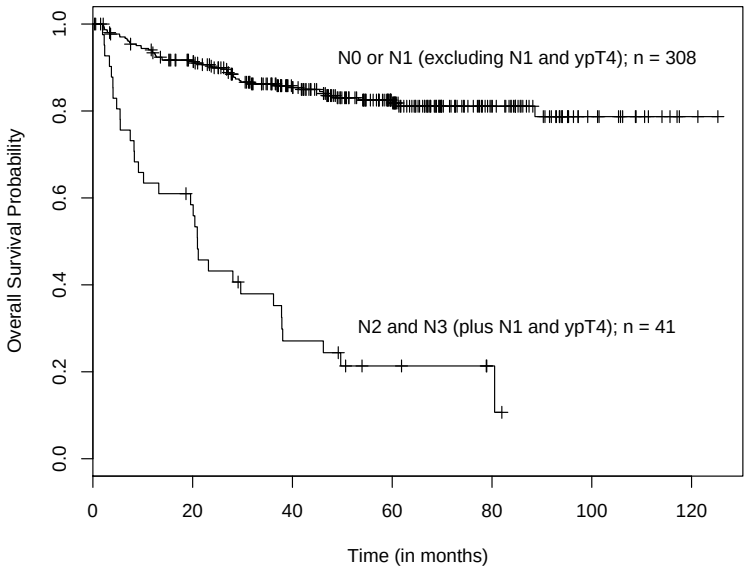
We attempt to identify high- and low-risk groups of rectal cancer patients by differentiation with respect to pathological T and N category (ordinal assessments of tumours and lymph nodes, respectively). The data are taken from the preoperative arm of the CAO/ARO/AIO-94 trial (Sauer *et al.* 2004) and comprise survival times for $n = 349$ patients treated with a neo-adjuvant chemoradiotherapy regime starting before surgery. All patients belong to M category (assessment of metastates) M0, 48 patients from category M1 were excluded from the analysis.

For this situation, we propose a new maximally selected statistic for a censored response and two ordered covariates with potential interactions. Log-rank scores are used as influence function h for the censored response and the potential partitions g are constructed from all combinations of the five T and three N categories. As both categories are ordered, only those partitions are used which are ordered in T given N and vice versa yielding 194 candidate partitions, 187 of which

Table 1: Pathological T and N category of 349 rectal cancer patients treated with a preoperative chemoradiotherapy.

	N category			
T category	N0	N1	N2 + N3	Total
ypT0	36	4	0	40
ypT1	16	5	0	21
ypT2	90	14	7	111
ypT3	107	33	29	169
ypT4	3	1	4	8
Total	252	57	40	349

Figure 2: Survival times of rectal cancer patients in two risk groups identified by a novel maximally selected log-rank statistic based on interactions of T and N category.



meet the sample size constraints. The maximum of the absolute values of the corresponding 187 standardized statistics is 8.69 with a p-value smaller than 0.0001. The partition chosen by the algorithm identifies all patients from category N2 or N3 as being under high risk and almost all patients from N0 and N1 as being under low risk. As an exception, a single patient with ypT4 and N1 is assigned to the high risk group as well—whether or not this decision is sensible or results from random variation cannot be judged based on one observation alone. Figure 2 depicts Kaplan-Meier estimates of the survival times in the two risk groups.

To relate these results to current practice, we employ the TNM system (Sobin and Wittekind 2002) for cancer classification. It defines three stages by fixed cutpoints in the interaction of T and N category: stage I vs. II is discriminated by the T category, stage II vs. III by the N category (an additional stage IV is based on the M category). Thus, TNM also uses the N category to distinguish more severe forms of cancer; however, it uses the split N $\leq$ N0 (associated with a much smaller standardized statistic of 7.31) while our procedure selects N $\leq$ N1 by maximizing over all partially ordered interactions (including all fixed interactions from the TNM stages). Placing category N1 in the low-risk group might be associated with application of chemoradiotherapy before rather than after surgical resection (for which the TNM staging is applied). This and other prognostic factors for preoperative chemoradiotherapy are currently under investigation (Rödel *et al.* 2007).

8. Discussion

Maximally selected statistics for the estimation of simple cutpoint models have been in use since many years. Many researchers appreciate a model that is easy to communicate and implement in practice. Of course, the trade-off between simplicity and accuracy has to be carefully investigated.

A new class of generalized maximally selected statistics based on the conditional inference framework of Strasser and Weber (1999) allows for a unified treatment of different kinds of maximally

selected statistics. Test procedures from this framework can be adapted to new test problems by specifying an influence function h, suitable for the scale level of the response, and setting up a set of candidate partitions g determined from the available covariates. As the number of potential partitions can become large, efficient algorithms are required for evaluating the distribution of the maximum statistic. For partitions based on cutpoints, we provide such an algorithm that computes the asymptotic distribution in linear time.

The implementation of (known and newly designed) maximally selected statistics only requires the specification of the binary candidate partitions, via a function g, and a problem-specific influence function h. Linear statistics $\mathbf{T}$ and the test statistic $T_{\max}$ can be computed in the R system (R Development Core Team 2007) utilizing the function `maxstat_test()` from package **coin** (Hothorn *et al.* 2006, 2007) in which approximations for the distribution of $T_{\max}$ are readily available, both via the asymptotic distribution and Monte-Carlo methods (also in the presence of blocks, e.g., in multicenter trials).

In summary, a unified treatment of maximally selected statistics for nominal, ordered, discrete and continuous numeric, censored and multivariate response variables as well as nominal, ordered and multivariate covariates to be dichotomized is now possible both conceptually and practically.

Acknowledgments

We would like to thank the editor, associate editor and a reviewer for helpful comments and A. Genz for pointing out the special product structure of the covariance matrix. C. Rödel kindly gave us permission to use the CAO/ARO/AIO–94 trial data. The work of T. Hothorn was supported by Deutsche Forschungsgemeinschaft (DFG) under grant HO 3242/1–3.

References

Betensky RA, Rabinowitz D (1999). "Maximally Selected χ^2 Statistics for $k \times 2$ Tables." *Biometrics*, **55**, 317–320.

Boulesteix AL (2006a). "Maximally Selected Chi-square Statistics and Binary Splits of Nominal Variables." *Biometrical Journal*, **48**, 838–848.

Boulesteix AL (2006b). "Maximally Selected Chi-square Statistics for Ordinal Variables." *Biometrical Journal*, **48**, 451–462.

Boulesteix AL, Strobl C (2007). "Maximally selected Chi-squared statistics and non-monotonic associations: An exact approach based on two cutpoints." *Computational Statistics & Data Analysis*, **51**, 6295–6306.

Boulesteix AL, Strobl C, Weidinger S, Wichmann HE, Wagenpfeil S (2007). "Multiple testing for SNP-SNP interactions." *Technical report*, LMU, München. URL `http://www.statistik.lmu.de/~carolin/`.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Wadsworth, California.

Buccisano F, Maurillo L, Gattei V, Del Poeta G, Del Principe MI, Cox MC, Panetta P, Consalvo MI, Mazzone C, Neri B, Ottaviani L, Fraboni D, Tamburini A, Lo-Coco F, Amadori S, Venditti A (2006). "The Kinetics of Reduction of Minimal Residual Disease Impacts on Duration of Response and Survival of Patients with Acute Myeloid Leukemia." *Leukemia*, **20**, 1783–1789.

De'ath G (2002). "Multivariate Regression Trees: A New Technique For Modeling Species-Environment Relationships." *Ecology*, **83**(4), 1105–1117.

Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C, Tosolini M, Camus M, Berger A, Wind P, Zinzindohoue F, Bruneval P, Cugnenc PH, Trajanoski Z, Fridman WH, Pages F (2006). "Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome." *Science*, **313**, 1960–1964.

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Genz A, Kahaner DK (1986). "The Numerical Evaluation of Certain Multivariate Normal Integrals." *Journal of Computational and Applied Mathematics*, **16**, 255–258.

Hájek J, Šidák Z, Sen PK (1999). *Theory of Rank Tests*. 2nd edition. Academic Press, London.

Hansen BE (1997). "Approximate Asymptotic p Values for Structural-Change Tests." *Journal of Business & Economic Statistics*, **15**, 60–67.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2007). **coin***: Conditional Inference Procedures in a Permutation Test Framework*. R package version 0.6-7. URL http://CRAN.R-project.org/.

Hothorn T, Lausen B (2003). "On the Exact Distribution of Maximally Selected Rank Statistics." *Computational Statistics & Data Analysis*, **43**(2), 121–137.

Huggett AJ (2005). "The Concept and Utility of 'Ecological Thresholds' in Biodiversity Conservation." *Biological Conservation*, **124**, 301–310.

Jennen C (1985). "Second-Order Approximations to the Density, Mean and Variance of Brownian First-Exit Times." *Annals of Probability*, **13**, 126–144.

Lausen B, Hothorn T, Bretz F, Schumacher M (2004). "Assessment of Optimal Selected Prognostic Factors." *Biometrical Journal*, **46**(3), 364–374.

Lausen B, Schumacher M (1992). "Maximally Selected Rank Statistics." *Biometrics*, **48**, 73–85.

Lausen B, Schumacher M (1996). "Evaluating the Effect of Optimized Cutoff Values in the Assessment of Prognostic Factors." *Computational Statistics & Data Analysis*, **21**(3), 307–326.

Meurant G (1992). "A Review on the Inverse of Symmetric Tridiagonal and Block Tridiagonal Matrices." *SIAM Journal on Matrix Analysis and Applications*, **13**(3), 707–728.

Miller R, Siegmund D (1982). "Maximally Selected Chi Square Statistics." *Biometrics*, **38**, 1011–1016.

Miwa T, Hayter AJ, Liu W (2000). "Calculations of Level Probabilities for Normal Random Variables with Unequal Variances with Applications to Bartholomew's Test in Unbalanced One-Way Models." *Computational Statistics & Data Analysis*, **32**, 17–32.

Müller J, Hothorn T (2004). "Maximally Selected Two-Sample Statistics as a New Tool for the Identification and Assessment of Habitat Factors with an Application to Breeding-Bird Communities in Oak Forests." *European Journal of Forest Research*, **123**, 219–228.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Rödel C, Hothorn T, Fietkau R, Liersch T, Wittekind C, Sauer R (2007). "Prognostic significance of TNM-staging after neoadjuvant chemoradiotherapy for rectal cancer." *Strahlentherapie und Onkologie*, **183**(Suppl. 1), 113–113.

Royston P, Altman DG, Sauerbrei W (2006). "Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea." *Statistics in Medicine*, **25**, 127–141.

Sauer R, Becker H, Hohenberger W, Rödel C, Wittekind C, Fietkau R, Martus P, Tschmelitsch J, Hager E, Hess CF, Karstens JH, Liersch T, Schmidberger H, Raab R (2004). "Preoperative Versus Postoperative Chemoradiotherapy for Rectal Cancer." *New England Journal of Medicine*, **351**, 1731–1740.

Sobin LH, Wittekind C (2002). *TNM Classification of Malignant Tumours*. 6th edition. Wiley-Liss, New York.

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Statistics." *Mathematical Methods of Statistics*, **8**, 220–250.

Worsley KJ (1982). "An Improved Bonferroni Inequality and Applications." *Biometrika*, **69**(2), 297–302.