Maximally selected Chi-squared statistics and non-monotonic associations: An exact approach based on two cutpoints

Anne-Laure Boulesteix[a,*], Carolin Strobl[b]

[a]*Department of Medical Statistics and Epidemiology, Technical University of Munich, Ismaningerstr. 22, D-81675 Munich, Germany*
[b]*Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany*

Abstract

Binary outcomes that depend on an ordinal predictor in a non-monotonic way are common in medical data analysis. Such patterns can be addressed in terms of cutpoints: for example, one looks for two cutpoints that define an interval in the range of the ordinal predictor for which the probability of a positive outcome is particularly high (or low). A Chi-squared test may then be performed to compare the proportions of positive outcomes in and outside this interval. However, if the two cutpoints are chosen to maximize the Chi-squared statistic, referring the obtained Chi-squared statistic to the standard Chi-squared distribution is an inappropriate approach. It is then necessary to correct the *p*-value for multiple comparisons by considering the distribution of the maximally selected Chi-squared statistic instead of the nominal Chi-squared distribution. The exact distribution of the Chi-squared statistic obtained with the two optimal cutpoints is derived based on combinatorial considerations. This approach is illustrated by a simulation study and an application to varicella data.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Umbrella ordering; Ordinal; Contingency table; Adjustment; Multiple testing

1. Introduction

Suppose we have a binary outcome Y ($Y = 0, 1$) and an at least ordinally scaled predictor variable X that is suspected to be associated with Y. In medical applications, there is often interest in testing independence of X and Y against an ordered alternative e.g., in dosis-response problems. Some widely used methods for testing for trends in $2 \times K$ ordered tables are, e.g., the Cochran–Armitage test, the Cochran–Mantel–Haenszel test, rank tests such as the Jonckheere–Terpstra test and the Wilcoxon rank sum test, or approaches based on isotonic regression (Robertson et al., 1988; Salanti and Ulm, 2003). In machine learning, such associations are often examined based on binary splits of the form $\{X \leqslant a\}$ vs. $\{X > a\}$. Patterns of the form $\{a < X \leqslant b\}$ vs. $\{X \leqslant a \cup X > b\}$ are also conceivable, for example if the probability $P(Y = 1 | X = x)$ is higher for $x \in]a, b]$ than for $x \in]-\infty, a] \cup]b, +\infty[$.

Such dependence structures are related to 'umbrella orderings' if the conditional probability $P(Y = 1 | x)$ is larger for intermediate x values than for small and large x values and to 'U-shapes' if vice versa. Umbrella orderings or U-shapes may be observed in medical research when X is a predictor such as the age (in a broad sense) and Y denotes e.g.,

---

* Corresponding author. Tel.: +49 89 4140 4347.
E-mail address: anne-laure.boulesteix@tum.de (A.-L. Boulesteix).

the occurrence of complications. For instance, many diseases are known to be more severe for both infants and elder patients than for young adults. Another interesting example is varicella. This disease is not equally serious for all age categories: it rather shows a non-monotonic pattern. Similarly, perinatal morbidity and mortality are higher for both premature and post-term babies than for babies born at term. Another typical example of umbrella ordering in medical research is the effect of different doses of chemicals on the occurrence of tumors: some chemicals show evidence of an increasing monotonic trend with a downturn for high doses due to the inhibition of the tumor development by the toxic effect (Hans and Dunson, 2005).

There have been various proposals to assess such downturns or upturns. This problem is also commonly denoted as change-point detection. A recent reference including a brief review of various methods for monotonic responses is Hans and Dunson (2005). They suggest a bayesian inference method that addresses explicitly the problem of downturns with applications to a carcinogenesis study. In the framework of maximally selected statistics, Lausen et al. (2002) generalize the asymptotic results of Lausen and Schumacher (1992) on maximally selected rank statistics to ordinal predictors and examine, e.g., umbrella alternatives.

Both approaches focus on the downturn, but do not give any information on the cutpoints defining the high-risk and low-risk intervals. Moreover, they might not be applicable to small sample sizes. In the last few years, cutpoint-based strategies have been sometimes criticized for ignoring a large part of the information contained in the data. Moreover, setting an artificial cutpoint when there is no cutpoint but rather a smooth transition is critical. While dichotomizing continuous predictors may be controversial depending on the considered problem (Royston et al., 2006), cutpoints might be useful in the case of ordinal predictors with few distinct values, or to support medical decisions and diagnostics. Moreover, cutpoint selection is a crucial issue in classification tree algorithms such as CART by Breiman et al. (1984) and, more generally, in all the machine learning methods based on recursive partitioning.

In this paper, we address the assessment of cutpoints defining low or high-risk intervals, based on the principle of maximally selected statistics. Our approach considers a special case of umbrella orderings that is known as the 'epidemic wave model' (Siegmund, 1986). From now on, we consider a variable X with $K > 2$ ordered categories which are denoted as $1, \ldots, K$. Suppose one selects the pair of cutpoints (k_1, k_2) that maximizes the Chi-squared statistic obtained from Table 1.

This resulting p-value must be interpreted with caution. Claiming that the cutpoints k_1 and k_2 are relevant because the p-value is low is incorrect. Indeed, the distribution of the maximally selected Chi-squared statistic is different from the nominal Chi-squared distribution with one degree of freedom on which the Chi-squared test is based. Maximally selected statistics and minimally selected p-values have been the subject of numerous articles in the case of one cutpoint. Miller and Siegmund (1982) show that the maximally selected Chi-squared statistic converges to a normalized Brownian bridge under the null-hypothesis of no association between X and Y. The distribution of the maximally selected Chi-squared statistic in the small sample case is examined by Halpern (1982) in a simulation study, while Koziol (1991) derives the exact distribution of maximally selected Chi-squared statistics using a combinatorial approach. Maximally selected rank statistics are investigated in Lausen and Schumacher (1992, 1996) and Hothorn and Lausen (2003). Holländer et al. (2004) address confidence intervals for the effect of prognostic factors after optimal cutpoint selection. The exact distribution of the maximally selected Chi-squared statistic in the context of a binary Y and an at least ordinally scaled X with ties is derived in Boulesteix (2006b). An exact approach to handle the case of optimally selected splits of a nominal variable is given in Boulesteix (2006a). The underlying idea of these papers is that the p-value obtained from the optimal cutpoint or split has to be adjusted to account for the multiple testing effect. Applications of the theory of maximally selected statistics to recursive partitioning algorithms are discussed by, e.g., Shih (2004) and Lausen et al. (2004). All these articles address the case of one optimally selected cutpoint.

Matters become much more complicated when several cutpoints are chosen optimally and simultaneously, e.g., for recombination detection in DNA sequences. Assessing the distribution of maximally selected statistics in this

Table 1
Contingency table obtained by cutting X at k_1 and k_2

	$X \leqslant k_1$ or $X > k_2$	$k_1 < X \leqslant k_2$	Σ
$Y = 0$	n_1	n_2	N_0
$Y = 1$	n_3	n_4	N_1

situation is a very difficult task. Hence, existing approaches are often based on simulations (Halpern, 2000). Another related method is Kuiper's goodness-of-fit test (Kuiper, 1960). In the two-sample case, it tests the equality of two continuous distribution functions (corresponding to $Y = 0$ and 1, respectively) based on empirical distributions. In the two-cutpoint framework described in the present article, Kuiper's test is expected to have higher power than the more usual Kolmogorov–Smirnov test. However, it is not as easy to interpret as maximally selected statistics. Practitioners expect simple conclusions such as 'the risk is significantly higher if $a < X \leqslant b$ than if $X \leqslant a$ or $X > b$' as supported by the (adjusted) p-value of the Chi-squared test. Another inconvenience of Kuiper's test is that it does not account for ties. In the case of ordinal variables with few categories, ties cannot be ignored. In the present article, we suggest an approach that overcomes these two problems.

We propose a new combinatorial approach to derive the exact distribution of the maximally selected Chi-squared statistic in the two-cutpoint framework. Our novel procedure is distribution-free and can be applied in the case of a binary Y and an at least ordinally scaled X. It is appropriate to analyze samples with moderate or small sizes with predictors taking only a few values, e.g., $X \in \{1, 2, 3, 4, 5, 6\}$, but becomes computationally intensive when K approaches N. It is easily generalizable to other association statistics for 2×2 contingency tables.

The rest of the paper is organized as follows. Our approach to derive the exact distribution of the maximally selected Chi-squared statistic in the two-cutpoints framework is presented in Section 2, including a discussion of computational aspects. In Section 3, the new approach is compared via simulations to the comparable method for maximally selected Chi-squared statistics with one cutpoint by Boulesteix (2006b). Section 4 gives an illustration through an application to varicella data.

2. Derivation of the exact distribution

2.1. Notations

Let $(x_i, y_i)_{i=1,\ldots,N}$ denote N independent observations of X and Y. N_0 and N_1 denote the numbers of observations with $y_i = 0$ and $y_i = 1$, respectively, and m_k ($k = 1, \ldots, K$) the number of observations with $x_i = k$, whereas m_{ck} ($c = 0, 1, k = 1, \ldots, K$) is the number of observations with $y_i = c$ and $x_i = k$. The association between X and Y may be visualized by plotting the graph m_{1k}/m_k, $k = 1, \ldots, K$. Extreme examples are depicted in Fig. 1 for $N_0 = N_1 = 30$ and $m_1 = m_2 = m_3 = m_4 = m_5 = m_6 = 10$. An approximately horizontal graph of type (a) indicates poor association between X and Y. Types (b) and (c) correspond to strong monotonic associations, whereas (d) and (e) display non-monotonic association patterns with two underlying cutpoints. We consider splits of X involving two cutpoints k_1 and k_2 of the
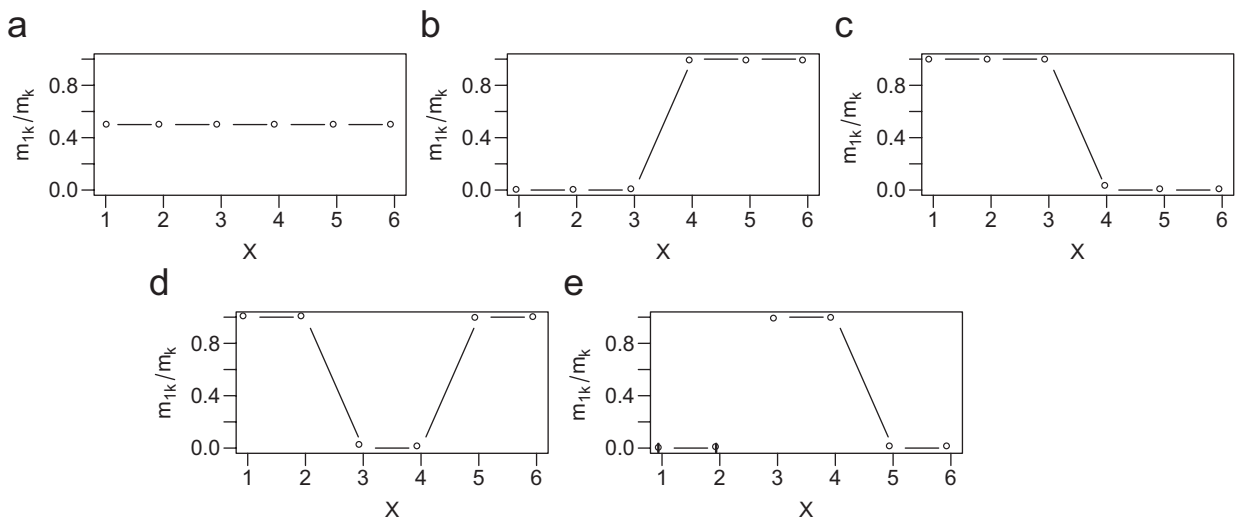

Fig. 1. Proportion of observations with $Y = 1$ in each category of X : m_{1k}/m_k, $k = 1, \ldots, K$. $N_0 = N_1 = 30$, $m_1 = m_2 = m_3 = m_4 = m_5 = m_6 = 10$.

form $\{k_1 < X \leqslant k_2\}$ vs. $\{X \leqslant k_1 \cup X > k_2\}$. The set $\mathcal{K}$ of the possible pairs of cutpoints is denoted as

$$\mathcal{K} = \{(k_1, k_2) | 1 \leqslant k_1 \leqslant K - 1; k_1 + 1 \leqslant k_2 \leqslant K\}.$$

Note that the splits involving only one cutpoint of the type $\{X \leqslant k_1\}$ vs. $\{X > k_1\}$ are a special case corresponding to $k_2 = K$. The usual Chi-squared statistic for 2×2 contingency tables computed from Table 1 is denoted as $\chi^2_{k_1, k_2}$. It can be written as

$$\chi^2_{k_1, k_2} = \frac{N(n_1 n_4 - n_2 n_3)^2}{(n_1 + n_2)(n_3 + n_4)(n_1 + n_3)(n_2 + n_4)}. \tag{2.1}$$

In this paper, we consider the Chi-squared statistic obtained by selecting the pair of cutpoints $(k_1, k_2) \in \mathcal{K}$ maximizing $\chi^2_{k_1, k_2}$:

$$\chi^2_{\max} = \max_{(k_1, k_2) \in \mathcal{K}} \chi^2_{k_1, k_2}. \tag{2.2}$$

The rest of Section 2 deals with the computation of the exact distribution of $\chi^2_{\max}$ under the null-hypothesis of no association between X and Y, given $N_0, N_1, m_1, \ldots, m_K$. Note that N_0, N_1, and $m_1, \ldots, m_K$ can be seen as fixed distribution parameters. For simplification, we use the notation $F(d) = P_{H_0}(\chi^2_{\max} \leqslant d)$ throughout the paper.

2.2. The naive exact approach

Let us consider the $\binom{N}{N_1}$ ways to draw N_1 out of N observations, which are denoted as 'configurations' in the rest of this section. Let $\mathcal{C}(d)$ denote the set of the configurations yielding $\chi^2_{\max} > d$ and $c(d)$ its cardinal number. The probability $F(d) = P_{H_0}(\chi^2_{\max} \leqslant d)$ is obtained as

$$F(d) = P_{H_0}(\chi^2_{\max} \leqslant d) = 1 - \frac{c(d)}{\binom{N}{N_1}},$$

since all the configurations are equally likely under the null-hypothesis. The naive exact approach to compute $c(d)$ consists of enumerating all the $\binom{N}{N_1}$ configurations and computing $\chi^2_{\max}$ for each of them. Since $\chi^2_{\max}$ depends only on $m_{11}, \ldots, m_{1K}$ and not on the arrangement of the observations with $Y = 1$ within each category, the computational complexity may be reduced by enumerating the possible vectors $(m_{11}, \ldots, m_{1K})$ instead of all $\binom{N}{N_1}$ configurations. By possible vectors, we mean vectors of positive integers summing to N_1, such that $m_{1k} \leqslant m_k$, for $k = 1, \ldots, K$. For a fixed vector $(m_{11}, \ldots, m_{1K})$, the number of configurations is given as $\prod_{k=1}^{K} \binom{m_k}{m_{1k}}$. Enumerating all the possible vectors $(m_{11}, \ldots, m_{1K})$ and computing the value of $\chi^2_{\max}$ and the number of configurations for each of them is computationally prohibitive, even for moderate N and K. Storage requirements turn out to exceed the capacity of modern computers, since a huge integer has to be stored for each of the possible vectors $(m_{11}, \ldots, m_{1K})$, whose number grows with N^K.

In the next section, a faster algorithm for computing $c(d)$ is presented.

2.3. A novel fast algorithm

The novel algorithm is based on two ideas: (i) the reformulation of the inequation $\chi^2 > d$ in terms of boundary functions (see Section 2.3.1), (ii) the conversion of the two-cutpoint problem into several one-cutpoint problems (see Section 2.3.2).

2.3.1. Boundary functions

Suppose we split the available sample of size N into two complementary sets A and $\overline{A}$ of size N_A and $N - N_A$, respectively. Let m_{1A} denote the number of observations from A with $Y = 1$. The Chi-squared statistic yielded by this split is given as

$$\chi^2 = \frac{N(m_{1A}(N_0 - N_A + m_{1A}) - (N_A - m_{1A})(N_1 - m_{1A}))^2}{N_0 N_1 N_A (N - N_A)}.$$

Via expensive but simple computations (Boulesteix, 2006b), it can be shown that

$$\chi^2 > d \Leftrightarrow \begin{cases} m_{1A} > f_+(N_A) \\ \text{or} \\ m_{1A} < f_-(N_A), \end{cases} \tag{2.3}$$

where f_+ and f_- are functions that depend on N_0, N_1 and d:

$$f_{\chi^+}(t) = \frac{N_1 t}{N} + \frac{N_0 N_1 \sqrt{d}}{N} \sqrt{\frac{i}{N}\left(1 - \frac{i}{N}\right)\left(\frac{1}{N_0} + \frac{1}{N_1}\right)},$$

$$f_{\chi^-}(t) = \frac{N_1 t}{N} - \frac{N_0 N_1 \sqrt{d}}{N} \sqrt{\frac{i}{N}\left(1 - \frac{i}{N}\right)\left(\frac{1}{N_0} + \frac{1}{N_1}\right)}.$$

Note that the generalization of our method to other maximally selected criteria (or minimally selected p-values) is done by replacing f_{χ^+} and f_{χ^-} by appropriate functions f_+ and f_- derived from the definition of the considered association statistic. For instance, f_+ and f_- are derived by Strobl et al. (2007) for the Gini gain criterion (Breiman et al., 1984) used for split selection in many recursive partitioning algorithms.

The next section presents an efficient algorithm to compute $c(d)$ based on Eq. (2.3).

2.3.2. Converting the two-cutpoint problem into a one-cutpoint problem

The principle underlying our algorithm consists of decomposing $\mathscr{C}(d)$ into disjoint sets by recoding X into pseudo-variables $X^{(1)}, \ldots, X^{(K-1)}$, as illustrated below for the case $K = 6$. Recoding is performed such that the smallest and the largest values of X are coded using consecutive numbers. For $k = 1, \ldots, K - 1$, let $X^{(k)}$ denote the variable taking the value

$$X^{(k)} = \sigma^{(k)}(X),$$

where $\sigma^{(k)}$ is the permutation defined by

$$\sigma^{(k)}(i) = i \qquad \text{if } i \leqslant k,$$
$$= K - i + k + 1 \quad \text{if } i > k.$$

Note that we have $X = X^{(K-1)}$. As an example, for $K = 6$, the six categories 1, 2, 3, 4, 5, 6 are recoded successively as 1, 6, 5, 4, 3, 2 ($X^{(1)}$), 1, 2, 6, 5, 4, 3 ($X^{(2)}$), 1, 2, 3, 6, 5, 4 ($X^{(3)}$) and 1, 2, 3, 4, 6, 5 ($X^{(4)}$). The double inequation $a < X \leqslant b$ is then equivalent to $X^{(a)} > a + (K - b)$, for all $a = 1, \ldots, K - 1$ and $b = a + 1, \ldots, K$. Using the pseudo-variables $X^{(1)}, \ldots, X^{(K-1)}$, we have thus transformed our two-cutpoint problem into $K - 1$ one-cutpoint problems.

$\mathscr{C}(d)$ can be decomposed as $K - 1$ disjoint subsets

$$\mathscr{C}(d) = \bigcup_{k=1}^{K-1} \mathscr{C}_k(d), \tag{2.4}$$

where $\mathscr{C}_k(d)$ ($k = 1, \ldots, K$) denotes the subset of configurations fulfilling the following conditions:

A1. There exists a split of the variable $X^{(k)}$ yielding $\chi^2 > d$.
   For example, if $K = 6$ and $k = 2$, at least one of the splits $\{1\}\{2, 6, 5, 4, 3\}$, $\{1, 2\}\{6, 5, 4, 3\}$, $\{1, 2, 6\}\{5, 4, 3\}$, $\{1, 2, 6, 5\}\{4, 3\}$, $\{1, 2, 6, 5, 4\}\{3\}$ has to yield $\chi^2 > d$.
A2. For all $k' < k$, the splits of the variable $X^{(k')}$ yield $\chi^2 \leqslant d$ (hence, $\mathscr{C}_1(d), \ldots, \mathscr{C}_K(d)$ are disjoint).
   For example, if $K = 6$ and $k = 2$, the splits $\{1\}\{6, 5, 4, 3, 2\}$, $\{1, 6\}\{5, 4, 3, 2\}$, $\{1, 6, 5\}\{4, 3, 2\}$, $\{1, 6, 5, 4\}\{3, 2\}$ and $\{1, 6, 5, 4, 3\}\{2\}$ (corresponding to splits of $X^{(1)}$) have to yield $\chi^2 \leqslant d$.

Since $\mathscr{C}_1(d), \ldots, \mathscr{C}_K(d)$ are disjoint, we have

$$c(d) = \sum_{k=1}^{K-1} c_k(d),$$

where $c_k(d)$ is defined as $c_k(d) = |\mathscr{C}_k(d)|$. For $k = 1$, A2 is not relevant and $c_k(d)$ is the number of configurations satisfying A1. It can be efficiently computed based on the method for maximally selected Chi-squared statistics for ordinal variables proposed by Boulesteix (2006b), since $X^{(1)}$ is an ordinal variable. We refer to Boulesteix (2006b) for a description of the algorithm.

The rest of this section presents a new algorithm to compute $c_k(d)$ for $k > 1$ for $d \geqslant 0$. For a fixed $k = 1, \ldots, K - 1$, $\mathscr{C}_k(d)$ may also be decomposed into $K - k$ disjoint subsets $\mathscr{C}_{k,k}(d), \ldots, \mathscr{C}_{k,K-1}(d)$:

$$\mathscr{C}_k(d) = \bigcup_{i=k}^{K-1} \mathscr{C}_{k,i}(d), \tag{2.5}$$

where $\mathscr{C}_{k,i}(d)$ denotes the subset of configurations out of $\mathscr{C}_k(d)$ for which the two following conditions are fulfilled.

B1. The split $\{X^{(k)} \leqslant i\}$ vs. $\{X^{(k)} > i\}$ yields $\chi^2 > d$.
     For example, if $K = 6$, $k = 2$ and $i = 4$, the split $\{1, 2, 6, 5\}\{4, 3\}$ has to yield $\chi^2 > d$.
B2. For all $i' < i$, the split $\{X^{(k)} \leqslant i'\}$ vs. $\{X^{(k)} > i'\}$ yields $\chi^2 \leqslant d$.
     For example, if $K = 6$, $k = 2$ and $i = 4$, the splits $\{1\}\{2, 6, 5, 4, 3\}$, $\{1, 2\}\{6, 5, 4, 3\}$ and $\{1, 2, 6\}\{5, 4, 3\}$ have to yield $\chi^2 \leqslant d$.

In Eq. (2.5), the index covers $k, \ldots, K - 1$. It can be explained as follows. If the split $\{X^{(k)} \leqslant i\}$ vs. $\{X^{(k)} > i\}$ $(i < k)$ yields $\chi^2 > d$, then the split $\{X^{(i)} \leqslant i\}$ vs. $\{X^{(i)} > i\}$ also yields $\chi^2 > d$ and the considered configurations are not in $\mathscr{C}_k(d)$ but in $\mathscr{C}_i(d)$. Note that this is a consequence of the definition of the permuted variables $X^{(1)}, \ldots, X^{(K-1)}$.

For fixed i and $k > 1$, $c_{k,i}(d) = |\mathscr{C}_{k,i}(d)|$ is computed as follows:

$$c_{k,i}(d) = \sum_{i_1 \in I_{k,i}(d)} \binom{m_{\sigma^{(k)}(1)}}{i_1} \cdot \left(\sum_{i_2 \in I_{k,i}(d,i_1)} \binom{m_{\sigma^{(k)}(2)}}{i_2} \cdots \left(\sum_{i_K \in I_{k,i}(d,i_1,\ldots,i_{K-1})} \binom{m_{\sigma^{(k)}(K)}}{i_K} \right) \cdots \right), \tag{2.6}$$

where the integers $i_1, \ldots, i_K$ correspond to the numbers of observations with $Y = 1$ in the categories $\sigma^{(k)}(1), \ldots,$ $\sigma^{(k)}(K)$, respectively, and $I_{k,i}(d, i_1, \ldots, i_j)$ may be seen as the allowed interval for i_{j+1}, given the numbers $i_1, \ldots, i_j$ of observations with $Y = 1$ within the categories $X = \sigma^{(k)}(1), \ldots, \sigma^{(k)}(j)$.

For fixed k, i and j, the interval $I_{k,i}(d, i_1, \ldots, i_j)$ is obtained by eliminating from $\{0, \ldots, m_{\sigma^{(k)}(j+1)}\}$ the values for which at least one of the following conditions is violated:

C1. The split $\{X^{(k)} \leqslant i\}$ vs. $\{X^{(k)} > i\}$ yields $\chi^2 > d$ (corresponding to B1).
C2. The splits $\{X^{(k')} \leqslant i'\}$ vs. $\{X^{(k')} > i'\}$, for $k' < k$ and $i' = k', \ldots, K - 1$ yield $\chi^2 \leqslant d$ (corresponding to A2).
C3. The splits $\{X^{(k)} \leqslant i'\}$ vs. $\{X^{(k)} > i'\}$, for $i' < i$, yield $\chi^2 \leqslant d$ (corresponding to B2).

Note that, in general, it may be impossible to determine whether C1–3 are fulfilled or not without fixing $i_{j+1}, \ldots, i_{K-1}$. At this point we only eliminate from $\{0, \ldots, m_{\sigma^{(k)}(j+1)}\}$ those values for which we can already tell that they lead to violation of C1–3. From a theoretical point of view, the result of the computation of $c_{k,i}(d)$ would be the same if C1–3 were checked only in the very end, once all $i_1, \ldots, i_{K-1}$ are fixed. But then the method would be essentially almost as complex as the naive method. Immediately eliminating those values from $I_{k,i}(d, i_1, \ldots, i_j)$ for which we already know that they violate C1, C2 or C3 is a computational trick that reduces the complexity considerably.

The intervals $I_{k,i}(d, i_1, \ldots, i_j)$ are derived using the functions f_+ and f_- defined in Section 2.3.1. Let us explain the principle based on the example of $I_{2,4}(d, i_1, i_2, i_3)$. For $K = 6$ and $k = 2$, we have $\sigma^{(2)}(1) = 1$, $\sigma^{(2)}(2) = 2$, $\sigma^{(2)}(3) = 6$, $\sigma^{(2)}(4) = 5$, $\sigma^{(2)}(5) = 4$, $\sigma^{(2)}(6) = 3$. $I_{2,4}(d, i_1, i_2, i_3)$ is the allowed interval for the number of observations with $Y = 1$ in category $X = \sigma^{(2)}(4) = 5$, given the numbers i_1, i_2, i_3 of observations with $Y = 1$ in categories 1, 2, 6. C1 holds if the

split $\{X^{(2)} \leqslant 4\}$ vs. $\{X^{(2)} > 4\}$ yields $\chi^2 > d$. Out of the $m_1 + m_2 + m_6 + m_5$ observations with $X^{(2)} \leqslant 4$, $i_1 + i_2 + i_3 + i_4$ observations have $Y = 1$. Thus, C1 holds if and only if

$$i_4 > f_+(m_1 + m_2 + m_6 + m_5) - (i_1 + i_2 + i_3)$$

or

$$i_4 < f_-(m_1 + m_2 + m_6 + m_5) - (i_1 + i_2 + i_3).$$

These inequations yield the values of i_4 for which C1 does not hold. Similarly, for given i_1, i_2, i_3, violation of C2 and C3 can be simply reformulated in terms of i_4, thus yielding the interval $I_{2,4}(d, i_1, i_2, i_3)$. The other intervals $I_{k,i}(d, i_1, \ldots, i_j)$ are derived based on the same principle.

Computation time can be spared by computing and storing $\binom{m_k}{j}$, for $k = 1, \ldots, K$ and $j = 0, \ldots, m_k$ before applying Eq. (2.6).

After computation of $c_{k,i}(d)$, for $k = 1, \ldots, K - 1$ and $i = k, \ldots, K - 1$, $c(d)$ is obtained as their sum: since $\mathscr{C}_{k,k}(d), \ldots, \mathscr{C}_{k,K-1}(d)$ are disjoint, we have

$$c_k(d) = \sum_{i=k}^{K-1} c_{k,i}(d),$$

and thus

$$c(d) = \sum_{k=1}^{K-1} \sum_{i=k}^{K-1} c_{k,i}(d).$$

Finally, $F(d)$ is given as

$$F(d) = P_{H_0}(\chi^2_{\max} \leqslant d) = 1 - \frac{\sum_{k=1}^{K-1} \sum_{i=k}^{K-1} c_{k,i}(d)}{\binom{N}{N_1}}.$$

Note that our method computes the distribution function $F(d)$ at a single value d. If the full distribution is needed, the algorithm should be run several times. In the case of a very small N and very small K, it might be faster to use the naive method, which directly yields the full distribution of the maximally selected Chi-squared statistic. However, as discussed in the next section, the naive method becomes unfeasible for increasing N and K. In the next section, we discuss the advantages of our novel algorithm over the naive exact method in terms of computational complexity and memory requirements.

2.4. Computation time

In this section, the computation time of our novel algorithm is compared to the naive approach for various parameter combinations. Table 2 gives the elapsed time (rounded to the nearest second) as output by the R function `system.time` for both approaches.

- In contrast to the naive approach, the computation time for the new algorithm increases with d. This can be explained as follows. For small values of d, the set $\mathscr{C}_1(d)$ is large. $\mathscr{C}_2(d), \ldots, \mathscr{C}_{K-1}(d)$ are then small, because the sets $\mathscr{C}_k(d)$ are disjoint. Since the computation of $c_1(d)$ based on the method described in Boulesteix (2006b) is much faster than that of $c_k(d)$ $(k > 1)$, the overall computation time is larger for large d values.
- With both approaches, the computation time is larger if the X categories have the same numbers of observations than in the unbalanced case.
- With both approaches, the computation time is slightly larger if $N_0 = N_1$ than if $N_0 \neq N_1$.
- In all situations, the novel algorithm is much more efficient than the naive approach. The difference between novel algorithm and naive approach increases with N and K.

Table 2
Elapsed time (rounded to the nearest 1/100 second) for the naive approach (left) and our new algorithm (right) to compute $F(d) = P_{H_0}(\chi^2_{\max} \leqslant d)$, given $N_0, N_1, m_1, \ldots, m_K$

$m_1, \ldots, m_K$	N_0, N_1	d	Naive exact approach	Novel exact algorithm
10,10,10,10	20,20	0	5.5	0.10
		2	5.5	0.14
		10	5.5	0.19
4,16,4,16	20,20	0	3.0	0.03
		2	3.0	0.08
		10	3.0	0.13
10,10,10,10	10,30	0	2.6	0.01
		2	2.6	0.08
		10	2.6	0.08
20,20,20,20	40,40	0	40	0.08
		2	40	0.56
		10	40	1.2
8,8,8,8,8	20,20	0	55	0.07
		2	55	0.8
		10	55	1.2
10,10,10,10,10	25,25	0	90	0.08
		2	90	1.5
		10	90	3.0
8,8,8,8,8,8	24,24	0	810	0.11
		2	810	5.7
		10	810	14
10,10,10,10,10,10	30,30	0	1500	0.14
		2	1500	12
		10	1500	42

For larger values of N and K, the naive approach rapidly becomes prohibitive. In contrast, the new exact algorithm can be easily applied to sample sizes larger than 100 and to X variables with up to eight categories. It can roughly be explained as follows:

- The novel algorithm takes into account the ordinality of X by using the fast method for ordinal splits given in Boulesteix (2006b) in the computation of $c_1(d)$.
- Only the configurations yielding $P(\chi^2_{\max} > d)$ are enumerated and counted.
- Through the use of the boundary functions f_+ and f_-, $\chi^2_{\max}$ is not computed for each vector $(m_{11}, \ldots, m_{1K})$.

In the most favorable case (small d), the complexity of the novel algorithm is essentially polynomial in N since the problem reduces to computing $c_1(d)$, see Strobl et al. (2007). With growing d, complexity increases but remains far under N^K, since the size of the intervals $I_{k,i}(d, i_1, \ldots, i_j)$ decreases strongly with d. Note that the algorithm applies to the case of ordinal variables with a small number of categories only, because it becomes expensive with increasing K.

3. Simulation study

We implemented our method in the language R. Our implementation is included in the freely available R package exactmaxsel as a function Ford2. Before starting the power study addressed in the next section, we outline how the correctness of our novel combinatorial method can be assessed based on simulations. For fixed marginal conditions

Table 3
Performance of both methods for the one-cutpoint design

Effect	One-cutpoint method			Novel method		
	Weak	Med.	Strong	Weak	Med.	Strong
Marginal	6	19	52	7	14	43
Central	15	41	74	12	33	70

$N_0, N_1, m_1, \ldots, m_k$, a large number of data sets, say, 10 000 are generated under the null-hypothesis and the maximal Chi-squared statistic is derived for each of the 10 000 data sets. This yields an estimate of the distribution function of the maximally selected Chi-squared statistic given $N_0, N_1, m_1, \ldots, m_K$. For the same values of N_0, N_1 and $m_1, \ldots, m_K$, the theoretical distribution function is computed via our novel combinatorial approach. For example, after installation of the `exactmaxsel` package, the value of the distribution function at $d = 1.5$, for $N_0 = N_1 = 30$ and $m_1 = \cdots = m_5 = 12$ is obtained by

```
> library(exactmaxsel)

> Ford2(1.5,n0 = 30,n1 = 30,A = c(12,12,12,12,12),statistic = "chi2")
```

Extensive simulations involving different values of $N_0, N_1, m_1, \ldots, m_K$, are conducted. The obtained empirical distribution is always perfectly consistent with the theoretical distribution computed using our novel combinatorial procedure (due to space constraints the results are not shown but are easily replicable using the function call stated above).

A simulation study is conducted to evaluate the power of our novel method to detect non-monotonic association with a two-cutpoint pattern. The performance of our new method is compared to that of the method designed for one cutpoint by (Boulesteix, 2006b). For all simulations, we set the total sample size to $N = 20$ and generate the ordinal predictor X from a multinomial distribution with $K = 6$ categories and equal probabilities for each category. Similar results can be obtained with other settings (different values for K and N). The conditional distribution of Y given X is varied across the simulation experiments. We examine two settings. In the first setting, predictors with one cutpoint (thus inducing a monotonic) are simulated. In the second setting, predictors with two cutpoints (thus inducing a non-monotonic ordering) are simulated. The simulation designs are described in detail below.

(1) *One-cutpoint design*: A single cutpoint is fixed on the range of X. On the left of this cutpoint, the response Y is sampled from a Bernoulli distribution with low probability of success $P(Y = 1) = p_l$. On the right of the cutpoint, Y is sampled from a Bernoulli distribution with a high probability of success $P(Y = 1) = p_r$. In the different experimental conditions, the cutpoint is set either at a marginal position (between the first and second category of X) or at a central position (between the third and fourth category of X). The difference between p_l and p_r is varied in the simulation experiments. The values of p_l and p_r are set to simulate a weak effect (0.2, 0.4), a medium effect (0.2, 0.6) or a strong effect (0.2, 0.8) of X.
(2) *Two-cutpoint design*: Two cutpoints are fixed on the range of X, either both at marginal positions on the same side (between the first and second and between the second and third category of X), both at symmetric central positions (between the second and third and between the fourth and fifth category of X) or both at symmetric marginal positions (between the first and second and between the fifth and sixth category of X). The response Y is sampled from a Bernoulli distribution with low probabilities of success p_l and p_r on the left of the left cutpoint and on the right of the right cutpoint, respectively, and high probability of success p_m between the two cutpoints. The values of p_l, p_m and p_r are set to simulate a weak effect (0.2, 0.4, 0.2), a medium effect (0.2, 0.6, 0.2), a strong effect (0.2, 0.8, 0.2) or a mixed effect (0.2, 0.8, 0.6).

The percentage of simulation iterations (out of 100) in which association is detected at the standard significance level of 0.05 is displayed in Tables 3 (one-cutpoint design) and 4 (two-cutpoint design). This percentage may be seen as an indicator of the power of the considered method to discover the association patterns. It can be observed from Table 3 that our method can often successfully detect association in the two-cutpoint design. In the one-cutpoint design,

Table 4
Performance of both methods for the two-cutpoint design

Effect	One-cutpoint method				Novel method			
	Weak	Med.	Strong	Mixed	Weak	Med.	Strong	Mixed
Both marginal	6	15	32	31	7	30	62	24
Symmetric central	4	12	27	45	16	30	63	45
Symmetric marginal	7	11	38	41	9	26	58	41

Table 5
Varicella complications

Age category (in years)	1:0–1	2:1–2	3:2–3	4: > 3
No complications ($Y = 0$)	10	7	9	59
Complications ($Y = 1$)	6	19	12	48

Table 3 shows that our method sometimes detects association, but the power is of course lower than using the simpler method designed for one cutpoint by Boulesteix (2006b). Unsurprisingly, the power is higher for strong associations than for weak associations, and for central cutpoints than for marginal cutpoints. Table 4 shows that our novel method performs well to detect association in the two-cutpoint design. The power of our approach is higher than the power of the method for one cutpoint in all experimental designs, except for the case of mixed effects with both cutpoints set at marginal positions (which may be seen as intermediate between the one-cutpoint and the two-cutpoint setting). The power improvement is particularly striking (up to +150%) in the case of symmetric central cutpoints. In a word, our method may not be used to detect association patterns in general settings, since it results in a loss of power if there is only one true cutpoint. However, it is able to detect two-cutpoint patterns with high power, even for very small sample sizes, where methods assuming monotonic dependence fail.

4. The varicella study

Varicella (chickenpox) is a highly communicable disease caused by the varicella-zoster virus. Although it is commonly regarded as a mild childhood illness, serious complications can occur. The risk of serious consequences is believed to depend on the age of the patient (Banz et al., 2003). For analyzes using our new approach, we consider a data set including $N = 170$ children between 0 and 18 years who were diagnosed with varicella. $N_1 = 85$ of them had complications, whereas the remaining $N_0 = 85$ children are controls (no complications). This data set was sampled from a larger data set presented by Wagenpfeil et al. (2004) in the context of a large retrospective epidemiological study. Table 5 gives the number of cases without and with complications in each age category.

The maximal Chi-squared statistic is obtained for the cutpoints $k_1 = 1$ and $k_2 = 3$ and the corresponding p-value is $p_{raw} = 1.0 \times 10^{-2}$. This approach overestimates the association between age and risk of complications, because the p-value is not corrected for optimal choice of cutpoints. Our novel method yields the corrected p-value $p = 3.8 \times 10^{-2}$. This result suggests that varicella is more serious for children between one and three years than for younger children (who may be protected by maternal antibodies) and elder children. Approaches that assume a monotonic trend (or one cutpoint) yield larger p-values: using the approach based on maximally selected Chi-squared statistics for ordinal predictors with one cutpoint by Boulesteix (2006b) gives a p-value of $p = 0.17$, whereas the classical Chi-squared test for trend in proportions (as implemented in the R function `prop.trend.test`) yields a p-value of 0.27, thus both failing to detect association. In a word, the distribution of maximally selected Chi-squared statistics in the context of two cutpoints may be used to correct a minimally selected p-value and has higher power than monotonic approaches to detect association in the case of a non-monotonic association with two underlying cutpoints.

5. Discussion

In this article, we propose a novel combinatorial method for computing the exact distribution of the maximally selected Chi-squared statistic in the context of a non-monotonic association with a binary response. The method can be used to adjust the p-value of the Chi-squared test for multiple comparisons when two cutpoints are selected from the range of X to maximize the Chi-squared statistic. As shown in the real data example, our approach provides an efficient tool to assess the statistical significance of a pair of cutpoints. It can detect association in the case of an umbrella ordering with two underlying cutpoints, where methods that assume monotonic dependence (one cutpoint) fail. Unsurprisingly, the simulations also show that our method results in a loss of power if there is only one true cutpoint. Thus, it is not appropriate as a general test of independence.

From a practical point of view, our new method may be useful in two important situations. First, it avoids the biased reporting of low p-values, when the cutpoints are chosen optimally. Secondly, if the investigator suspects a dependence pattern between X and Y but cannot assess it with classical monotonic approaches, our method may be helpful to confirm the (non-monotonic) association between X and Y. In practice, investigators sometimes notice two candidate cutpoints based on descriptive plots and want to know if this pattern is relevant or only due to chance. This problem may be addressed with our method. Note that, if one applies successively the one-cutpoint and the two-cutpoint methods or choose the method after observation of the data, adjustment for multiple testing is recommended. As a two-cutpoint method, it could also be applied to outcomes associated with points around a circle, e.g., in astronomy. Another potential application is recursive partitioning. In the last few years, procedures based on maximally selected statistics have been successively applied to regression and classification trees, for instance in Shih (2004), Lausen et al. (2004), Strobl et al. (2007), to avoid the variable selection bias outlined by e.g., Loh and Shih (1997). Our method could become a powerful exact tool to assess complex splits in p-value adjusted trees.

References

Banz, K., Wagenpfeil, S., Neiss, A., Goertz, A., Staginnus, U., Vollmar, J., Wutzler, P., 2003. The cost-effectiveness of routine childhood varicella vaccination in Germany. Vaccine 21, 1256–1267.

Boulesteix, A.L., 2006a. Maximally selected chi-square statistics and binary splits of nominal variables. Biometrical J. 48, 838–848.

Boulesteix, A.L., 2006b. Maximally selected chi-square statistics for ordinal variables. Biometrical J. 48, 451–462.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, J.C., 1984. Classification and Regression Trees. Wadsworth, Monterey, CA.

Halpern, A.L., 2000. Multiple-changepoint testing for an alternating segments model of a binary sequence. Biometrics 56, 903–908.

Halpern, J., 1982. Maximally selected chi square statistics for small samples. Biometrics 38, 1017–1023.

Hans, C., Dunson, D.B., 2005. Bayesian inferences on umbrella orderings. Biometrics 61, 1018–1026.

Holländer, N., Sauerbrei, W., Schumacher, M., 2004. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. Statist. Medicine 23, 1701–1713.

Hothorn, T., Lausen, B., 2003. On the exact distribution of maximally selected rank statistics. Comput. Statist. Data Anal. 43, 121–137.

Koziol, J.A., 1991. On maximally selected chi-square statistics. Biometrics 47, 1557–1561.

Kuiper, N., 1960. Tests concerning random points on a circle. Proc. Konink Nederl Akad. Wetensch. A 63, 38–47.

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. Comput. Statist. Data Anal. 21, 307–326.

Lausen, B., Lerche, R., Schumacher, M., 2002. Maximally selected rank statistics for dose-response problems. Biometrical J. 44, 131–147.

Lausen, B., Hothorn, T., Bretz, F., Schumacher, M., 2004. Assessment of optimal selected prognostic factors. Biometrical J. 46, 364–374.

Loh, W., Shih, Y., 1997. Split selection methods for classification trees. Statist. Sinica 7, 815–840.

Miller, R., Siegmund, D., 1982. Maximally selected chi square statistics. Biometrics 38, 1011–1016.

Robertson, T., Wright, F., Dykstra, R., 1988. Order restricted statistical inference. Wiley, New York.

Royston, P., Altman, D.G., Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Statist. Medicine 25, 127–141.

Salanti, G., Ulm, K., 2003. Tests for trend in binary response. Biometrical J. 45, 277–291.

Shih, Y.S., 2004. A note on split selection bias in classification trees. Comput. Statist. Data Anal. 45, 457–466.

Siegmund, D., 1986. Boundary crossing probabilities and statistical applications. Ann. Statist. 14, 361–404.

Strobl, C., Boulesteix, A.L., Augustin, T., 2007. Unbiased split selection for classification trees based on the Gini Index. Comput. Statist. Data Anal., in press, doi:10.1016/j.csda.2006.12.030.

Wagenpfeil, S., Neiss, A., Banz, K., Wutzler, P., 2004. Empirical data on the varicella situation in Germany for vaccination decisions. Clinical Microbiology and Infection 10, 425–430.