LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Monika Jelizarow, Ulrich Mansmann, Jelle J. Goeman

A Cochran-Armitage-type and a score-free global test for multivariate ordinal data

Technical Report Number 168, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A Cochran-Armitage-type and a score-free global test for multivariate ordinal data

Monika Jelizarow[1]*, Ulrich Mansmann[1,2] and Jelle J. Goeman[3]

[1] Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians University Munich, Marchioninistr. 15, 81377 Munich, Germany

[2] Department of Statistics, Ludwig-Maximilians University Munich, Ludwigstr. 33, 80539 Munich, Germany

[3] Department for Health Evidence, Radboud University Medical Center, Postbus 9101, 6500 HB Nijmegen, The Netherlands

Abstract

We propose a Cochran-Armitage-type and a score-free global test that can be used to assess the presence of an association between a set of ordinally scaled covariates and an outcome variable within the range of generalized linear models. Both tests are developed within the framework of the well-established 'global test' methodology and as such are feasible in high-dimensional data situations under any correlation and enable adjustment for covariates. The Cochran-Armitage-type test, for which an intimate connection with the traditional score-based Cochran-Armitage test is shown, rests upon explicit assumptions on the distances between the covariates' ordered categories. In contrast, the score-free test parametrizes these distances and thus keeps them flexible, rendering it ideally suited for covariates measured on an ordinal scale. As confirmed by means of simulations, the Cochran-Armitage-type test focuses its power on set-outcome relationships where the distances between the covariates' categories are equal or close to those assumed, whereas the score-free test spreads its power over the full range of possible set-outcome relationships, putting more emphasis on monotonic than on non-monotonic ones. Based on the tests' power properties, it is discussed when to favour one or the other, and the practical merits of both of them are illustrated by an application in the field of rehabilitation medicine. Our proposed tests are implemented in the `R` package `globaltest`.

Keywords: Cochran-Armitage test for trend; generalized linear model; global test; logit model; multivariate ordinal data

*Corresponding author: Monika Jelizarow, e-mail: jelizarow@ibe.med.uni-muenchen.de

1. Introduction

Global hypothesis tests for possibly high-dimensional data have become an important topic in statistical research. Primarily, this has been driven by the need for methodology that allows to test predefined sets of microarray-based gene expression data for association with some clinical parameter (Draghici et al., 2003; Goeman et al., 2004; Mansmann and Meister, 2005; Kong et al., 2006; Hummel et al., 2008). The main argument put forward by researchers has been that it may sometimes be more worthwhile to draw inferential conclusions about the sets as a whole than about the individual genes, both in view of interpretability of results and power. From the statistical viewpoint, gene expression levels are metrically scaled (or, to be more precise, ratio scaled) variables. Consequently, the plethora of tests proposed in this context (see Ackermann and Strimmer (2009) for an overview) may likewise be applied to sets of metric variables stemming from other contexts. The potential benefit of global tests, however, reaches beyond problems on the metric scale.

In medical applications where categorical variables are widespread, it is particularly ordinal variables that can often be meaningfully structured into sets. Examples include questions in psychomedical diagnostic tests (e.g. structured into sets by the subdimension they describe), side effects in drug safety studies (e.g. structured into sets by the body function they affect), items in questionnaire-based studies on functional limitations and disabilities (e.g. structured into sets by means of the International Classification of Functioning, Disability and Health (ICF) (Ustün et al., 2003)) and single-nucleotide polymorphisms (SNPs) in next-generation sequencing studies (e.g. structured into sets by genes). Typically, such prior knowledge is not exploited inferentially, and one simply performs well-established univariate tests for each variable. When the objective is to assess the presence of an association with some binary variable, for example, the most widely used univariate test for ordinal variables is the two-sided Cochran-Armitage (CA) test for trend (Cochran, 1954; Armitage, 1955) which, in medical statistics, is usually better known in the one-sided formulation of Freidlin et al. (2002). As with gene expression data, however, in the case of ordinal data it may likewise be preferable to shift the unit of analysis from individual variables to whole sets of variables. It is therefore of practical interest to develop a methodology that allows to address such problems.

The literature concerned with global tests for ordinal data is sparse. For the two-sample case, Klingenberg et al. (2009) proposed a permutation test for stochastic order between the ordinal variables' marginal distributions. Recently, besides discussing Hotelling-type tests along the lines of Agresti and Klingenberg (2005) which treat ordinal data as nominal, Jelizarow et al. (2014) generalized this test from one-sided to

two-sided problems. Furthermore, they showed that, under working independence between the variables in the set to be tested, the test statistic of Klingenberg et al. (2009) is equivalent to the sum of variable-specific one-sided CA test statistics over the whole set. Their own test statistic equals the sum of variable-specific two-sided CA test statistics. The tests of Klingenberg et al. (2009) and Jelizarow et al. (2014) can thus be seen as permutation-based generalizations of the CA test to higher dimensions. This fact renders them an intuitive choice for set-based analyses of ordinal data, yet they have their limitations. Firstly, they are confined to problems where the set of interest shall be tested for association with some binary variable. This leaves many possible set relationships with non-binary variables unexplored. Secondly, they do not allow for adjustment for potential confounders. In practice where observational studies are common, however, the possibility of making such adjustments is of utmost importance, in order that false positive findings can be prevented.

The present paper develops two global tests for ordinal data which overcome the above limitations. The tests are based on different assumptions regarding the distances between the variables' ordered categories, rendering them useful in different practical situations. The first part of the paper introduces the statistical framework within which both tests are being constructed. In particular, this is the framework of the 'global test' of Goeman et al. (2004, 2006) which was originally proposed for the analysis of sets of genes. Within the broad context of generalized linear models (GLMs) (McCullagh and Nelder, 1989), the global test exploits the duality between association and prediction: if the set of interest is associated with some other variable, it will improve prediction of that variable. Adopting the terminology of prediction models, the considered null hypothesis is that none of the covariates in the set is associated with the outcome variable, and the alternative hypothesis is that at least one of the covariates in the set shows such an association. Adjustment for other covariates is feasible, provided that their number is smaller than the sample size, which is the standard case in practice.

The second part of the paper elaborates and discusses the two tests proposed. The first test is simply the original global test for metric data applied to scores that need to be assigned a priori to the covariates' categories, for example 1 to 'low pain', 2 to 'moderate pain' and 4 to 'severe pain' if one believes that the distance between 'moderate pain' and 'severe pain' is twice the distance between 'low pain' and 'moderate pain'. We shall refer to this test as 'CA-type' test, since the CA test is also based on prespecified scores. It turns out that, with data standardized to unit variance, this test is a natural generalization of the traditional two-sided CA test to higher dimensions, covariate-adjusted scenarios and all types of outcome variables that are within the range of GLMs. Immediate connections with other methods are pointed out. While the CA-type test expects the user to explicitly choose scores, and making a choice of scores implies making assumptions on the distances between the covariates' categories, the

second test which we shall refer to as 'score-free' test is unpredjudiced regarding these distances. As such, it is ideally suited for ordinal covariates because, by definition, the distances between their categories are generally unknown. The unpredjudicedness is achieved through an appropriate dummy-based coding scheme for the ordinal observations which uses only the ordering of the categories. An appealing property of this test is that the test result does not depend on any reference category in the coding scheme.

The third part of the paper examines the behaviour of the two tests by means of simulations, illustrates their application with data from rehabilitation medicine, and provides practical recommendations on when to favour one or the other. Our proposed tests are implemented in the `R` package `globaltest` which can be obtained from www.bioconductor.org.

2. The 'global test' framework

For a sample of n independent subjects, suppose that we have an $n \times 1$ outcome vector $\boldsymbol{y}$, an $n \times q$ design matrix $\boldsymbol{Z}$ which contains realizations of the covariates we would like to adjust for (e.g. typical potential confounders such as age and sex), and an $n \times p$ design matrix $\boldsymbol{X}$ which contains realizations of the covariates we would like to make inferences about. Suppose further that q is smaller than n, whereas p may exceed n. The data situation may thus be high-dimensional. Under the assumption that the covariates and the outcome variable relate to each other via the GLM, we have

$$g(E(\boldsymbol{y})) = \mathbf{1}\gamma_0 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{X}\boldsymbol{\beta}, \tag{1}$$

where $g(\cdot)$ is the canonical link function for the exponential family distribution of the components of $\boldsymbol{y}$, for example the identity function when the outcome variable is continuous (e.g. some blood parameter) or the logit function when the outcome variable is binary (e.g. some disease subtype). $\mathbf{1}$ is an $n \times 1$ vector of ones, γ_0 denotes an intercept term, $\boldsymbol{\gamma}$ is an unknown $q \times 1$ vector of regression coefficients for the covariates in $\boldsymbol{Z}$, and $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients for the covariates in $\boldsymbol{X}$. Based on the observed data, we are interested whether the set of covariates in $\boldsymbol{X}$ as a whole is associated with the outcome $\boldsymbol{y}$, after adjustment for the effect of the covariates in $\boldsymbol{Z}$. This problem can be expressed through the hypotheses

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \quad \text{against} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}. \tag{2}$$

Problem (2) is that for which Goeman et al. (2004, 2006) developed the 'global test', based on ideas of le Cessie and van Houwelingen (1995). In particular, they derived a score test statistic that can be employed whatever the dimensionality of the alternative

hypothesis is, provided that the respective null hypothesis is low-dimensional. This is in contrast to the classical score, Wald or likelihood ratio test statistic: they all break down when the number of model parameters under the alternative of interest exceeds the number of subjects in the sample. In explicit terms, the test statistic of Goeman et al. (2004, 2006) has the form

$$S = (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is the expectation of $\boldsymbol{y}$ under the null hypothesis. Because $\boldsymbol{\mu}$ is unknown, its maximum likelihood estimate $\hat{\boldsymbol{\mu}} = g^{-1}(\mathbf{1}\hat{\gamma}_0 + \boldsymbol{Z}\hat{\boldsymbol{\gamma}})$ is plugged in, with $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}$ being the null model coefficients estimated via an iteratively reweighted least squares algorithm. The resultant test statistic

$$\hat{S} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \tag{3}$$

is thus a quadratic form in the residuals of the null model. For this quadratic form, Goeman et al. (2011) analytically derived an approximate null distribution which is conditional on $\boldsymbol{X}$ and thus remains valid for any correlation between the covariates in the set considered. By means of simulations, this null distribution was shown to perform well with respect to type I error rate control even when the sample size is moderate to small. Alternatively, the test statistic's exact null distribution may be obtained via permutation, yet this procedure is computationally more demanding and, more importantly, it is only valid for problem (2) if the null covariates and the covariates in the set to be tested are independent of each other. For significance assessment, the test statistic's permutation null distribution should therefore only come into question if such an independence assumption seems plausible or, trivially, if no covariates are present under the null hypothesis. In this paper we shall use the approximate null distribution of Goeman et al. (2011) throughout.

The global test exhibits several properties (P1–P6) making it amenable to broad and efficient use in practice. As previously mentioned, it is applicable both in the case of low-dimensional and high-dimensional alternatives (P1), it allows for covariate adjustment without further assumptions (P2), it is valid even under correlation (P3), and it can be performed at low computational costs, since an analytical approximation of the test statistic's null distribution is at hand (P4). Besides that, it possesses an optimality property, which follows from the fact that it has been constructed as a score test. In particular, it has optimal average power to detect alternatives uniformly distributed on the p-dimensional ball $\|\boldsymbol{\beta}\| \leq \epsilon$, for $\epsilon \downarrow 0$. In less technical terms, among all possible tests, the global test maximizes the average power against alternatives that are in a neighbourhood of the null hypothesis (P5). On average, it is thus the best test to use if it is expected that all or most covariates in the set are only weakly associated with the outcome variable. It is important to note, however, that this optimality property is

meant in terms of the chosen parametrization of the covariates under the alternative; changing the parametrization means changing the shape of the neighbourhood of the null hypothesis where the test is optimal. Finally, the test statistic (3) can be written as $\hat{S} = \sum_{k=1}^{p} \left[\boldsymbol{x}_k^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}})\right]^2$, that is, the sum of covariate-specific test statistics over the whole set, where $\boldsymbol{x}_k$ is the kth column of $\boldsymbol{X}$ (P6). We shall see later on in Sections 3 and 5 that this property proves to be useful in various respects. Noting that, at convergence of the null model, it holds $(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, where $\boldsymbol{I}$ denotes the n-dimensional identity matrix, $\boldsymbol{H} = \tilde{\boldsymbol{Z}} \left(\tilde{\boldsymbol{Z}}^\top \boldsymbol{W} \tilde{\boldsymbol{Z}}\right)^{-1} \tilde{\boldsymbol{Z}}^\top \boldsymbol{W}$ with $\tilde{\boldsymbol{Z}} = (\boldsymbol{1}|\boldsymbol{Z})$ is the asymmetric hat matrix of the null model, and $\boldsymbol{W} = \text{diag}(\phi v(\hat{\boldsymbol{\mu}}))$ is the covariance matrix of $\boldsymbol{y}$ under the null hypothesis, with ϕ being the dispersion parameter and $v(\cdot)$ the variance function of the distribution of the components of $\boldsymbol{y}$, the kth covariate-specific test statistic can in turn be written as $\hat{S}_k = \left[\boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}})\right]^2$. From this representation we can immediately see that the contribution of each covariate to the overall test statistic is determined by its residual variance, adjusted for the null covariates. Whether this implicit weighting is appropriate or not depends on the application, such that some standardization might become necessary. We come back to this issue in Section 3.4. For further interpretations of the test statistic (3) we refer to Goeman et al. (2004, 2006), and to Goeman et al. (2004) and Solari et al. (2012) for connections with penalized likelihood and random effects methods.

Essentially, the framework of the global test is defined by (1)–(3), and all tests constructed within it enjoy the properties P1–P6. For sets of metrically scaled covariates, several such tests have already been implemented, each of which is suited for a different outcome type: a global test for the linear model (for continuous outcomes) (Goeman et al., 2004), the logit model (for binary outcomes) (Goeman et al., 2004), the multinomial logit model (for multi-class outcomes), the Poisson model (for count outcomes), and an extended global test for the Cox proportional hazards model (for survival outcomes) (Goeman et al., 2005). In Section 3 we discuss how, within the above framework, this versatile methodology can be made applicable to sets of ordinally scaled covariates.

3. Handling ordinal covariates under the alternative

3.1. Preliminaries

In what follows, suppose that the covariates in the set of interest are ordinal, and let c_k denote the number of categories of the kth covariate. For convenience of notation, let the ordered categories of unknown distance be labelled with numbers 1 to c_k. (In the data set considered in Section 5, for example, most of the covariates describe func-

tional limitations and disabilities, and the numbers 1 to 3 stand for the categories 'no impairment', 'mild to moderate impairment' and 'severe to complete impairment'.) For x_{ik}, the ith realization of the kth covariate, we thus have: $x_{ik} \in \{1, \ldots, c_k\}$.

Technically, the ordinal covariates' special character manifests itself in the fact that their realizations typically need to be recoded in order to enable proper specification of the model under the alternative. Direct use of the labels would imply the assumption that the covariates' categories are equally-spaced. Given that the numbers 1 to c_k are arbitrary and merely meant to indicate which of the categories have been observed, this may not always be desirable. Hence, if we want to render the global test methodology sensitive towards the covariates' ordinal nature, we need to recode the x_{ik}s appropriately. Two approaches to do so are presented in Sections 3.2 and 3.3, resulting in two different tests for sets of ordinal data which both enjoy the properties P1–P6 described in the previous section.

3.2. CA-type approach with prespecified scores

The first approach codes observations on an ordinal scale in the same fashion as does the CA test for trend, hence the name CA-type approach. Essentially, this means that the numbers 1 to c_k are transformed into scores that need to be assigned a priori to the ordinal covariates' categories, and the observed scores are then treated as if they were metric observations. Our motivation to consider such a score-dependent approach within the global test framework stems from the wide popularity of the CA test in statistical practice and particularly in medical applications. Formally, the transformation rule that characterizes the CA-type approach can be expressed by

$$\tilde{x}_{ik} = s_k(v) \text{ if } x_{ik} = v, \tag{4}$$

where $v = 1, \ldots, c_k$ indexes the ordered categories and $s_k(v)$ denotes the score assigned to the vth category of the kth covariate. It is easy to see that direct use of the numbers 1 to c_k is a special case of (4). The CA-type test statistic then is

$$\hat{S}_{\text{CA}} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{5}$$

where $\tilde{\boldsymbol{X}}$ is the score-transform of the design matrix $\boldsymbol{X}$ in terms of (4). Thus, the test statistic (5) is the original test statistic (3) applied to prespecified scores.

A special variant of the test statistic (5) arises when the outcome variable is binary, the null model contains only an intercept, and the columns $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_p$ of $\tilde{\boldsymbol{X}}$ are standardized to have unit variance. In particular, under these conditions, the test statistic (5) is equivalent to the sum of covariate-specific two-sided CA test statistics. The proof

is a straightforward calculation and is given in Appendix A.1. We can immediately conclude from this relationship that, with $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_p$ standardized to unit variance, the resultant CA-type test is a proper generalization of the traditional two-sided CA test in three important directions: to higher dimensions, to covariate-adjusted scenarios and to all types of outcome variables that are within the range of GLMs. As such, it can likewise be seen as a generalization of the earlier mentioned test of Jelizarow et al. (2014). The standardization of the columns of $\tilde{\boldsymbol{X}}$, and its implications, will be further discussed in Section 3.4.

For the validity of the CA-type test, the concrete choice of scores is not relevant, provided that this choice has been made without inspection of the data observed. When it comes to the test's power, however, the choice of scores is crucial. The crux is that the scores reflect the suspected relationship between the covariates in the set to be tested and the outcome variable. For example, choosing equally-spaced scores for all covariates in the set reflects the suspicion that the relationship is linear, that is, that the outcome changes linearly between two adjacent categories of at least one covariate in the set. If the suspicion is correct, the CA-type test will be powerful. If it is not correct, that is, if the choice of scores is poor, it may happen that the test has no power at all. We shall illustrate this point by means of simulations in Section 4.

In connection with the choice of scores, two issues deserve particular emphasis. Firstly, the CA-type test has the desirable property that two sets of scores $\{(s_k(1), \ldots, s_k(c_k))\}_{k=1}^{p}$ and $\{(s'_k(1), \ldots, s'_k(c_k))\}_{k=1}^{p}$ lead to the same test result if constants $t, u \in \mathbb{R}$ exist such that $s'_k(v) = t \cdot s_k(v) + u$ for all v and k. The outcome of the test is thus the same for scores that are linear transforms of each other, such as (1, 2, 4) and (3, 5, 9) or (10, 20, 40). Practically speaking, this means that the test result solely depends on the kind of the suspected relationship and not on the — to some extent subjective — numerical scale that has been chosen to reflect it. This property may come as a surprise because, obviously, the test statistic (5) is not invariant to every linear transformation of the scores used. The reason why the outcome of the test nevertheless is so lies in the way in which the test statistic needs to be rescaled before its approximate null distribution can be derived (Goeman et al., 2011). For details on the rescaling we refer to Goeman et al. (2011), and here limit ourselves to just mentioning its welcome consequences. Secondly, because the CA-type test is a two-sided test and as such does not depend on the sign of the true regression coefficients for the covariates in the set of interest, it will not be sensitive towards the direction of the suspected relationship of each covariate with the outcome variable. For illustration, for some set that only contains ordinal covariates with three categories, any of the 2^p possible mixtures of the strictly monotonically increasing scores (1, 2, 4) and the strictly monotonically decreasing scores (-1, -2, -4) will lead to the same test result. This should be kept in mind in order to prevent false inferential conclusions.

The CA-type test is useful whenever the research interest focuses on the detection of relatively specific alternatives. In such situations, the fact that the test requires specification of scores for all covariates in the set to be tested, and that making a choice of scores means making assumptions on the distances between the covariates' categories, will seldom pose considerable problems. Rather, it can be taken advantage of in order to direct the power of the test towards the desired alternative. When the research interest is broader in the sense that many different alternatives are considered equally important, however, '[...] scientists may feel that the assignment of scores is slightly unscrupulous, or at least they are uncomfortable about it. [...]' (Cochran, 1954). A test that is useful in such situations is discussed in the next section.

3.3. Score-free approach

The second approach to handling ordinality dispenses with scores altogether, hence the name score-free approach. It codes ordinal observations by using the dummy-based coding scheme of Walter et al. (1987), sometimes called split coding (Gertheiss et al., 2011). This means that the numbers 1 to c_k are transformed such that the ordinal covariates are no more represented one-dimensionally but multi-dimensionally by groups of dummies, with each group corresponding to one ordinal covariate. As opposed to classical dummies, the dummies used here contain information on the ordering of the covariates' categories. In explicit terms, the transformation rule that characterizes the score-free approach is

$$d_{ik\tilde{v}} = \begin{cases} 1 & \text{if } x_{ik} > \tilde{v} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $d_{ik\tilde{v}}$ is the ith component of $\boldsymbol{d}_{k\tilde{v}}$, which is the $\tilde{v}$th dummy vector for the kth covariate, and $\tilde{v} = 1, \ldots, \tilde{c}_k$ with $\tilde{c}_k := c_k - 1$. The score-free test statistic then is

$$\hat{S}_{\text{SF}} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{D}\boldsymbol{D}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{7}$$

where $\boldsymbol{D} = \left(\boldsymbol{D}_1 | \ldots | \boldsymbol{D}_p\right)$ is the dummy-transform of the design matrix $\boldsymbol{X}$ in terms of (6), with $\boldsymbol{D}_k = \left(\boldsymbol{d}_{k1} | \ldots | \boldsymbol{d}_{k\tilde{c}_k}\right)$ denoting the kth group of dummy vectors. Because the $\boldsymbol{D}_k$s are $n \times \tilde{c}_k$ matrices, we have $\tilde{c}_k$ (rather than one) model parameters for the kth covariate, so that the dimension of the alternative in (2) increases from $1 + q + p$ to $1 + q + \sum_{k=1}^{p} \tilde{c}_k$. We may thus encounter an alternative that is high-dimensional even when the data situation in itself is low-dimensional. As pointed out in Section 2, however, test statistics constructed within the global test framework can be used whatever the dimensionality of the alternative hypothesis is, and therefore no problems occur from that.

The $\boldsymbol{D}_k$s obtained through (6) are easy to interpret: the first dummy vector tells us whether the sample has been classified higher than into the first category, the second dummy vector tells us whether the sample has been classified higher than into the second category, and so on. The respective model parameters are similarly easy to interpret: $\beta_{k\tilde{v}}$, the $\tilde{v}$th regression coefficient for the kth covariate, describes the distance between category $\tilde{v}$ and $\tilde{v}+1$, that is, the difference between the effects of category $\tilde{v}$ and $\tilde{v}+1$. Effectively, this means that the first category is taken to be the reference category, and that the effects of the first and the second category are assumed to be more similar than the effects of the first and the third category, which in turn are assumed to be more similar than the effects of the first and the fourth category, and so on. Stated differently, it is expected that the outcome changes rather smoothly than jaggedly across the categories, which is intuitively plausible for covariates measured on an ordinal scale. It is important to emphasize, however, that no assumptions are made on the particular size of the $\beta_{k\tilde{v}}$s. The resultant score-free test is therefore ideally tailored to ordinal data: it incorporates the ordering of the covariates' categories, but at the same time it is unprejudiced regarding the distances between them. This is well reflected in the power properties of the test, as simulations in Section 4 will confirm: the range of alternatives it can detect varies from linear to umbrella-like relationships between the covariates in the set of interest and the outcome variable, with monotonic relationships being more likely to be detected than non-monotonic relationships.

It has just been said that the transformation rule (6) effectively takes the first category to be the reference category, and that it defines dummies under the assumption of 'smoothness'. Analogous transformation rules or coding schemes may be written up with any other of the categories as reference category. A general formulation is

$$d_{ik\tilde{v}}^{(r)} = \begin{cases} -1 & \text{if } x_{ik} \leq \tilde{v} \wedge \tilde{v} < r \\ 1 & \text{if } x_{ik} > \tilde{v} \wedge \tilde{v} \geq r \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $r \in \{1, \ldots, c_k\}$ is the chosen reference category. It is easy to see that (8) reduces to (6) when $r = 1$. The interpretation of the respective $\boldsymbol{D}_k^{(r)}$s is slightly more intricate than above. For example, for $c_k = 3$ and $r = 2$, the first dummy vector tells us whether the sample has been classified lower than into the second category, and the second dummy vector tells us whether the sample has been classified higher than into the second category. At first sight, this may suggest that different choices of the reference category lead to different test statistics and hence to potentially different inferential conclusions. This, however, is not the case, which is convenient because the choice of the reference category is often arbitrary. In particular, it is readily verified that the different nature of the $\boldsymbol{D}_k^{(r)}$s does not affect the interpretation of the $\tilde{v}$th regression coefficient as the distance between category $\tilde{v}$ and $\tilde{v}+1$. We thus have $\beta_{k\tilde{v}}^{(r)} = \beta_{k\tilde{v}}$ for

all r, meaning that the parametrization of the model under the alternative does not depend on the choice of the reference category. Intuitively, it is therefore clear that any score-free test statistic $\hat{S}_{\text{SF}}^{(r)}$ which is derived based on (8) must be equivalent to the test statistic (7), provided that the null model includes at least an intercept. For a formal proof of this valuable invariance property see Appendix A.2. The score-free test may thus be regarded as a test that randomly picks one category on the ordinal scale and, starting from there, parametrizes the distances between adjacent categories, thereby keeping them flexible.

3.4. Ordinal covariates on different scales

In practice, the most frequently encountered situation is that where all covariates in the set to be tested are measured on the same ordinal scale, that is, where $c_k = c$ for all k. This section briefly discusses practical solutions to potential issues that may arise in situations where the covariates are measured on different ordinal scales.

An important property of the test statistics (5) and (7) is that they can be decomposed into covariate-specific contributions. For the former, the contribution of each covariate to the overall test statistic is determined by its residual variance, adjusted for the null covariates. For the latter, the covariate-specific contribution is determined by the summed residual variances of the respective dummies, likewise adjusted for the null covariates. This becomes apparent from the fact that the test statistics can be written as $\hat{S}_{\text{CA}} = \sum_{k=1}^{p} \left[\tilde{\boldsymbol{x}}_k^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}})\right]^2$ and $\hat{S}_{\text{SF}} = \sum_{k=1}^{p} \sum_{\tilde{v}}^{\tilde{c}_k} \left[\boldsymbol{d}_{k\tilde{v}}^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}})\right]^2$, respectively, where $\boldsymbol{H}$ is the hat matrix of the null model (see the penultimate paragraph of Section 2). In general, this implicit weighting of the covariates is desirable: covariates with high residual variance usually carry more potentially important information than those with low residual variance, so they should have more influence on the test result. However, when the covariates are measured on different ordinal scales, this weighting will in some way be distorted by the fact that covariates with many categories are more likely to lead to high residual variance than covariates with few categories. Given that the metric level of measurement is more informative than the ordinal one, and that the finer the ordinal scale the closer it is to the metric scale, one may argue that it is only intuitive to give more weight to covariates with many categories than to covariates with few categories. In some instances, however, one might want to correct for the imbalance between the ordinal scales used. This can be accomplished by standardizing each of the covariates to unit variance before the CA-type or the score-free test is being performed. For the CA-type test statistic $\hat{S}_{\text{CA}}$, this means that we need to replace $\tilde{\boldsymbol{x}}_k$ by $\tilde{\boldsymbol{x}}_k' = \tilde{\boldsymbol{x}}_k / \left[n^{-1} \tilde{\boldsymbol{x}}_k^\top \left(\boldsymbol{I} - \boldsymbol{H}'\right) \tilde{\boldsymbol{x}}_k\right]^{1/2}$, where $\boldsymbol{H}' = n^{-1}\mathbf{1}\mathbf{1}^\top$. Note that this leads directly to the generalized CA test discussed in the second paragraph of Section 3.2; the generalized CA test is thus 'scale-corrected' by construction. For the score-free test statistic

$\hat{S}_{\mathrm{SF}}$, standardization of each covariate to unit variance means that we need to replace $\boldsymbol{d}_{k\tilde{v}}$ by $\boldsymbol{d}'_{k\tilde{v}} = \boldsymbol{d}_{k\tilde{v}} / \left[n^{-1}\mathrm{trace}\left(\boldsymbol{D}_k^\top \left(\boldsymbol{I} - \boldsymbol{H}'\right)\boldsymbol{D}_k\right)\right]^{1/2}$.

4. Simulations

4.1. Simulation set-up

In Section 3 we have stated that the CA-type test would be useful in situations where the research interest focuses on the detection of relatively specific alternatives, and that the score-free test in turn would be useful in situations where many different alternatives are considered equally important, that is, where the research interest is rather broad. In this section we present a small simulation study which we conducted with the objective of illustrating and further clarifying these statements. In particular, we examined the performance of the CA-type and the score-free test for different set-outcome relationships. The CA-type test was based on the equally-spaced scores 1 to c_k throughout, and both tests were used in their ordinary form with unstandardized covariates.

Throughout the study, the outcome variable was binary and 0/1-coded, the set to be tested comprised $p = 100$ independent ordinally scaled covariates with the same number $c = 3$ of categories, and there were no covariates to be adjusted for. The sample sizes considered were $n = 20, 40, 60, 80, 100$. Our major interest thus lay in high-dimensional data scenarios. Within this general set-up, we studied five different set-outcome relationships: linear, non-strictly monotonic, asymmetric umbrella, umbrella and mixed. For completeness, we further studied the null case of no relationship, even though, in principle, good type I error rate control can be expected due to the fact that our tests have been constructed within the global test framework. To obtain data sets for which the different relationships can be found, we used that, in the set-up considered, the set-outcome relationship is determined by the trend in the binomial proportions of sample units with outcome 1 (and 0, respectively) across the categories of each of the 100 covariates. With $(b^1_{k1}, b^1_{k2}, b^1_{k3}) =: \boldsymbol{b}^1_k$ (and $(b^0_{k1}, b^0_{k2}, b^0_{k3}) =: \boldsymbol{b}^0_k$, respectively) denoting the kth covariate's binomial proportions, where $b^1_{kv}, b^0_{kv} \in (0,1)$ and $b^1_{kv} + b^0_{kv} = 1$ for $v = 1, 2, 3$, the particular patterns of binomial proportions that we examined in our study were

(a) S0 (null case): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{100} = (0.5, 0.5, 0.5)$,

(b) S1 (linear): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{24} = (0.4, 0.5, 0.6)$;
$\boldsymbol{b}^1_{25} = \cdots = \boldsymbol{b}^1_{100} = (0.5, 0.5, 0.5)$,

(c) S2 (non-strictly monotonic): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{24} = (0.35, 0.55, 0.55)$;

$$\boldsymbol{b}_{25}^1 = \cdots = \boldsymbol{b}_{100}^1 = (0.5, 0.5, 0.5),$$

(d) S3 (asymmetric umbrella): $\boldsymbol{b}_1^1 = \cdots = \boldsymbol{b}_{24}^1 = (0.4, 0.6, 0.5)$;
$\boldsymbol{b}_{25}^1 = \cdots = \boldsymbol{b}_{100}^1 = (0.5, 0.5, 0.5)$,

(e) S4 (umbrella): $\boldsymbol{b}_1^1 = \cdots = \boldsymbol{b}_{24}^1 = (0.45, 0.65, 0.45)$;
$\boldsymbol{b}_{25}^1 = \cdots = \boldsymbol{b}_{100}^1 = (0.5, 0.5, 0.5)$ and

(f) S5 (mixed): $\boldsymbol{b}_1^1 = \cdots = \boldsymbol{b}_6^1 = (0.4, 0.5, 0.6)$;
$\boldsymbol{b}_7^1 = \cdots = \boldsymbol{b}_{12}^1 = (0.35, 0.55, 0.55)$;
$\boldsymbol{b}_{13}^1 = \cdots = \boldsymbol{b}_{18}^1 = (0.4, 0.6, 0.5)$;
$\boldsymbol{b}_{19}^1 = \cdots = \boldsymbol{b}_{24}^1 = (0.45, 0.65, 0.45)$;
$\boldsymbol{b}_{25}^1 = \cdots = \boldsymbol{b}_{100}^1 = (0.5, 0.5, 0.5)$.

For S1–S5, the number of informative covariates in the set was thus chosen as 24. For each desired pattern, random data sets were generated as follows. Firstly, the binary outcome was drawn from a Bernoulli distribution with probability of success equal to 0.5. Secondly, conditionally on the outcome, realizations of each of the 100 ordinal covariates were drawn from covariate-specific independent multinomial distributions such that the desired pattern of binomial proportions resulted; multinomial distributions that satisfy this condition were determined based on Bayes' theorem. The power (type I error rate) was then estimated from 10000 random data sets as the average rejection rate of false (true) null hypotheses, and the desired significance level was $\alpha = 0.05$. The simulation margin of error thus amounted to approximately $\pm$ 0.44 %. The results from our simulation experiments are reported in Section 4.2.

4.2. Simulation results

Table 2 summarizes the average rejection rates obtained with our two tests for the simulation scenarios S0–S5. Under the null hypothesis of no association between the set and the outcome variable (scenario S0), both tests offer good type I error rate control: nearly all deviations of the actual type I error rate from the nominal one lie within the simulation margin of error, which confirms the general usability of the approximate null distribution of Goeman et al. (2011). Under the different alternative hypotheses of interest (scenarios S1–S5), we find for the CA-type test that its power increases the better the prespecified scores reflect the true set-outcome relationship. This is intuitively plausible and typical for score-dependent methods for ordinal data, such as for the traditional univariate CA test. The dependence of the CA-type test's power properties on the choice of scores becomes particularly evident when we contrast the results for S1 and S4 with each other. To recap, we have chosen the scores (1, 2, 3) throughout the set, which reflects linearity of the suspected relationship between the covariates

in the set and the outcome variable. This is exactly the kind of set-outcome relationship that is true for S1. As Table 2 shows, this accurate match between the prespecified scores and the true set-outcome relationship renders the CA-type test powerful, even slightly more powerful than the score-free test. For S4, in contrast, the CA-type test has basically no power at all. Apparently, this is owing to the fact that here the degree of misspecification of scores is fairly large, since S4 represents an umbrella-like set-outcome relationship. (A side remark: to have power to detect umbrella-like set-outcome relationships, we would have had to choose umbrella-shaped scores such as, for example, (1, 2, 1).) As can be further seen from Table 2, an entirely different picture than for the CA-type test is obtained for the score-free test. In particular, our results indicate that the latter has power irrespective of what kind of relationship the covariates in the set exhibit with the outcome variable, and that the power to detect monotonic set-outcome relationships (scenarios S1 and S2) exceeds the power to detect non-monotonic ones (scenarios S3 and S4). This specific behaviour of the score-free test has been confirmed by various further simulation experiments that we conducted on this issue (not shown here).

The simplistic character of the scenarios S1–S4 has helped to illustrate the power properties of the CA-type and the score-free test. Scenarios of this kind are, however, unlikely to be encountered in practice. A more realistic scenario is represented by S5 where some of the covariates in the set are monotonically related to the outcome variable, whereas others show a non-monotonic relationship. As could have been expected from the results for S1–S4, here the score-free test has more power than the CA-type test. Nevertheless, the score-free test will not *per se* be the better choice. In particular, the fact that the CA-type test requires correctly specified scores to be powerful makes it useful in applications where only a specific type of set-outcome relationship is considered important.

Table 2: Average rejection rates for the simulation scenarios S0–S5 (see Section 4.1 for detailed descriptions).

	CA-type test*					Score-free test				
Sample size n	20	40	60	80	100	20	40	60	80	100
S0	0.053	0.051	0.056	0.054	0.049	0.055	0.054	0.056	0.052	0.052
S1	0.194	0.435	0.689	0.861	0.948	0.185	0.415	0.662	0.838	0.937
S2	0.180	0.388	0.632	0.812	0.916	0.185	0.402	0.651	0.828	0.925
S3	0.087	0.147	0.224	0.311	0.393	0.136	0.280	0.476	0.648	0.792
S4	0.056	0.060	0.064	0.073	0.075	0.093	0.143	0.217	0.303	0.408
S5	0.119	0.230	0.380	0.520	0.663	0.147	0.302	0.506	0.683	0.827

* based on the equally-spaced scores (1, 2, 3) throughout the set

5. Application to data on functional limitations and disabilities in multiple sclerosis

5.1. Data set and question of interest

To illustrate the application of the tests presented in Section 3, we analyzed data from the multi-centre cross-sectional study on functional limitations and disabilities in multiple sclerosis (MS) of Holper et al. (2010). The study was conducted in overall four rehabilitation centres in Germany and Switzerland from 2007 to 2008, and it was based on the International Classification of Functioning, Disability and Health (ICF) (Ustün et al., 2003). In brief, the ICF is an extensive classification framework which allows for the description of individual, social and environmental aspects of functioning and disability both across health conditions and for specific health conditions such as MS. The description is realized by means of relevant selections from an overall pool of more than 1400 health-related ordinally scaled items called ICF categories, henceforth referred to as ICF covariates. The latter can be structured into four non-overlapping sets, the so-called ICF components: 'body functions' (b), 'body structures' (s), 'activities and participation' (d) and 'environmental factors' (e).

The considered data set includes $n = 93$ individuals of which 33 were diagnosed with the MS form primary progressive MS (PP MS) and 60 with the MS form secondary progressive MS (SP MS). Aside from disease-related and socio-demographic details on the individuals, the data set provides information on each individual's functioning and disability status captured by means of $p = 129$ ICF covariates that are considered particularly relevant for MS patients. Of this total, 34 ICF covariates belong to the ICF component b (e.g. 'orientation functions' (b114), 'sensation of pain' (b280) and 'sexual functions' (b640)), 13 to the ICF component s (e.g. 'spinal cord and related structures' (s120), 'structure of reproductive system' (s630) and 'structure of lower extremity' (s750)), 51 to the ICF component d (e.g. 'writing' (d170), 'washing oneself' (d510) and 'doing housework' (d640)) and 31 to the ICF component e (e.g. 'climate' (e225), 'immediate family' (e310) and 'transportation services, systems and policies' (e540)). For the complete list of the ICF covariates involved see Holper et al. (2010). As has been recommended by Bostan et al. (2012) for the five-level ordinal scale originally used in the ICF components b, s and d, we coarsened both the five-level and the nine-level scale originally used in the ICF component e to three levels: for the ICF covariates in b, s and d, the numbers 1, 2 and 3 label the categories 'no impairment', 'mild to moderate impairment' and 'severe to complete impairment', whereas for the ICF covariates in e they label the categories 'facilitator', 'neither barrier nor facilitator' and 'barrier'.

Based on the above data, we tested for association between each of the ICF compo-

nents and the MS form (coded with 0 for PP MS and 1 for SP MS). Statistical analyses of ICF-based data that exploit the prior knowledge on the structure of the data have previously been advocated by Jelizarow et al. (2014). Merely for the purpose of illustration of differences between the CA-type and the score-free test, we applied them both, noting that in practice one should decide for one or the other, which is sensibly done based upon power considerations. We used the CA-type test with the equally-spaced scores (1, 2, 3) throughout, and we adjusted our analysis for age, sex and sum score from the Beck Depression Inventory (BDI) II (Beck et al., 1996). Because we were interested in testing the four ICF components simultaneously, it was necessary to adjust the respective p-values for multiplicity. We did so by means of the Bonferroni-Holm procedure (Holm, 1979). Alternatively, one could think of multiplicity adjustment procedures that respect the fact that the ICF components are of different size, yet here we prefer to treat the ICF components on the same footing for the sake of simplicity. The results obtained are discussed below in Section 5.2.

5.2. Test results

Table 3 displays the Bonferroni-Holm adjusted p-values for the ICF components b, s, d and e, obtained with the CA-type and the score-free test for the logit model. At the standard level of significance $\alpha = 0.05$, the CA-type and the score-free test lead to the same inferential conclusions for b, s and d: while b and d are found to be significantly associated with the MS form, no such association can be revealed for s. It can thus be said that PP MS and SP MS patients differ in their overall pattern of restrictions of body functions as well as activities and participation. When it comes to the ICF component e, the CA-type test clearly maintains the null hypothesis of no association with the MS form, whereas the score-free test rejects it. Recalling the simulation results on power from Section 4.2, this may indicate that the ICF component e comprises ICF covariates that exhibit a non-monotonic relationship with the MS form. Figure 1 helps to clarify whether this is the case: it shows the ICF covariate-specific contributions to the test statistics $\hat{S}_{\mathrm{CA}}$ (left panel) and $\hat{S}_{\mathrm{SF}}$ (right panel) for the entire set e. If now an ICF covariate is non-monotonically related to the MS form, its influence on $\hat{S}_{\mathrm{CA}}$ is likely to be smaller compared to its influence on $\hat{S}_{\mathrm{SF}}$. Among the 31 ICF covariates included in e, it becomes readily visible from the figure that this is particularly true for the ICF covariate 'light' (e240). A look into the data in fact suggests the presence of a non-monotonic relationship: the binomial proportions across the categories of the ICF covariate e240 are 0.21, 0.54 and 0.24 for PP MS patients and, consequently, 0.79, 0.46 and 0.76 for SP MS patients. The fact that this is fairly close to an umbrella-like relationship explains why the influence of e240 on $\hat{S}_{\mathrm{CA}}$ is considerably less pronounced than on $\hat{S}_{\mathrm{SF}}$. Noting that non-monotonic relationships seem to be present for 17 further ICF covariates in e

(see hatched bars in Figure 1), and that numerous of these ICF covariates belong to the most influential ones in the set, it is of little surprise that here our two tests have lead to different inferential conclusions.

Differences between PP MS and SP MS patients with respect to functional limitations and disabilities in the course of the disease have previously been reported in the medical literature (A. Thompson, 2004; Amato et al., 2006). On the basis of individual ICF covariates, however, the presence of such differences could so far not be confirmed; merely some descriptive observations in that direction were made (Holper et al., 2010). In contrast to that, our results show that, on the basis of ICF components, proper statistical evidence in favour of the phenomenology communicated in the medical literature can be provided. This well exemplifies the potential practical benefit of the tests developed in this paper.

As an additional but rather informal step, we performed the CA-type and the score-free test separately for each ICF covariate, even though the classical univariate scenario is not that by which the tests' development has been motivated. For comparison, we performed ICF covariate-specific likelihood ratio tests, based on both the CA-type and the score-free approach to handling ordinality. As with the analysis of the ICF components, we adjusted for age, sex and BDI score. After Bonferroni-Holm correction of the covariate-specific p-values, we find that for none of the 129 ICF covariates a statistically significant effect can be detected, irrespective of which of the four tests is being used. Our univariate results are thus in line with the earlier mentioned univariate results of Holper et al. (2010).

Table 3: Multiplicity-adjusted p-values via Bonferroni-Holm for the ICF components b, s, d and e.

	CA-type test	Score-free test
Body functions (b)	0.030	0.021
Body structures (s)	0.345	0.328
Activities and participation (d)	0.049	0.023
Environmental factors (e)	0.145	0.039

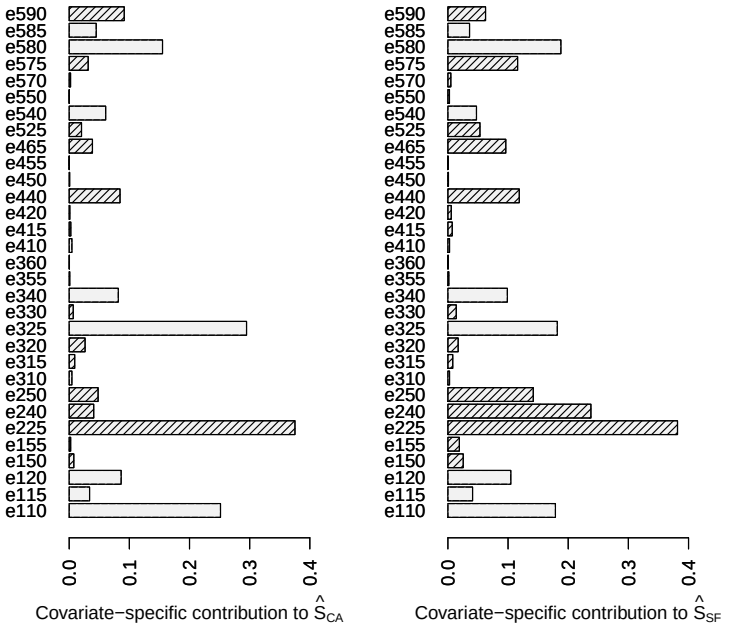

Figure 1: Covariate-specific contributions to the test statistics $\hat{S}_{\text{CA}}$ and $\hat{S}_{\text{SF}}$ for the ICF component e. Hatched bars belong to those ICF covariates for which the data suggest a non-monotonic relationship with the MS form.

6. Discussion and conclusion

Motivated by the wide occurence of structured ordinal data in medical applications, we have developed two tests that enable researchers to assess the presence of an association between a set of ordinal covariates and an outcome variable within the range of GLMs. Feasibility independent of the dimensionality of the alternative hypothesis, validity under any correlation and the possibility of covariate-adjustment render the tests widely useful in practice. Our first test, the score-based CA-type test, expects the user to make assumptions on the distances between the covariates' categories, and its power is then directed towards the set-outcome relationship that is in line with these assumptions. Under mild conditions, we have shown that this test is a proper generalization of the traditional CA test to higher dimensions, covariate-adjusted scenarios and GLM-specific outcomes. Our second test, the score-free test, respects the ordering of the covariates' categories while dispensing with assumptions on the distances between them, and its power is spread over the full range of set-outcome relationships, with more emphasis put on monotonic than on non-monotonic ones. In practice, whether to employ the CA-type or the score-free test depends on whether some

specific alternative or many different alternatives are considered important, such that recommendations can only be made with reference to concrete applications.

One scenario where the score-free test promises to be more appropriate than the CA-type test is when sets of SNPs in genetic association studies of complex diseases are to be tested. To test individual SNPs in case-control situations, it is common practice to use the traditional CA test, where the scores are chosen such that they reflect the underlying genetic model (Freidlin et al., 2002; Balding, 2006). To test sets of SNPs, the SNP-specific CA test statistics are often combined into one test statistic for the entire set, and critical values are obtained via some resampling procedure (Balding, 2006; Hoh and Ott, 2003). This popularity of the CA test for the analysis of SNP data speaks for the usefulness of the CA-type and, as a special variant, the generalized CA test in this context. For complex diseases, however, the genetic model is typically unknown, and the choice of scores hence unclear. To overcome this issue, one can perform separate tests for each genetic model and then build some weighted average of the respective results. As pointed out by Balding (2006), it will mostly be sensible to choose the weights such that greater plausibility of the additive model is reflected but that the resultant test still has power to detect effects that are far from additive. The fact that this corresponds to the power properties of the score-free test without that weights need to be specified argues for future explorations of this test in the context of SNP set analyses.

Although standard univariate problems are not those for which our tests have originally been intended for, it is important to emphasize that the latter may be valuable in such situations as well. It should be kept in mind, however, that the tests proposed are score tests and as such only have optimal average power when the deviation from the null hypothesis is small, that is, when the effect of the covariate considered is weak.

Given that, in medical applications and beyond, sets of ordinal covariates are more frequently encountered than sets of nominal covariates, we have focused on the former in this paper. Besides the CA-type and the score-free test, however, the `R` package `globaltest` (which can be obtained from www.bioconductor.org) likewise implements a global test that is tailored to covariates measured on a nominal scale. Application of this test to sets of ordinal covariates can be sensible, yet only in instances where monotonic and non-monotonic set-outcome relationships are considered equally important.

Finally, the tests proposed are not only useful by themselves but, in addition, can be fruitfully combined with multiplicity adjustment procedures for, for example, hypotheses that can be structured in a tree by some expert knowledge (Meinshausen, 2008; Goeman and Solari, 2010; Goeman and Finos, 2012). As the inferential exploitment of such and even more comprehensive prior information becomes more and more popular, future research problems will call for extensions of the CA-type and the

score-free test for more complex models, such as for the cumulative logit model for ordinally scaled outcomes.

Acknowledgements

MJ is grateful to the German National Academic Foundation for a doctoral scholarship by which this work was supported. We thank Alarcos Cieza for helpful discussions on the ICF, and for providing us with the ICF-based data from Holper et al. (2010).

A. Proofs for Sections 3.2 and 3.3

A.1. Relationship between the CA-type test statistic and the CA test statistic

Consider the test statistic (5). Because the latter can be written as the sum of covariate-specific test statistics over the whole set, we can examine the univariate case without loss of generality. Let the outcome variable be binary and 0/1-coded. With $\tilde{x}_{ik}$ and y_i the ith component of $\tilde{\boldsymbol{x}}_k$ and $\boldsymbol{y}$, respectively, be $\sum_{i=1}^{n} \delta_{y_i 1} =: n_2$ the number of subjects with outcome 1 ('cases'), $n - n_2 =: n_1$ the number of subjects with outcome 0 ('controls'), $\sum_{i=1}^{n} \delta_{\tilde{x}_{ik} s_k(v)} y_i =: n_{2kv}$ the number of cases with $\tilde{x}_{ik} = s_k(v)$ and $\sum_{i=1}^{n} \delta_{\tilde{x}_{ik} s_k(v)} - n_{2kv} =: n_{1kv}$ the number of controls with $\tilde{x}_{ik} = s_k(v)$. For the logit model without null covariates and with the columns of $\tilde{\boldsymbol{X}}$ standardized to have unit variance, the test statistic (5) can be written as the weighted sum $\hat{S}' = \sum_{k=1}^{p} w_k \left[\tilde{\boldsymbol{x}}_k^\top (\boldsymbol{y} - \mathbf{1}\frac{n_2}{n})\right]^2$, where $w_k = \left\{\frac{1}{n}\left[\tilde{\boldsymbol{x}}_k^\top \tilde{\boldsymbol{x}}_k - \frac{1}{n}\left(\tilde{\boldsymbol{x}}_k^\top \mathbf{1}\right)^2\right]\right\}^{-1}$. We can write the kth covariate-specific test statistic $\hat{S}'_k = \left[\tilde{\boldsymbol{x}}_k^\top (\boldsymbol{y} - \mathbf{1}\frac{n_2}{n})\right]^2$ as

$$
\begin{aligned}
\hat{S}'_k &= \left[\sum_{i=1}^{n} \tilde{x}_{ik}\left(y_i - \frac{n_2}{n}\right)\right]^2 \\
&= \left[\sum_{i=1}^{n}\sum_{v=1}^{c_k} \delta_{\tilde{x}_{ik} s_k(v)} s_k(v)\left(y_i - \frac{n_2}{n}\right)\right]^2 \\
&= \left[\sum_{v=1}^{c_k} s_k(v)\left(\sum_{i=1}^{n} \delta_{\tilde{x}_{ik} s_k(v)} y_i - \sum_{i=1}^{n} \delta_{\tilde{x}_{ik} s_k(v)} \frac{n_2}{n}\right)\right]^2 \\
&= \left[\sum_{v=1}^{c_k} s_k(v)\left(n_{2kv} - \frac{n_2}{n} n_{1kv} - \frac{n_2}{n} n_{2kv}\right)\right]^2 \\
&= \left[\sum_{v=1}^{c_k} s_k(v)\left(\frac{n_1 n_{2kv}}{n} - \frac{n_2 n_{1kv}}{n}\right)\right]^2 .
\end{aligned}
$$

Likewise, we can write the kth covariate-specific 'weight' w_k as

$$\begin{aligned} w_k &= \left\{ \frac{1}{n} \left[\sum_{i=1}^{n} \tilde{x}_{ik}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \tilde{x}_{ik} \right)^2 \right] \right\}^{-1} \\ &= \left\{ \frac{1}{n^2} \left[n \sum_{i=1}^{n} \sum_{v=1}^{c_k} \delta_{\tilde{x}_{ik} s_k(v)} s_k^2(v) - \left(\sum_{i=1}^{n} \sum_{v=1}^{c_k} \delta_{\tilde{x}_{ik} s_k(v)} s_k(v) \right)^2 \right] \right\}^{-1} \\ &= \left\{ \frac{1}{n^2} \left[n \sum_{v=1}^{c_k} s_k^2(v)(n_{1kv} + n_{2kv}) - \left(\sum_{v=1}^{c_k} s_k(v)(n_{1kv} + n_{2kv}) \right)^2 \right] \right\}^{-1}. \end{aligned}$$

It is now easy to see that, up to a constant factor, $w_k \hat{S}'_k$ is equivalent to the square of the one-sided CA test statistic (see for example Freidlin et al. (2002) for this most frequently used formulation of the latter), which in turn is the two-sided CA test statistic. Thus, $\hat{S}'$ is equivalent to the sum of traditional two-sided covariate-specific CA test statistics. (The constant factor corresponds exactly to that by means of which $\hat{S}'$ is rescaled in order to be able to compute its approximate null distribution (Goeman et al., 2011).)

A.2. Invariance of the score-free test statistic to the choice of the reference category

Consider the transformation rule (8), and let $\hat{S}_{\mathrm{SF}}^{(r)} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{D}^{(r)} \boldsymbol{D}^{(r)\top} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ be the respective score-free test statistic, for some reference category $r \in \{1, \ldots, c_k\}$. To prove that $\hat{S}_{\mathrm{SF}}^{(r)}$ is invariant to the choice of the reference category, we must first rewrite it. Let $\boldsymbol{I}$ and $\boldsymbol{H}$ be defined as in the penultimate paragraph of Section 2, and let $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{Z}} \left(\tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}} \right)^{-1} \tilde{\boldsymbol{Z}}^\top$ denote the projection matrix that $\boldsymbol{H}$ becomes in the case of the linear model with normally distributed errors. Using that $(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, and noting that $\boldsymbol{H}\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{H}}$ and therefore $(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H}) \left(\boldsymbol{I} - \tilde{\boldsymbol{H}} \right)$, we can write $\hat{S}_{\mathrm{SF}}^{(r)}$ in the more cumberstone form

$$\hat{S}_{\mathrm{SF}}^{(r)} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \tilde{\boldsymbol{H}}) \boldsymbol{D}^{(r)} \boldsymbol{D}^{(r)\top} (\boldsymbol{I} - \tilde{\boldsymbol{H}})(\boldsymbol{I} - \boldsymbol{H})^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}).$$

Let $\boldsymbol{d}_{k\tilde{v}}^{(r)}$ be the $\tilde{v}$th dummy vector for the kth covariate. We notice that all that happens when we go from $\boldsymbol{d}_{k\tilde{v}}^{(r)}$ to $\boldsymbol{d}_{k\tilde{v}}^{(r+1)}$ is that the entries of the rth dummy vector are subtracted by 1. In equations, this means

$$\boldsymbol{d}_{k\tilde{v}}^{(r+1)} = \boldsymbol{d}_{k\tilde{v}}^{(r)} - \mathbf{1}\delta_{r\tilde{v}},$$

where $\delta_{r\tilde{v}} = 1$ if $r = \tilde{v}$ and $\delta_{r\tilde{v}} = 0$ otherwise. Because the vector of ones is in the null space of the projection defined by $(\boldsymbol{I} - \tilde{\boldsymbol{H}})$, it follows immediately that $(\boldsymbol{I} - \tilde{\boldsymbol{H}})\boldsymbol{D}^{(r)}$ is invariant to the choice of the reference category, provided that the null model is non-empty. Consequently, any choice of the reference category will lead to the same test statistic, which completes the proof.

References

A. Thompson (2004). Overview of primary progressive multiple sclerosis (PPMS): similarities and differences from other forms of MS, diagnostic criteria, pros and cons of progressive diagnosis. *Multiple Sclerosis 10*, S2–S7.

Ackermann, M. and K. Strimmer (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics 10*, 47.

Agresti, A. and B. Klingenberg (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C 54*, 691–706.

Amato, M. P., V. Zipoli, and E. Portaccio (2006). Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies. *Journal of Neurological Sciences 245*, 41–46.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics 11*, 375–386.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics 7*, 781–791.

Beck, A. T., R. A. Steer, and G. K. Brown (1996). *Manual for the Beck Depression Inventory-II*. San Antonio: The Psychological Cooperation.

Bostan, C., C. Oberhauser, and A. Cieza (2012). Investigating the dimension functioning from a condition-specific perspective and the qualifier scale of the International Classification of Functioning, Disability and Health based on Rasch analyses. *American Journal of Physical Medicine and Rehabilitation 91(suppl)*, S129–S140.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics 10*, 417–451.

Draghici, S., P. Khatri, R. P. Martins, G. C. Ostermeier, and S. Krawetz (2003). Global functional profiling of gene expression. *Genomics 81*, 98–104.

Freidlin, B., G. Zheng, Z. Li, and J. L. Gastwirth (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity 53(3)*, 146–152.

Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society: Series C 60*, 377–395.

Goeman, J. J. and L. Finos (2012). The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology 11(1)*, 1–18.

Goeman, J. J., J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21(9)*, 1950–1957.

Goeman, J. J. and A. Solari (2010). The sequential rejection principle of familywise error control. *Annals of Statistics 38(6)*, 3782–3810.

Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20(1)*, 93–99.

Goeman, J. J., S. A. van de Geer, and H. C. van Houwelingen (2006). Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society: Series B 68*, 477–493.

Goeman, J. J., H. C. van Houwelingen, and L. Finos (2011). Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika 98*, 381–390.

Hoh, J. and J. Ott (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics 4*, 701–709.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Holper, L., M. Coenen, A. Weise, G. Stucki, A. Cieza, and J. Kesselring (2010). Characterization of functioning in multiple sclerosis using the ICF. *Journal of Neurology 257(1)*, 103–113.

Hummel, M., R. Meister, and U. Mansmann (2008). GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics 24(1)*, 78–85.

Jelizarow, M., A. Cieza, and U. Mansmann (2014). Global permutation tests for multivariate ordinal data: alternatives, test statistics and the null dilemma. *Journal of the Royal Statistical Society: Series C (Applied Statistics). doi: 10.1111/rssc.12070*.

Klingenberg, B., A. Solari, L. Salmaso, and F. Pesarin (2009). Testing Marginal Homogeneity Against Stochastic Order in Multivariate Ordinal Data. *Biometrics 65*, 452–462.

Kong, S. W., W. T. Pu, and P. Park (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics 22*, 2373–2380.

le Cessie, S. and H. C. van Houwelingen (1995). Testing the fit of regression models via score tests in random effects models. *Biometrics 51*, 600–614.

Mansmann, U. and R. Meister (2005). Testing differential gene expression in functional groups. *Methods of Information in Medicine 44*, 449–453.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman & Hall.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika 95*, 265–278.

Solari, A., S. le Cessie, and J. J. Goeman (2012). Testing goodness of fit in regression: a general approach for specified alternatives. *Statistics in Medicine 31*, 3656–3666.

Ustün, T. B., S. Chatterji, J. Bickenbach, N. Kostanjsek, and M. Schneider (2003). The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. *Disability and Rehabilitation 25(11-12)*, 565–571.

Walter, S. D., A. R. Feinstein, and C. K. Wells (1987). Coding ordinal independent variables in multiple regression analysis. *American Journal of Epidemiology 125*, 319–323.