



HAL
open science

Nonparametric tests for repeated observations with ordered categorical data

Huiting Huang, Fortunato Pesarin

► **To cite this version:**

Huiting Huang, Fortunato Pesarin. Nonparametric tests for repeated observations with ordered categorical data. *Annales de l'ISUP*, 2019, 63 (2-3), pp.155-166. hal-03603890

HAL Id: hal-03603890

<https://hal.science/hal-03603890v1>

Submitted on 10 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pub. Inst. Stat. Univ. Paris

63, fasc. 2-3, 2019, 155-166

Numéro spécial en l'honneur des 80 ans de Denis Bosq /

Special issue in honour of Denis Bosq's 80th birthday

Nonparametric tests for repeated observations with ordered categorical data

Huiling Huang, Fortunato Pesarin

Department of Statistical Sciences, University of Padova-Italy

Abstract: This article deals with nonparametric permutation testing methods for repeated observations with multivariate ordered categorical data. Our specific interest is on testing for stochastic dominance, i.e. for a set of restricted alternatives. Several solutions to the univariate stochastic dominance case based on restricted maximum likelihood ratio tests have been proposed in the literature. These solutions require quite demanding and stringent assumptions and so are generally criticized: their asymptotic null distributions are mixtures of central chi-squared variables with weights depending on the underlying population distribution F and so the related accuracy is at least difficult or even impossible to assess. Further, testing for stochastic dominance in multivariate settings by the likelihood approach is known to be a much more difficult problem. By working within the conditioning on a set of sufficient statistics under the null hypothesis and the nonparametric combination (NPC) of dependent permutation tests it is possible to find exact solutions to some of the related problems. Solutions for multivariate two-sample designs guided by a medical application problem are provided.

1. Introduction

Problems of testing for ordered categorical variables are frequently met in many application disciplines, e.g.: social sciences, psychology, genetics, genomics, clinical trials, marketing, the environment, industrial quality control, and so forth, where at most a countable number of $V \geq 1$ of such variables are observed on each individual unit.

Testing of hypotheses for ordered categorical variables is known to be quite a difficult problem when the interest is for stochastic dominance, that is for a set of restricted alternatives. Stochastic dominance problems are of specific interest in application contexts since are frequently encountered in practice and present peculiar difficulties, especially within the framework of likelihood ratio tests (e.g. [3], [5–9], [11–13], [16], [18], [22], [23], [33], [34], [36–38]). Several solutions to the univariate case have been proposed within the restricted maximum likelihood ratio test. These solutions are generally criticized, because they: i) require quite stringent assumptions; ii) are known asymptotically for relatively small number C of categories; iii) the asymptotic null distributions are mixtures of central chi-squared variables where

AMS 2000 subject classifications: Primary 62G10, 62H17, 62M02; secondary 62G09

Keywords and phrases: conditional inference, multivariate permutation testing, nonparametric combination, restricted alternatives, stochastic dominance

the weights essentially depend on the underlying unknown population distribution F and so the related degree of accuracy and the rate of convergence are at least difficult to characterize and to assess. Thus, when available their utilization is in most cases doubtful both in practice and in theory.

Moreover, also due to the extreme difficulty to express the likelihood function [17], the multivariate case is considered as almost impossible to be analyzed within the likelihood approach, especially when V is larger than 2 or when sample size is smaller than the number of population parameters (e.g. [10], [12], [13], [19], [34], [36]). By working within the conditionality principle of inference and the nonparametric combination (NPC) of dependent permutation tests (e.g. [1], [2], [10], [14], [15], [24–31]), there are exact solutions to most important of related problems.

The NPC approach works as a general methodology for many multivariate settings, e.g. as with cases when sample sizes are smaller (even much smaller) than V , when there are non-ignorable missing and/or censored data, when some of the variables are quantitative and others categorical, when the null hypothesis is expressed as an interval of equivalent points, when the data come from selection-bias experiments, and so forth. In particular, it results especially useful when testing for a set of restricted alternatives with experimental data, i.e. for stochastic dominance, in which context it shows specific efficacy providing for exact admissible solutions that also enjoy nice asymptotic properties. NPC methods require that the given testing problems are equivalently broken-down into a set of simpler sub-problems (Roy [32]) to each of which a permutation test is available, provided that these partial tests are jointly analyzed. Thus, the NPC works within the permutation theory. Permutation tests are conditional inferential methods where conditioning is on a set of sufficient statistics under the null hypothesis for the underlying distribution F . For most problems of practical interest, especially when F lies outside the regular exponential family of distributions, the minimal sufficient statistics is just the actual pooled data set \mathbf{X} . Within the NPC methodology it is particularly noticeable that researchers are not required to model and estimate the dependence parameters on variables and on these partial tests because, due to sufficiency of \mathbf{X} , the related methods become invariable on all underlying parameters. In addition, the related solutions are often easy to obtain even for complex problems like those arising from dominance alternatives with ordered categorical variables.

The exact determination of permutation null distributions is obtained by the complete examination of all permutations of the data \mathbf{X} , $\Pi(\mathbf{X})$ say. Of course, this way becomes impractical when sample sizes are not very small and for complex problems (with the data of section 2, $\Pi(\mathbf{X})$ contains some 2.447×10^{11} V -dimensional points). In practice, however, they can be estimated, to the desired degree of accuracy, by a conditional Monte Carlo method consisting of a random sample of R elements from $\Pi(\mathbf{X})$. This solution is especially recommended for NPC methods (e.g. [25], [27], [28], [30]). Main NPC routines are implemented in R, Python, SAS[®], StatXact[®], MATLAB[®], etc.

Section 2 introduces a motivating medical example; section 3 examines the ba-

sic methodology for restricted alternatives in two-sample designs; solutions to the example are in section 4; some concluding remarks are in section 5.

2. A motivating example

This motivating medical example, from [4] and [20], concerns a clinical trial on shoulder tip pain scores observed on every patient at $V = 6$ time points after a laparoscopic surgery. The scores are classified into $C = 5$ ordered categories ranging from $c_1 \equiv 1$ (low) to $c_5 \equiv 5$ (high). Two treatments, a drug Y and a placebo N , were randomly assigned to $n = 41$ eligible patients, $n_1 = 22$ of which received Y and $n_2 = 19$ received N . Since pain distributions are expected to be naturally time-decreasing, i.e. $X_t \stackrel{d}{\geq} X_q$, $1 \leq t < q \leq V$, the pooled data set

$$\mathbf{X} = \{X_{jti}, i = 1, \dots, n_j, t = 1, \dots, V, j = 1, 2\}$$

appears as a list of trajectories of repeated observations that behave as a two-sample of *independent discrete non-stationary stochastic processes* with *unknown time-converging* (maybe monotonically) *transition matrices*. According to this interpretation, the testing process has to be necessarily set up *without assuming time-invariant distributions*.

In addition to the analyses discussed in [4], [20], [21] and [35] it is required to establish if patients taking Y exhibit, across time, a stochastically lower level of pain than those taking N . That is, while considering the multivariate distributional equality $\mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2$ as the basic assumption, onwards called the *working null*, to test for the *joint stochastic dominance* hypotheses of marginal distributions:

$$H_0 : \bigcap_{t=1}^V (X_{1t} \stackrel{d}{=} X_{2t}) \equiv \bigcap_{t=1}^V \bigcap_{h=1}^{C-1} [F_{1t}(c_h) = F_{2t}(c_h)],$$

against the alternative:

$$H_1 : \bigcup_{t=1}^V (X_{1t} \stackrel{d}{<} X_{2t}) \equiv \bigcup_{t=1}^V \bigcup_{h=1}^{C-1} [F_{1t}(c_h) > F_{2t}(c_h)],$$

where $F_{jt}(c_h) = \Pr\{X_{jt} \leq c_h\}$, $j = 1, 2$, $h = 1, \dots, C - 1$ are partial (marginal) CDFs, and where, of course, $F_{jt}(C) = 1$, $t = 1, \dots, V$, $j = 1, 2$. It is worth noting that the problem has been broken-down, according to Roy [32], into $(C - 1) \times V = 24$ simpler partial sub-problems. Also noticeable is that the related testing problem for a set of restricted alternatives requires a number of *multi-one-sided* partial test statistics. Note that H_0 is the tested null and that the working null $\mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2$ has the role to justify the exchangeability property.

As far as we know, in no case this problem can be analyzed within likelihood-based methods unless extremely stringent conditions, generally too difficult or even almost

impossible to justify, were introduced. This is a direct consequence of two main aspects: i) the full set of involved parameters, corresponding to the number of cells in a multivariate contingency table representation, has $C^V - 1 = 15624$ dimensions (that number becomes 640584 if parameters are subject-specific, i.e. random; Joe [17]); ii) the set of alternatives is restricted to lie in the $V \cdot (C - 1) = 24$ -dimensional positive orthant. So, the likelihood way cannot be validly considered.

With clear meaning of the symbols, the problem: $H_{0th} : F_{2t}(c_h) = F_{1t}(c_h)$ against $H_{1th} : F_{2t}(c_h) > F_{1t}(c_h)$, $h = 1, \dots, C - 1$, $t = 1, \dots, V$, can be tested by the joint comparison of $(C - 1) \times V$ random frequencies, like $\hat{F}_{2t}(c_h)$ and $\hat{F}_{1t}(c_h)$, where, of course, the resulting 24 partial tests are dependent in quite a complex way, with dependence coefficients depending on all underlying unknown parameters. That is why for their analyses it is necessary to stay within the permutation approach and the NPC.

The related data are in the following Table 1:

TABLE 1
Data on tip shoulder pain observed along time in 2 groups of patients

Treatment	Y						N						
	Time						Time						
ID	1	2	3	4	5	6	ID	1	2	3	4	5	6
1	1	1	1	1	1	1	23	5	2	3	5	5	4
2	3	2	1	1	1	1	24	1	5	3	4	5	3
3	3	2	2	2	1	1	25	4	4	4	4	1	1
4	1	1	1	1	1	1	26	4	4	4	4	4	3
5	1	1	1	1	1	1	27	2	3	4	3	3	2
6	1	2	1	1	1	1	28	3	4	3	3	3	2
7	1	3	2	1	1	1	29	3	3	4	4	4	3
8	2	2	1	1	1	1	30	1	1	1	1	1	1
9	1	1	1	1	1	1	31	1	1	1	1	1	1
10	3	1	1	1	1	1	32	1	5	5	5	4	3
11	1	1	1	1	1	1	33	1	3	2	2	1	1
12	2	1	1	1	1	2	34	2	2	3	4	2	2
13	1	2	2	2	2	2	35	2	2	1	3	3	2
14	3	1	1	1	3	3	36	1	1	1	1	1	1
15	2	1	1	1	1	1	37	1	1	1	1	1	1
16	1	1	1	1	1	1	38	5	5	5	4	3	3
17	1	1	1	1	1	1	39	3	3	3	3	1	1
18	2	1	1	1	1	1	40	5	4	4	4	2	2
19	4	4	2	4	2	2	41	1	3	3	3	3	1
20	4	4	4	2	1	1							
21	1	1	1	2	1	1							
22	1	1	1	2	1	2							

3. The two-sample basic problem

In order to find suitable general solutions to the methodological testing problems raised by the motivating medical example, our proposal is to stay within the theory

of conditional inference where the conditioning is on a set of sufficient statistics, under the working null hypothesis $\mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2$, for the underlying distribution F . This implies to stay within the permutation theory and the NPC of dependent tests. Indeed, with a clear meaning of the symbols, indicating by $p_F(\mathbf{X})$ the likelihood associated with the population distribution F , the likelihood of any two-sample data $\mathbf{X} = (\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1}; \mathbf{X}_{21}, \dots, \mathbf{X}_{2n_2})$ is

$$p_F(\mathbf{X}) = \prod_{i=1}^{n_1} p_{F_1}(\mathbf{X}_{1i}) \prod_{i=1}^{n_2} p_{F_2}(\mathbf{X}_{2i}),$$

where $\mathbf{X}_{ji} = (X_{j1i}, \dots, X_{jVi})$ are individual V -dimensional time trajectories. This, when $F_1 = F_2$, as stated by the working null hypothesis, is invariable with respect to any rearrangements \mathbf{X}^* , i.e. permutations, of observed trajectories \mathbf{X} . That is: $p_F(\mathbf{X}) = p_F(\mathbf{X}^*)$ is a permutation invariable likelihood. Of course, such a property is not true under the alternative where $F_{1t} \neq F_{2t}$, for some t . Moreover, provided that the likelihood is positive on the observed data set, i.e. $p_F(\mathbf{X}) > 0$, *the pooled data \mathbf{X} under the working null hypothesis is always a set of sufficient statistics for the underlying F* . In addition, especially when p_F lies outside the regular exponential family or n is smaller than the number of parameters of F , \mathbf{X} is minimal sufficient. Thus, conditioning on \mathbf{X} makes any inference to be independent of F .

Actually, *the null conditional probability, given \mathbf{X} , of every event A member of a suitable family of events \mathcal{A} , is independent of F ; specifically: $\forall F$ and $\forall A \in \mathcal{A}$, $\Pr\{\mathbf{X}^* \in A; F|\mathbf{X}\} = \Pr\{\mathbf{X}^* \in A|\mathbf{X}\}$.*

This makes permutation inferences distribution-free and nonparametric. In particular, it is to put into evidence that the nonparametric property is especially important with respect to the set of dependence coefficients involved by the set of partial tests into which the testing problem is broken-down.

Indeed, *if $\mathbf{T} = (T_1, \dots, T_K)$ are $K \geq 1$ permutation statistics (e.g. tests) and if $\varphi: \mathcal{R}^K \rightarrow \mathcal{R}^1$ is any measurable function, then the permutation null distribution of φ is independent of F ; noticeably, it is independent of all dependence parameters underlying \mathbf{T} . To be specific:*

$$\begin{aligned} \Pr\{\varphi(T_1^*, \dots, T_K^*) \leq z; F|\mathbf{X}\} &= \Pr\{\varphi(T_1^*, \dots, T_K^*) \leq z; |\mathbf{X}\} \\ &= \Pr[\varphi_{\mathbf{T}}^{-1}(z)|\mathbf{X}] = \frac{\#\{\mathbf{X}^* \in \varphi_{\mathbf{T}}^{-1}(z)\}}{M}, \end{aligned}$$

since, due to measurability of φ , $\varphi_{\mathbf{T}}^{-1}(z) \in \mathcal{A}$ for every $z \in \mathcal{R}^1$.

We emphasize that this latter property is central for the NPC approach where the combining functions should satisfy general testing properties (e.g. [25], [27], [30]). Also, it is important to emphasize that the number K into which the hypothesis is broken-down can be unrelated to the number V of variables (time repetitions) and with the number of parameters defining F .

In practice, indicating by $\mathbf{X} = \{\mathbf{X}(i), i = 1, \dots, n; n_1, n_2\}$ the $n_1 + n_2 = n$ data trajectories, where it is intended that the first n_1 vectors in the list are from the

first sample and the rest from the second, a random permutation $\mathbf{X}^* \in \Pi(\mathbf{X})$ can be obtained as $\mathbf{X}^* = \{\mathbf{X}(u_i^*), i = 1, \dots, n; n_1, n_2\}$, where $\mathbf{u}^* = \{u_1^*, \dots, u_n^*\}$ is a random permutation of unit labels $\mathbf{u} = \{1, \dots, n\}$. Thus, the permuted sub-table at time t , $t = 1, \dots, V$, associated with \mathbf{X}^* is computed as

$$\{f_{jth}^* = \#(X_{jti}^* \in c_h), h = 1, \dots, C, j = 1, 2\},$$

where: $\#(\cdot)$ is the number of elements of \mathbf{X}^* that satisfy (\cdot) ; $\mathbf{X}_{1i}^* = \mathbf{X}(u_i^*)$ for $i \leq n_1$ and $\mathbf{X}_{2i}^* = \mathbf{X}(u_i^*)$ for $n_1 < i \leq n$. Of course, each permuted sub-table is such that $f_{\cdot th} = f_{1th} + f_{2th} = f_{1th}^* + f_{2th}^* = f_{\cdot th}^*$, $h = 1, \dots, C$, and so marginal frequencies $f_{\cdot th}$, as well as cumulative marginal frequencies $N_{\cdot th} = N_{1th} + N_{2th} = N_{\cdot th}^*$ with $N_{jth} = \sum_{s \leq h} f_{jts}$, are permutation invariable quantities. Thus, permuted EDFs are:

$$\hat{F}_{jt}^*(c_h) = \sum_{s \leq h} f_{jts}^*/n_j, h = 1, \dots, C, t = 1, \dots, V, j = 1, 2.$$

4. The one-dimensional testing problem

Let us fix our attention to one time point t , first. The related partial testing problem is then:

$$H_{0t} : X_{1t} \stackrel{d}{=} X_{2t} \equiv \bigcap_{h=1}^{C-1} [F_{1t}(c_h) = F_{2t}(c_h)],$$

against

$$H_{1t} : X_{1t} \stackrel{d}{<} X_{2t} \equiv \bigcup_{h=1}^{C-1} [F_{1t}(c_h) > F_{2t}(c_h)],$$

where, according to Roy [32] *union-intersection* methodology, the hypotheses have been broken-down into $C - 1$ simpler one-sided sub-problems. Of course, each sub-problem is concerning the comparison of two unknown frequencies. Clearly, that problem can be properly solved by joint analysis of $C - 1$ dependent partial tests and then by their NPC.

The $C - 1$ partial test statistics we propose are:

$$T_{th}^* = (\hat{F}_{1th}^* - \hat{F}_{2th}^*) [\bar{F}_{\cdot th}(1 - \bar{F}_{\cdot th})]^{-\frac{1}{2}}, h = 1, \dots, C - 1,$$

where: $\hat{F}_{jth}^* = \hat{F}_{jt}^*(c_h) = N_{jth}^*/n_j$, $j = 1, 2$, $\bar{F}_{\cdot th} = N_{\cdot th}/n$ are permutation and marginal EDFs. Note that EDFs \hat{F}_{jth}^* are maximum likelihood unbiased estimates of population CDFs $F_{jth}(c_h)$, $h = 1, \dots, C - 1$, $j = 1, 2$. Also note that: i) large values of each partial test T_{th}^* are significant; ii) the $C - 1$ partial tests T_{th}^* are positively dependent; iii) 0 is assigned to expressions with the form 0/0; iv) each T_{th}^* is a reformulation of Fisher's exact probability test; v) each T_{th}^* is conditionally

optimal with conditional variance $\sigma_{T_{th}^*}^2 = n_1 n_2 (n - 1) / n^2$ not dependent on (t, h) ; vi) asymptotically each T_{th}^* is normally distributed.

Their NPC can be achieved according to the methods discussed in [25], [27], [30]. The simplest way of combination is, for instance, by their direct sum:

$$T_{ADt}^* = \sum_{h=1}^{C-1} \left(\hat{F}_{1th}^* - \hat{F}_{2th}^* \right) [\bar{F}_{.th}(1 - \bar{F}_{.th})]^{-\frac{1}{2}}.$$

It is worth noting that T_{ADt} looks like the Anderson-Darling goodness-of-fit test (sum of dependent standardized partial statistics) for directional alternatives. T_{ADt} is provided with some nice properties (e.g. [25], [27], [30]): i) as all partial tests T_{th}^* are unbiased, it is unbiased; ii) as at least one of T_{th}^* is consistent, it is consistent; iii) as its combined acceptance region is convex, it is admissible; iv) it is an admissible combination of conditionally optimal partial tests, and so there does not exist any other test G , for the same hypotheses and within the same conditions, that is uniformly better than it.

The p -value-statistic related to the pair (T_{ADt}, \mathbf{X}) is defined as $\lambda_{ADt} = \Pr\{T_{ADt}^* \geq T_{ADt}^o | \mathbf{X}\}$, where the conditioning on actual data set \mathbf{X} is emphasized and $T_{ADt}^o = T_{ADt}(\mathbf{X})$ represents the observed value of T_{ADt} on data \mathbf{X} . According to the general testing rule, if $\lambda_{ADt} \leq \alpha$ then H_0 is rejected at significance level $\alpha > 0$.

To justify the NPC solution, let us consider the representation, related to a general problem with K partial tests, R random permutations and combining function ψ , that follows:

T_1^o	T_{11}^*	\cdots	T_{1r}^*	\cdots	T_{1R}^*
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
T_K^o	T_{K1}^*	\cdots	T_{Kr}^*	\cdots	T_{KR}^*
			\downarrow		
T_ψ^o	$T_{\psi 1}^*$	\cdots	$T_{\psi r}^*$	\cdots	$T_{\psi R}^*$

where: the first column of first sub-table contains the observed values of K partial tests, i.e. $T_k^o(\mathbf{X})$, $k = 1, \dots, K$; the r -th column contains the values of the K partial tests at the r -th random permutation \mathbf{X}_r^* , $r = 1, \dots, R$. The first element of second sub-table contains the observed value of a combined tests obtained by the combining function ψ , i.e. $T_\psi^o = \psi(T_1^o, \dots, T_K^o)$; the r -th element is the permutation value of combined test ψ at the r -th data permutation.

If the working null hypothesis would be true the sub-matrix $\{T_{kr}^*\}_{K \times R}$ simulates the K -dimensional null distribution of K partial permutation tests. Accordingly, the sub-vector $\{T_{\psi r}^*\}_R$ simulates the null permutation distribution of combined test ψ . Thus, the statistic $\hat{\lambda}_\psi = \#(T_{\psi r}^* \geq T_\psi^o) / R$ gives an unbiased and strongly consistent estimate, as R diverges, of the p -value statistic of combined test T_ψ . If H_0 would not be true, at least one of partial tests would give larger observed values than in H_0 and so, if combining function ψ is non-decreasing in each argument, the p -value

statistic satisfies the relation: $\hat{\lambda}_{\psi H_1} \leq \hat{\lambda}_{\psi H_0}$ uniformly for every data set \mathbf{X} and for every underlying distribution F . The latter implies the unbiasedness property and so justifies that if $\hat{\lambda}_{\psi} \leq \alpha$ then H_0 is rejected. Formal proofs of these properties are in [25], [27], [28], [30].

The same unidimensional problem (e.g. [1], [29]) can also be tackled by assigning non-decreasing W scores to ordered classes, e.g. as $c_h \rightarrow w_h$, with $w_1 \leq w_2 \leq \dots \leq w_C$, and $w_1 < w_C$. In such a case the data are transformed into $w_{hi} = w_h \cdot \mathbf{1}(X(i) = c_h)$, $i = 1, \dots, n$, where $\mathbf{1}(\cdot)$ is the counting function. Thus, the permutation solution is nothing else than a comparison of sample means of scores: $T_W^* = \bar{w}_2^* - \bar{w}_1^*$. One further solution is by transforming data X_{jti} into ranks $R_{jti} = \#(X_{jtl} \leq X_{jti})$ or mid-ranks $M_{jti} = \#(X_{jtl} < X_{jti}) + \#(X_{jtl} = X_{jti})/2$, $i = 1, \dots, n_j$, $t = 1, \dots, V$, $j = 1, 2$, and then to proceed, in the spirit of Mann and Whitney, by comparing mean of ranks and of mid-ranks: $T_{Rt}^* = \bar{R}_{2t}^* - \bar{R}_{1t}^*$ and $T_{Mt}^* = \bar{M}_{2t}^* - \bar{M}_{1t}^*$, respectively.

Clearly, although unbiased, consistent and easy to interpret, these three further solutions suffer from the arbitrary substitution of categorical data with numerical quantities. We will see, however, that with the data of the example their inferences are closely concordant in practice.

5. The multidimensional problem

Let us now consider the multidimensional hypotheses under testing: where the null hypothesis $H_0 : \bigcap_{t=1}^V \bigcap_{h=1}^{C-1} [F_{1t}(c_h) = F_{2t}(c_h)]$ against the set of restricted alternatives $H_1 : \bigcup_{t=1}^V \bigcup_{h=1}^{C-1} [F_{1t}(c_h) > F_{2t}(c_h)]$. This is nothing else than a simple extension of the Anderson-Darling goodness-of-fit type solution, shown for the unidimensional case. Thus, with a clear meaning of the symbols, its solution is achieved by the test statistic:

$$T_{AD}^* = \sum_{t=1}^V \sum_{h=1}^{C-1} \left(\hat{F}_{1th}^* - \hat{F}_{2th}^* \right) [\bar{F}_{\cdot th}(1 - \bar{F}_{\cdot th})]^{-\frac{1}{2}}.$$

Similarly to the unidimensional formulation, if the alternative is true then at least one summand assumes values not smaller than under H_0 . So, that test is unbiased, consistent and admissible. In place of the direct combination of V partial tests T_{ADt}^* , i.e. one Anderson-Darling test for each variable, it is possible to think of a more general combination like $T_{\psi}^* = \psi(T_{AD1}^*, \dots, T_{ADV}^*)$. The most commonly used of combining functions ψ is Fisher's $T_F = -2 \sum_t \log(\lambda_{ADt})$, where λ_{ADt} are the p -value statistics of T_{ADt}^* .

Likewise, the unidimensional setting, the multidimensional problem (e.g. [1], [29]) can also be tackled by assigning non-decreasing W_t scores to ordered classes, e.g. as $c_{th} \rightarrow w_{th}$, $h = 1, \dots, C$, $t = 1, \dots, V$, where $w_{t1} \leq \dots \leq w_{tC}$, with at least one strict inequality $\forall t$. In such a case the data are transformed into $w_{thi} = w_{th} \cdot \mathbf{1}(X_{jti} = c_{th})$. Thus, the permutation solution is nothing else than a multivariate comparison of

sample means of scores:

$$T_{\psi W}^* = \psi[(\bar{w}_{12}^* - \bar{w}_{11}^*), \dots, (\bar{w}_{V2}^* - \bar{w}_{V1}^*)].$$

And so on also for $T_{\psi R}^*$ and $T_{\psi M}^*$ with ranks and mid-ranks, respectively.

6. Solution to the example

The analyses of data of the medical example, based on $R = 100\,000$ random permutations, for T_{AD}^* , T_W^* , and T_M^* give the listed results:

t	1	2	3	4	5	6	$\hat{\lambda}$
T_{AD}^*	0.05848	0.00044	0.00005	0.00002	0.00024	0.00622	0.000035
T_W^*	0.09036	0.00045	0.00005	0.00002	0.00026	0.00739	0.000025
T_M^*	0.10795	0.00035	0.00007	0.00003	0.00026	0.00736	0.000075

where scores are $w_k = k$, $k = 1, \dots, K$. These results exhibit quite strong evidence for the distributional higher pain scores, i.e. stochastic dominance, of patients taking N with respect to those taking Y . We also see that three tests give almost exactly the same results.

In order to take account, at least approximately, of some unobserved covariates and of baseline data (i.e. scores W at time $t = 1$), we also conducted a multivariate dominance analysis on individual score differences $D_{ti} = W_{ti} - W_{1i}$, $t = 2, \dots, 6$, $i = 1, \dots, n$. The related results, still based on $R = 100\,000$ random permutations, with Fisher's combined test $T_F^* = \psi_F[(\bar{D}_{22}^* - \bar{D}_{21}^*), \dots, (\bar{D}_{62}^* - \bar{D}_{61}^*)]$ are:

t	2	3	4	5	6	$\hat{\lambda}_F$
T_F^*	0.03729	0.00631	0.00162	0.07842	0.47577	0.01731

From these results it seems that at time 6 pain score difference from baseline of two groups of patients manifest non-significant diversity; whereas on all other observation times, say 2 to 5, there seems to remain a significant greater pain on patients taking N with respect to those taking Y . This latter conclusion relies on a partial analysis bounded to times $2 \div 5$, whose combined test leads to $\hat{\lambda}_{F,2-5} = 0.00504$.

7. Concluding remarks

The NPC of dependent test method is suitable and effective for many multivariate testing problems which are very difficult or even impossible to solve within traditional likelihood parametric frameworks.

One major feature of the NPC, provided that the permutation testing theory applies, is that one must pay attention to a set of partial tests into which the problem is equivalently broken-down (Roy [32]). It is required that each partial test has to be appropriate for the related sub-hypotheses, because the underlying dependence

relation structure is nonparametrically captured by the conditional combining procedure. This aspect is of great importance especially for non-normal and categorical variables, in which dependence relations are generally difficult to define and, even when well-defined, are too hard to cope with especially with multivariate ordered categorical variables. The researcher is required to make sure that all partial tests are marginally appropriate, i.e. unbiased, at least one is consistent and larger values of each partial test are significant for the specific sub-alternative. The latter are, indeed, sufficient conditions for the NPC that are generally easy to check.

As the NPC method is conditional on a set of sufficient statistics, it shows good general power behavior. Monte Carlo experiments on ordered categorical data (reported in: [1], [10], [19], [29]), show that the direct and Fisher's combining functions on Anderson-Darling partial tests T_{ADt}^* and T_M^* have generally good power behavior both for balanced and unbalanced samples and, in some situations, they are slightly more powerful than the Mann-Whitney-like tests T_W^* and T_R^* . Moreover, NPC tests for restricted alternative compared with parametric competitors, when these exist, are relatively more efficient and much less demanding in terms of underlying assumptions.

References

- [1] ARBORETTI, G. R. and BONNINI, S. (2009). Some new results on univariate and multivariate permutation tests for ordinal categorical variables under restricted alternatives. *Stat. Methods Appl.* **18** 221–236. MR2515480
- [2] BASSO, D., PESARIN, F., SALMASO, L., and SOLARI, A. (2009). *Permutation tests for stochastic ordering and ANOVA: theory and applications with R*. Springer-Verlag, New York.
- [3] BAZYARI, A. and PESARIN, F. (2013). Parametric and permutation testing for multivariate monotonic alternatives. *Stat. Comput.* **23** 639–652. MR3094804
- [4] BRUNNER, E. and LANGER, F. (2000). Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biom. J.* **42** 663–675. MR1792530
- [5] COHEN, A., KEMPERMAN, J. H. B., and SACKROWITZ, H. B. (2000). Properties of likelihood inference for order restricted models. *J. Multivariate Anal.* **72** 50–77. MR1747423
- [6] COHEN, A., MADIGAN, D., and SACKROWITZ, H. B. (2003). Effective directed tests for models with ordered categorical data. *Aust. N. Z. J. Stat.* **45** 285–300. MR1999512
- [7] COLOMBI, R. and FORCINA, A. (2016). Testing order restrictions in contingency tables. *Metrika* **79** 73–90.
- [8] DARDANONI, V. and FORCINA, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context *J. Amer. Statist. Assoc.* **93** 1112–1123. MR1649205
- [9] DYKSTRA, R., KOCHAR, S., and ROBERTSON, T. (1995). Inference for likelihood

- ratio ordering in the two-sample problem. *J. Amer. Statist. Assoc.* **90** 1034–1040. MR1354020
- [10] FINOS, L., PESARIN, F., SALMASO, L., and SOLARI, A. (2008). Exact inference for multivariate ordered alternatives. *Stat. Methods Appl.* **17** 195–208. MR2395118
- [11] GÖKPINAR, F., GÖKPINAR, E., and BAYRAK, H. (2017). Permutation approach for ordinal preference data. *Comm. Statist. Simulation Comput.* **46** 2321–2332. MR3625284
- [12] HIROTSU, C. (1982). Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* **69** 567–577. MR695203
- [13] HIROTSU, C. (1986). Cumulative chi-squared statistic as a tool for testing goodness of fit. *Biometrika* **73** 165–173. MR836444
- [14] HUANG, H. (2019). Permutation tests for stochastic ordering with ordered categorical data. *PhD thesis, University of Padova, Italy.*
- [15] HUANG, H., PESARIN, F., ARBORETTI, R., and CECCATO, R. (2019). Multivariate permutation tests for ordered categorical data. *Manuscript submitted for publication.*
- [16] JELIZAROW, M., CIEZA, A., and MANSMANN, U. (2015). Global permutation tests for multivariate ordinal data: alternatives, test statistics and the null dilemma. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 191–213. MR3293925
- [17] JOE, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London. MR1462613
- [18] KATERI, M. and AGRESTI, A. (2013). Bayesian inference about odds ratio structure in ordinal contingency tables. *Environmetrics* **24** 281–288. MR3083961
- [19] KLINGENBERG, B., SOLARI, A., SALMASO, L., and PESARIN, F. (2009). Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics* **65** 452–462. MR2751468
- [20] LUMLEY, T. (1996). Generalized Estimating Equations for Ordinal Data: A Note on Working Correlation Structures. *Biometrics* **52** 354–361.
- [21] MUNZEL, U. and TAMHANE, A. C. (2002). Nonparametric multiple comparisons in repeated measures designs for data with ties. *Biom. J.* **44** 762–779.
- [22] NEUHÄUSER, M. (2006). An exact test for trend among binomial proportions based on a modified Baumgartner-Weiß-Schindler statistic. *J. Appl. Stat.* **33** 79–88. MR2222067
- [23] PERLMAN, M. D. and WU, L. (2002). A defense of the likelihood ratio test for one-sided and order-restricted alternatives. *J. Statist. Plann. Inference* **107** 173–186. MR1927762
- [24] PESARIN, F. (1994). Goodness of fit testing for ordered discrete distributions by resampling techniques. *Metron* **52** 57–71.
- [25] PESARIN, F. (2001). *Multivariate Permutation Tests: with Application to Biostatistics*. John Wiley & Sons, Ltd, Chichester
- [26] PESARIN, F. (2004). Alcuni problemi di verifica delle ipotesi per variabili categoriali. *Statistica, LXIV* 367–386.

- [27] PESARIN, F. (2006). Permutation test: Multivariate. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., New York.
- [28] PESARIN, F. (2015). Some elementary theory of permutation tests. *Comm. Statist. Theory Methods* **44** 4880–4892. [MR3424814](#)
- [29] PESARIN, F. and SALMASO, L. (2006). Permutation tests for univariate and multivariate ordered categorical data. *Austrian J. Stat.* **35** 315–324.
- [30] PESARIN, F. and SALMASO, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, Ltd, Chichester.
- [31] PESARIN, F., SALMASO, L., CARROZZO, E., and ARBORETTI, R. (2016). Union-intersection permutation solution for two-sample equivalence testing. *Stat. Comput.* **26** 693–701. [MR3489865](#)
- [32] ROY, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statistics* **24** 220–238. [MR57519](#)
- [33] SAMPSON, A. R. and WHITAKER, L. R. (1989). Estimation of multivariate distributions under stochastic ordering. *J. Amer. Statist. Assoc.* **84** 541–548. [MR1010343](#)
- [34] SILVAPULLE, M. J. and SEN, P. K. (2005). *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. John Wiley & Sons, Inc, Hoboken, NJ.
- [35] TROENDLE, J. F. (2002). A likelihood ratio test for the nonparametric Behrens-Fisher problem. *Biom. J.* **44** 813–824. [MR1934965](#)
- [36] WANG, Y. (1996). A likelihood ratio test against stochastic ordering in several populations. *J. Amer. Statist. Assoc.* **91** 1676–1683. [MR1439109](#)
- [37] WEIß, C. H. (2019). Distance-based analysis of ordinal data and ordinal time series. *J. Amer. Statist. Assoc.* 1–37.
- [38] ZHANG, J. T., GUO, J., ZHOU, B., and CHENG, M. Y. (2020). A simple two-sample test in high dimensions based on L^2 -norm. *J. Amer. Statist. Assoc.* **115** 1011–1027.

Department of Statistical Sciences
 via C. Battisti, 241
 35121 PADOVA Italy
 e-mail: huang@stat.unipd.it
pesarin@stat.unipd.it