Statistica Sinica Preprint No: SS-2024-0255	
Title	Testing for Treatment Effect in Multitreatment Case
Manuscript ID	SS-2024-0255
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0255
Complete List of Authors	Pier Luigi Conti, Livia De Giovanni and Ayoub Mounim
Corresponding Authors	Livia De Giovanni
E-mails	ldegiovanni@luiss.it
Notice: Accepted author version.	

Testing for treatment effect in multitreatment case

Pier Luigi Conti, Livia De Giovanni, Ayoub Mounim

Università di Roma "La Sapienza" (pierluigi.conti@uniroma1.it), Luiss University (ldegiovanni@luiss.it), Luiss University (amounim@luiss.it)

Abstract: In this paper the problem of testing for the presence/absence of a (multi-level)treatment effect is considered. A new test-statistic, essentially based on the same principles as the classical Kruskal-Wallis test, is introduced, and its theoretical properties are studied. Test-statistics for stochastic dominance problems are also studied. The good behaviour of the proposed test in terms of both significance level and power, with respect to other commonly used test procedures, is showed through a simulation study. Finally, an application to real data is provided.

Key words and phrases: Kruskal-Wallis, Multitreatment, Nonparametrics, Stochastic dominance.

1. Introduction

Assessing whether a treatment with different levels has effect on an outcome is of primary interest in Statistics. In an experimental framework, based on a controlled random assignment of subjects to treatment levels, a fundamental tool is ANOVA (ANalysis Of VAriance), essentially devoted to compare different independent samples, each corresponding to a different treatment level. When Normality of observations cannot be assumed, it is customary to use the Kruskal-Wallis test, that plays a role similar to that of the F-test in case of Normal, homoskedastic observations.

The situation is considerably more intricate in observational studies, in particular when the assignment-to-treatment mechanism is not controlled and "purely random". Due to the presence of confounding covariates, there could be relevant differences among subjects receiving different treatment levels.

In causal studies, the problem of comparing a single treatment *vs.* a control is widely studied. Bootstrap-based tests to assess the distributional consequences of a treatment on some outcome variable in (possibly) non-randomized assignment-to-treatment mechanisms, when a binary instrument is available for the researcher, are studied in Abadie (2002). Permutation tests are studied in Ding (2017), Wu and Ding (2018). Nonparametric tests, that are essentially extension of Kolmogorov-Smirnov and Wilcoxon tests, are proposed in Conti and De Giovanni (2022).

The "usual" approach to the problem of comparing the effects of a treatment with several levels seems to consist in focusing on pairwise ATE comparisons; cfr. Yang et al. (2016), Li and Li (2019). Simultaneous consideration of all pairwise comparisons, for instance combined through the Bonferroni rule, is not studied. The importance of defining a common support region when studying multi-level treatments, where differences in the implementation of certain approaches can vary the causal estimands and change the study population to which inference is generalizable, as well, is dealt with in Lopez and Gutman (2017). A technique, Vector Matching, for generating matched sets balanced with respect to covariates distribution when there are more than two categorical treatments, and that addresses some of the pitfalls of the current method, is introduced and compared to previously proposed algorithms to reduce the bias on observed covariates. The causal estimands for the proposed and existing matching algorithms is the pairwise average treatment effect, population and sample (PATE, SATE).

The main goal of the present paper is to develop some ANOVA-type tools for observational studies, where the assignment-to-treatment mechanism depends on uncontrolled covariates. In more detail, our first goals is to develop a test that plays the same role as the Kruskal-Wallis test in experimental studies. When the hypothesis of

no treatment effect is rejected, it is also of interest to find out whether there is any particular pattern of the effects for different levels of treatment. For instance, there could be a treatment level whose effect is highest in terms of outcome, or there could be some order among the effects of the different treatment levels, again in terms of outcome. This kind of analysis implies that it is first necessary to define what it means that a level of treatment is higher that another one in terms of outcome. This entails first the introduction of a stochastic dominance relationship between outcomes corresponding to different levels of treatment. Secondly, statistical tests are proposed for testing hypotheses concerning stochastic dominance relationships.

According to the above remarks, the main contributions offered by the present paper are essentially two.

- Construction of a new non-parametric test for comparing the effect of different treatment levels. This would provide a non-parametric ANOVA-type tool, based on a suitable extension of the classical Kruskal-Wallis test to observational studies. This test is of asymptotically exact significance level, and is proved to be superior if compared to test based on pairwise comparisons.

- Construction of new statistical tests to asses whether a treatment level dominates one or more other levels, in terms of the distribution of corresponding outcomes. As already remarked, if compared to the case of a single level *vs.* a control (cfr. Conti and De Giovanni (2022), Donald and Hsu (2014)), the case of several treatment levels offers different stochastic dominance patterns. It is interesting to note that the stochastic dominance relationship among treatment levels actually allows to order them, either partially or totally. In addition, the ordering between levels of treatment is not imposed a priori; we simply study whether it exists on the basis of the corresponding effects on outcomes.

The paper is organized as follows. Section 2 is devoted to a basic description of the problem. In Section 3 asymptotic preliminary results, on which all subsequent results rest, are provided. In Section 4 a Kruskal-Wallis type test is studied, and Section 5 deals with various problems of testing for stochastic dominance. Finally, in Section 6 a simulation study to compare our proposal to other existing in the literature is performed, and in Section 7 and application to real data is considered. Technical parts and proofs of results are deferred to Supplementary Materials.

2. Problem description

Consider n independent units, each receiving a treatment T with $K+1$ levels, denoted by 0, 1, ..., K. The level 0 conventionally corresponds to the absence of treatment, *i.e.* to the control group. Next, denote by $Y_{(k)}$, $k = 0, 1, \ldots, K$ the *potential outcome* of a subject when treatment is at level k. The *observed outcome* for a unit is then

$$Y = \sum_{k=0}^{K} Y_{(k)} I_{(T=k)}, \tag{1}$$

where

$$I_{(T=k)} = \begin{cases} 1 \text{ if } T = k \\ 0 \text{ if } T \neq k \end{cases}$$

is the indicator function of the event $T = k$.

From now on, we will say that treatment T has no effect (in distribution) whenever $Y_{(0)}, Y_{(1)}, \ldots, Y_{(K)}$ have the same probability distribution. Denoting by $\stackrel{d}{=}$ the equality in distribution, the absence of treatment effect is equivalent to say that $Y_{(0)} \stackrel{d}{=} Y_{(1)} \stackrel{d}{=} \cdots \stackrel{d}{=} Y_{(K)}$. This is essentially a probabilistic version of the *sharp hypothesis* of absence

of treatment effect in experimental studies; cfr. Ding (2017), Wu and Ding (2018). The symbol $F_k(y) = P(Y_{(k)} \leq y)$, $y \in \mathbb{R}$, will denote in the sequel the distribution function (d.f.) of potential outcome $Y_{(k)}$, $k = 0, \ldots, K$.

As already said, the assignment-to-treatment mechanism is not a controlled, "purely random", mechanism. Due to the presence of confounding covariates, there could be relevant differences among subjects receiving different treatment levels. In the present paper, we focus on an assignment-to-treatment mechanism that only depends on *observed* covariates, with no unobserved confounders; the relevant covariates vector is denoted by $\boldsymbol{X} = (X_1 \cdots X_P)$. The kth (generalized) propensity score is the probability of receiving treatment at level k, conditionally on $\boldsymbol{X} = \boldsymbol{x}$,

$$p_k(\boldsymbol{x}) = P(T = k|\boldsymbol{X} = \boldsymbol{x}); \quad k = 0, \ldots, K. \tag{2}$$

Of course, the relationship $p_0(\boldsymbol{x})+p_1(\boldsymbol{x})+\cdots+p_K(\boldsymbol{x}) = 1$ holds. The basic assumptions on which the paper rests are listed below.

H1. *Strong unconfoundedness.* $T \perp\!\!\!\perp (Y_{(0)}, Y_{(1)}, \ldots, Y_{(K)})|\boldsymbol{X}$.

H2. *Common support.* There exists a positive real ε for which $\varepsilon \leq p_k(\boldsymbol{x}) \leq 1 - \varepsilon$ for each $\boldsymbol{x}$ and $k = 0, 1, \ldots, K$.

H3. *Compactness.* The support $\mathcal{X}$ of $\boldsymbol{X}$ is a compact subset of $\mathbb{R}^P$.

Observed data for n subjects are the triplets $(Y_i, T_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. The r.v.s $(Y_i, T_i, \boldsymbol{X}_i)$ are assumed to be independent and identically distributed (*i.i.d.*).

In both parametric / non- or semi-parametric situations, the inferential focus is often on determining if any of the $K + 1$ treatments does have a different effect. The

main hypothesis to be tested is the *absence of treatment effect*, namely

$$\begin{cases} H_0: & F_0 = F_1 = \cdots = F_K \\ H_1: & \text{The d.f.s } F_k\text{s do not coincide} \end{cases}. \tag{3}$$

Other hypotheses of interest will be considered in Section 5.

3. Preliminary results

In order to estimate the d.f.s F_ks, an important preliminary step consists in estimating the propensity scores $p_k(\boldsymbol{x})$, $k = 0, 1, \ldots, K$. Here we just confine ourselves to some simple, informal ideas. More precise results are in the Supplementary Material.

If the (generalized) propensity score $p_k(\boldsymbol{x})$ is smooth enough, it can be approximated as

$$p_k(\boldsymbol{x}) \approx p_k^w(\boldsymbol{x}) = \frac{\exp\{\boldsymbol{x}'_{vec,L}\boldsymbol{\pi}_{k,L}\}}{1 + \sum_{k=1}^{K} \exp\{\boldsymbol{x}'_{vec,L}\boldsymbol{\pi}_{k,L}\}}, \; k = 0, 1, \ldots, K \tag{4}$$

where $\boldsymbol{x}_{vec,L}$ is a L-dimensional vector of orthogonal polynomials in the covariates, and $\boldsymbol{\pi}_{k,L}$ is a vector of coefficients. In other terms, each covariate is considered as a polynomial, including both main effects and higher-order terms (corresponding to interactions among covariates).

Next, $\boldsymbol{\pi}_{k,L}$ are estimated by maximizing the "working likelihood" $\sum \log p_k^w(\boldsymbol{x}_i)$, with the constraint $\sum_k p_k^w(\boldsymbol{x}_i) = 1$. In symbols:

$$\widehat{\boldsymbol{\pi}}_{k,L} = \arg\max \sum_{k=0}^{K} \sum_{\substack{i=1 \\ T_i = k}}^{n} \log p_k^w(\boldsymbol{x}_i), \quad k = 1, \ldots, K.$$

As estimators of $p_k(\boldsymbol{x})$s, it is then natural to take

$$\widehat{p}_0(\boldsymbol{x}) = \frac{1}{1+\sum_{k=1}^{K}\exp\{\boldsymbol{x}'_{vec,L}\widehat{\boldsymbol{\pi}}_{k,L}\}}, \tag{5}$$

$$\widehat{p}_k(\boldsymbol{x}) = \frac{\exp\{\boldsymbol{x}'_{vec,L}\widehat{\boldsymbol{\pi}}_{k,L}\}}{1+\sum_{k=1}^{K}\exp\{\boldsymbol{x}'_{vec,L}\widehat{\boldsymbol{\pi}}_{k,L}\}}, k = 1, \ldots, K. \tag{6}$$

The rate of convergence for the above estimators is obtained in Kim (2013), Proposition 2.1 and subsequent discussion. As a result, as n, L tend to infinity, with L increasing "slowly" w.r.t. n, the following result holds

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\widehat{p}_k(\boldsymbol{x}) - p_x(\boldsymbol{x})| = o_p(n^{-1/4}),\ k = 0, 1, \ldots, K. \tag{7}$$

As remarked by a referee, a relevant practical problem is the choice of L and $\boldsymbol{x}_{vec,L}$. In the simulation (Section 6), as well as in the application (Section 7) we have adopted the same approach as in the Supplement of Firpo (2007). In detail, the order of the approximating polynomial for propensity score is obtained through cross-validation, where the optimal number of terms minimizes a Kullback–Leibler distance. Furthermore, the order in which the polynomial terms are added Is the same as in Firpo (2007).

Next, on the basis of the estimators (5)-(6), the following are considered

$$\widehat{F}_k(y) = \sum_{i=1}^{n} I_{(Y_i\le y)} w_{i,k},\ k = 0, 1, \ldots, K \tag{8}$$

where $w_{i,k} = I_{(T_i=k)}\widehat{p}_k(\boldsymbol{x}_i)^{-1} \big/ \sum I_{(T_j=k)}\widehat{p}_k(\boldsymbol{x}_j)^{-1}$.

Define next the $K+1$-variate empirical process

$$\boldsymbol{W}_n(y) = \begin{bmatrix} W_{0,n}(y) \\ \dots \\ W_{K,n}(y) \end{bmatrix} = \begin{bmatrix} \sqrt{n}(\widehat{F}_0(y) - F_0(y)) \\ \dots \\ \sqrt{n}(\widehat{F}_K(y) - F_K(y)) \end{bmatrix}. \tag{9}$$

The second result on which all subsequent sections rest, and which is in its turn an immediate generalization of Donald and Hsu (2014) and Proposition 1 in Conti and De Giovanni (2022), is reported below.

Proposition 1. *Suppose that assumptions H1-H3, are met, that* (7) *holds, that* $Y_{(k)}$ *possess finite second moments, that* $E[Y_{(k)}|\boldsymbol{x}]$ *are continuously differentiable, and that* $F_k(y)$, $F_k(y|\boldsymbol{x})$ *are continuous, for all* $k = 0, 1, \dots, K$. *Then, the sequence of stochastic processes* (9) *converges weakly, as* $n \to \infty$, *to a* $(K+1)$*-variate Gaussian process*

$$\boldsymbol{W}(y) = \begin{bmatrix} W_0(y) \\ \dots \\ W_K(y) \end{bmatrix}$$

with null mean function ($E[W_k(y)] = 0$ *for all* $k = 0, 1, \dots, K$) *and covariance kernel*

$$\boldsymbol{C}(y, t) = E[\boldsymbol{W}(y) \otimes \boldsymbol{W}(t)] = \begin{bmatrix} C_{00}(y, t) & C_{01}(y, t) & \cdots & C_{0K}(y, t) \\ \dots & & & \\ C_{K0}(y, t) & C_{K1}(y, t) & \cdots & C_{KK}(y, t) \end{bmatrix} \tag{10}$$

with

$$C_{kk}(y,\,t) \;=\; E\left[\frac{1}{p_k(\boldsymbol{X})}(F_k(y\wedge t|\boldsymbol{X}) - F_k(y|\boldsymbol{X})\,F_k(t|\boldsymbol{X}))\right] \\ +E[(F_k(y|\boldsymbol{X}) - F_k(y))\,(F_k(t|\boldsymbol{X}) - F_k(t))] \tag{11}$$

and

$$C_{hk}(t,\,y) = C_{kh}(y,\,t) = E[(F_k(y|\boldsymbol{X}) - F_k(y))\,(F_h(t|\boldsymbol{X}) - F_h(t))]. \tag{12}$$

Convergence is in the set $l^{\infty}_{K+1}(\mathbb{R})$ *of bounded functions* $\mathbb{R}\mapsto\mathbb{R}^{K+1}$ *equipped with the sup-norm.*

Remark. As remarked by the AE, the weighting scheme based on inverse weighting probability may be unstable in case of very small probability. A practical and common solution is based on trimming, *i.e.* on removing subjects with scores in between α and $1-\alpha$ ($\alpha = 0$ means no trimming); cfr. Crump et al. (2009) where, in the case $K = 1$, a theoretical analysis is performed, and the rule of thumb $\alpha = 0.1$ is proposed. Generalizations to the multi-level treatment case are in Yang et al. (2016), and asymmetric trimming is considered in Stürmer et al. (2010). The main trouble with trimming is that, despite the potential gain in estimation efficiency, the decision on how many subjects to exclude is *ad hoc*, and a large proportion of the sample could be discarded; furthermore, they are frequently sensitive to the cutoff level. An alternative proposal is the overlap weighting (OW) method in Li et al. (2018), where, again in the case $K = 1$, each subject weight is the probability the subject is assigned to the opposite group. A comparison of the performance of different methods is in Li et al. (2019).

4. Non-parametric Kruskal-Wallis type test for treatment comparison

4.1 Reconsidering the "classical" Kruskal-Wallis test

The main goal of the present section is to construct a fully non-parametric test for the hypothesis problem (3). The test proposed here is in some way inspired to the well-known Kruskal-Wallis test used in nonparametric one-way ANOVA; cfr. Hettmansperger and McKean (2011). To understand how the Kruskal-Wallis test works, consider $K+1$ sub-populations, with continuous d.f.s F_0, F_1, ..., F_K, respectively. Assume further that the population d.f. is a mixture of F_0, F_1, ..., F_K, with weights λ_0, λ_1, ..., λ_K:

$$F(y) = \sum_{k=0}^{K} \lambda_k F_k(y). \tag{13}$$

As a divergence parameter of F_ks to F, let us consider

$$\begin{aligned} \delta &= \sum_{k=0}^{K} \lambda_k \left(\int_{-\infty}^{+\infty} F(y)\, dF_k(y) - \int_{-\infty}^{+\infty} F(y)\, dF(y) \right)^2 \\ &= \sum_{k=0}^{K} \lambda_k \left(\int_{-\infty}^{+\infty} F(y)\, dF_k(y) - \frac{1}{2} \right)^2 \end{aligned} \tag{14}$$

where $\int F(y)\, dF(y) = 1/2$ because F_ks are continuous.

The idea behind the "classical" Kruskal-Wallis test can be described in relatively simple terms. Suppose $K+1$ independent samples of size n_0, ..., n_K are available for our $K+1$ sub-populations, take $n = n_0 + \cdots + n_K$, and assume that $n_k/n \to \lambda_k$ for each k. Denote by R_{ik} the rank of observation ith ($i = 1, \ldots, n_k$) of sample kth ($k = 0, 1, \ldots, K$) computed across the entire sample of all n pooled observations, by $\overline{R}_k = \sum_i R_{ik}/n_k$ the average rank of the kth sample, and by $\overline{R}$ the overall average

rank. The Kruskal-Wallis statistic is traditionally written as

$$G = (n-1)\sum_{k=0}^{K} n_k \left(\overline{R}_k - \overline{R}\right)^2 \Bigg/ \sum_{k=0}^{K}\sum_{i=1}^{n_k} \left(R_{ik} - \overline{R}\right)^2. \tag{15}$$

To see the relationship between the statistic (15) and (14), let $\widehat{F}_k^e$ be the empirical distribution function (e.d.f.) of the kth sample, and let $\widehat{F}^e$ be the overall e.d.f. for all n observations pooled together. Then, it is easy to see (15) can be equivalently re-written as:

$$G = n\, c_n \sum_{k=0}^{K} \frac{n_k}{n} \left(\int_{-\infty}^{+\infty} \widehat{F}^e(y)\, d\widehat{F}_k^e(y) - \frac{1}{2} \right)^2 + d_n$$

for appropriate constants c_n, d_n, tending to 1 and 0, respectively, as n_ks go to infinity. Hence, the Kruskal-Wallis statistic is equivalent to

$$\sum_{k=0}^{K} \frac{n_k}{n} \left\{ \sqrt{n} \left(\int_{-\infty}^{+\infty} \widehat{F}^e(y)\, d\widehat{F}_k^e(y) - \frac{1}{2} \right) \right\}^2. \tag{16}$$

Eqn. (16) suggests to take, as a Kruskal-Wallis type test for the hypotheses problem (3),

$$D_n = \sum_{k=0}^{K} \frac{n_k}{n} \left\{ \left(\sqrt{n} \int_{-\infty}^{+\infty} \widehat{F}(y)\, d\widehat{F}_k(y) - \frac{1}{2} \right) \right\}^2 \tag{17}$$

where $\widehat{F}_k$s are given by (8), and

$$n_k = \sum_{i=1}^{n} I_{(T_i=k)}, \quad k = 0, \ldots, K; \tag{18}$$

$$\widehat{F}(y) = \sum_{k=0}^{K} \frac{n_k}{n} \widehat{F}_k(y). \tag{19}$$

Note that D_n/n can be considered as an estimator of the divergence measure (14).

It is worth noticing that the divergence measure δ in (14), as well as the statistic D_n in (17), can be expressed in a different but equivalent form, useful for subsequent developments. Consider first

$$\theta_{jk} = \int_{-\infty}^{+\infty} F_j(y)\, dF_k(y),\ j,\, k = 0,\, 1,\, \ldots,\, K \tag{20}$$

which is the measure of the divergence between the two d.f.s F_j, F_k used in the (two-sample) Wilcoxon-Mann-Whitney statistic; cfr., for instance, Conti and De Giovanni (2022) and references therein. Since (as it is easily verified through an integration by parts) $\theta_{jk} = 1-\theta_{kj}$, using θ_{jk} is equivalent to use θ_{kj}. Moreover, if $j = k$ then $\theta_{kk} = 1/2$, provided F_k is continuous. Then, the measure of divergence δ can be written down in terms of θ_{jk}s as

$$\begin{aligned} \delta &= \sum_{k=0}^{K} \lambda_k \left(\int_{-\infty}^{+\infty} \sum_{j=0}^{K} \lambda_j F_j(y)\, dF_k(y) - \frac{1}{2} \sum_{j=0}^{K} \lambda_j \right)^2 \\ &= \sum_{k=0}^{K} \lambda_k \left\{ \sum_{\substack{j=0 \\ j\neq k}}^{K} \lambda_j \left(\theta_{jk} - \frac{1}{2} \right) \right\}^2 \end{aligned} \tag{21}$$

Note further that when $F_0 = F_1 = \cdots = F_K$, all θ_{jk} are equal to $1/2$, and hence $\delta = 0$. Otherwise, δ is positive.

4.2 Limiting distribution of the proposed Kruskal-Wallis type test

To find out the limiting distribution of the test-statistic D_n under H_0, define first the estimator of θ_{jk}

$$\widehat{\theta}_{jk} = \int_{-\infty}^{+\infty} \widehat{F}_j(y)\, d\widehat{F}_k(y),\quad j \neq k = 0,\, \ldots,\, K, \tag{22}$$

the vectors

$$\boldsymbol{\theta}_{kK} = [\theta_{k1} \cdots \theta_{kK}]', \ \widehat{\boldsymbol{\theta}}_{kK} = \left[\widehat{\theta}_{k1} \cdots \theta_{kK}\right]', \ k = 0, \ldots, K-1$$

and the $K(K+1)/2$-dimensional vectors

$$\boldsymbol{\theta}_{vec} = \left[\boldsymbol{\theta}'_{0K} \cdots \boldsymbol{\theta}'_{K-1K}\right], \ \widehat{\boldsymbol{\theta}}_{vec} = \left[\widehat{\boldsymbol{\theta}}'_{0K} \cdots \widehat{\boldsymbol{\theta}}'_{K-1K}\right]. \tag{23}$$

In terms of $\widehat{\theta}_{jk}$, D_n can be written as

$$D_n = \sum_{k=0}^{K} n_k \left\{ \sum_{\substack{j=0 \\ j \neq k}}^{K} \frac{n_j}{n} \left(\widehat{\theta}_{jk} - \frac{1}{2} \right) \right\}^2. \tag{24}$$

The path to find out the limiting distribution of D_n under H_0 is summarized below.

1. Study the limiting distribution of $\widehat{\boldsymbol{\theta}}_{vec}$, for general $F_0, \ldots, F_K$.

2. Obtain, as a consequence, the limiting distribution of D_n, under the null hypothesis H_0 (in this case all components of the vector $\boldsymbol{\theta}_{vec}$ are equal to 1/2).

Proposition 2. *Under the same conditions as Proposition* 1*, as* $n \to \infty$*,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{vec} - \boldsymbol{\theta}_{vec}) \xrightarrow{d} \boldsymbol{V} \tag{25}$$

where $\boldsymbol{V}$ *is a (non-singular)* $K(K+1)/2$*-dimensional Multivariate Normal r.v. with mean vector* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}_{\boldsymbol{V}}$ *with elements specified in* (S2.12)*.*

Proof. See Supplementary Material. □

Proposition 3. *Suppose the conditions as Proposition* 1 *are met. The following two statements hold.*

1. *If* $F_0 = F_1 = \cdots = F_K$, *then*

$$D_n \xrightarrow{d} \boldsymbol{V}'\boldsymbol{A}'\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{V} \text{ as } n \to \infty. \tag{26}$$

where $\boldsymbol{V}$ *is the* $K(K+1)/2$*-dimensional Multivariate Normal r.v. defined in Proposition* 2, *and* $\boldsymbol{A}\boldsymbol{\Lambda}$ *are the matrices defined in* (*S*2.16).

2. *If the hypothesis* $F_0 = F_1 = \cdots = F_K$ *is false and* $\delta > 0$, *then*

$$\lim_{n\to\infty} P\left(D_n > M\right) = 1 \ \ \forall\, M > 0. \tag{27}$$

Proof. See Supplementary Material. □

Proposition 3 shows that the test-statistic D_n is consistent provided that $\delta > 0$. Hence, exactly as in the "classical" Kruskal-Wallis test, it is not an omnibus test. Among (many) other cases, it is consistent under a shift treatment effect, *i.e.* when the effect of treatment consists in a shift of the potential outcome distribution.

Unlike the Kruskal-Wallis test, the limiting distribution of D_n under H_0 is not a Chi-square with $K+1$ degrees of freedom, but instead a linear combination of $K+1$ independent Chi-square each with 1 degree of freedom, and with coefficients related to the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{V}}$. In the next subsection, an approximation based on sub-sampling is considered.

4.3 Testing for the absence of treatment effect

The basic idea to test for the hypothesis of absence of treatment effect is to refer to the test-statistic D_n, and to reject H_0 whenever D_n takes a "large enough" value. As a consequence, the rejection region takes the form $D_n > const$. Such a rejection region

can be also interpreted in terms of simultaneous confidence region for θ_{jk}s. Define, as in $(S2.13)$,

$$\widetilde{D}_n = \sum_{k=0}^{K} \lambda_k \left\{ \sqrt{n} \sum_{\substack{j=0 \\ j \neq k}} \lambda_j \left(\hat{\theta}_{jk} - \theta_{jk} \right) \right\}^2 . \tag{28}$$

The limiting distribution of $\widetilde{D}_n$, for general θ_{jk}s, is studied in Proposition $S2$. Since $\theta_{jk} = 1/2$ when $F_0 = \cdots = F_K$, as n gets large, considering the rejection region $D_n > const$ is equivalent to construct a simultaneous confidence region for θ_{jk}s of the form $\{\boldsymbol{\theta}_{vec} : \widetilde{D}_n \leq const\}$, and to reject H_0 whenever the vector with all components equal to $1/2$ is not in the confidence region. If the confidence level of the above region is equal to $1 - \alpha$, the significance level of the test is α. Hence, denoting by $d_{1-\alpha}$ the $(1 - \alpha)$th quantile of the distribution of $\widetilde{D}_n$, the rejection region can be expressed in the form

$$D_n > d_{1-\alpha}$$

It remains to provide (at least) an approximation for $d_{1-\alpha}$. Sub-sampling technique is a simple but effective technique to approximate distribution functions and quantiles; cfr. Politis and Romano (1994). Among its nice features, it requires conditions considerably less stringent than bootstrap; see, for instance, Ch. 23 of van der Vaart (1998), where bootstrap is shown to work under a strengthened form of Hadamard differentiability. Furthermore, the computational burden of subsampling is frequently less heavy than bootstrap. Its use in the context of testing for causal effect in the treatment *vs.* control case (*i.e.* with $K = 1$) has been considered in Conti and De Giovanni (2022). Define $S_i = (Y_i, T_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, let $1 \leq m \leq n$, and consider the $n!/[m!(n-m)!]$

sub-samples of size m of $(S_1, \ldots, S_n)$. The sub-sampling procedure is described below.

1. Select (deterministically or randomly) M independent sub-samples of size m from the sample of S_is, $i = 1, \ldots, n$. In the sequel, the lth selected subsample will be denoted by $\boldsymbol{u}_l$, $l = 1, \ldots, M$. The M subsamples do not generally refer to a partition of the whole sample.

2. Denote by $\widehat{F}_{k,m;l}(y)$ the estimates of $F_k(y)$s, $k = 0, 1, \ldots, K$ obtained from lth sub-sample, let $\widehat{\theta}_{jk,m;l}$ be the corresponding estimates of θ_{jk}, and let $\widehat{\lambda}_{j,m;l} = \sum_{i \in \boldsymbol{u}_l} I_{(T_i=j)}/m$ be the estimate of λ_j obtained by the lth sub-sample.

3. Compute the statistic $\widetilde{D}_{m;l} = \sum_k \widehat{\lambda}_{k,m;l} \left\{ \sqrt{m} \sum_j \widehat{\lambda}_{j,m;l} \left(\widehat{\theta}_{jk,m;l} - \widehat{\theta}_{jk} \right) \right\}^2$, $l = 1, \ldots, M$.

4. Compute the empirical d.f. $\widehat{R}_{n,m}(z) = \sum_l I_{(\widetilde{D}_{m;l} \leq z)}/M$.

5. Compute the corresponding quantile $\widehat{R}^{-1}_{n,m}(p) = \inf\{z : \widehat{R}_{n,m}(z) \geq p\}$.

If $m \to \infty$, $m/n \to 0$ as $n \to \infty$, from Politis and Romano (1994), under the same conditions of Proposition 3, the following two results are obtained.

(i) $\sup_z \left| \widehat{R}_{n,m}(z) - P(\widetilde{D}_n \leq z) \right|$ tends in probability to 0 as n, m, M tend to infinity.

(ii) $\widehat{d}_{1-\alpha}$ tends in probability to $d_{1-\alpha}$, for every fixed $0 < \alpha < 1$.

As a consequence, at an asymptotically exact significance level α, the test for absence of treatment effect possesses rejection region

$$D_n > \widehat{d}_{1-\alpha}. \tag{29}$$

Remark. As correctly remarked by a referee, m is related not only to n (the sample size), but also to L the dimension of $\boldsymbol{x}_{vec}$. Since L is chosen prior to m, the larger L,

the larger m, which may lead to a higher computational burden. In our simulation the value $m = n^{0.8}$ has been used, with L chosen as in the supplement of Firpo (2007).

5. Testing for stochastic dominance

5.1 General aspects

In evaluating the effect of treatments, it is frequently of interest to test whether the treatments do have an effect on the whole distribution function of Y, namely whether changing the treatment level improves the behavior of the whole d.f. of Y. This implies considering some appropriate notion of stochastic dominance for treatments effects. We consider here *first order stochastic dominance*; crf. McFadden (1989), Anderson (1996). The d.f. F_k first-order stochastically dominates F_j if $F_k(y) \leq F_j(y)$ for all real y. In this case, the notation $Y_{(j)} \preceq Y_{(k)}$ will be used.

An important feature of the measure θ_{jk} (20) is that $\theta_{jk} \geq 1/2$ whenever $Y_{(j)} \preceq Y_{(k)}$ (the opposite is not necessarily true), *i.e.* θ_{jk} is *monotonic* w.r.t. first order stochastic dominance. This opens the road to its use as a test-statistics for stochastic dominance.

In the case $K = 1$, namely treatment *vs.* control, different tests for stochastic dominance have been proposed in the literature; cfr., in particular, Donald and Hsu (2014), Donald and Hsu (2016) for a nice treatment of the subject and further references, and Conti and De Giovanni (2022), where a test for stochastic dominance based on Wilcoxon statistics is studied. The main goal of the present section is to study various problems of testing for first order stochastic dominance.

5.2 Testing for the "best" treatment

The goal of this section is to construct a test for the hypothesis that a certain treatment level is the "best" one, in the sense that it stochastically dominates all other treatments.

With no loss of generality, it may be assumed that the treatment level to be tested as the best is K, so that the corresponding hypothesis problem is

$$\begin{cases} H_0: & Y_{(k)} \preceq Y_{(K)} \; \forall\, k = 0,\, 1,\, \ldots,\, K-1 \\ H_1: & Y_{(K)} \text{ does not dominate all other treatment effects} \end{cases}. \tag{30}$$

Due to the monotonicity of θ_{jk}s, the hypothesis problem (30) can be transformed in its weaker version

$$\begin{cases} H_0: & \theta_{k,K} \geq \frac{1}{2} \qquad \forall\, k = 0, \ldots K-1 \\ H_1: & \theta_{k-1,k} < \frac{1}{2} \quad \text{for some } k = 0, \ldots K-1 \end{cases}. \tag{31}$$

A simple idea to test for the hypothesis (31) at a significance level α would consist in constructing simultaneous confidence intervals for θ_{0K}, θ_{1K}, $\ldots$, $\theta_{K-1\,K}$ with overall level $1-\alpha$, and in rejecting H_0 if at least one of those intervals contains $1/2$.

From the arguments of Section S.2.3.1 in the Supplementary Material, it is seen that

$$\max_{0 \leq k \leq K-1} \sqrt{n}(\widehat{\theta}_{k\,K} - \theta_{k\,K}) \xrightarrow{d} V_1^* \;\text{ as } n \to \infty \tag{32}$$

where V_1^* possesses absolutely continuous distribution. Hence, there exists a unique $d_{1-\alpha}$ for which $P\left(V_1^* \leq d_{1-\alpha}\right) = 1-\alpha$, and in view of (32), as n gets large,

$$P\left(\max_{1 \leq k \leq K} \sqrt{n}(\widehat{\theta}_{k-1\,k} - \theta_{k-1\,k}) \leq d_{1-\alpha}\right) \simeq 1-\alpha. \tag{33}$$

From (32) and (33), for large n, the following relationships are obtained

$$\begin{aligned} 1-\alpha &\simeq P\left(\max_{0\le k\le K-1}\sqrt{n}(\widehat{\theta}_{k\,K}-\theta_{k\,K})\le d_{1-\alpha}\right) \\ &= P\left(\sqrt{n}(\widehat{\theta}_{k\,K}-\theta_{k\,K})\le d_{1-\alpha}\;\forall\, k=0,\,1,\,\ldots,\,K-1\right) \\ &= P\left(\theta_{k\,K}\ge\widehat{\theta}_{k\,K}-\frac{d_{1-\alpha}}{\sqrt{n}}\;\forall\, k=0,\,1,\,\ldots,\,K-1\right) \end{aligned}$$

from which it is easy to conclude that the K intervals

$$\left[\widehat{\theta}_{k\,K}-\frac{d_{1-\alpha}}{\sqrt{n}},\,1\right],\quad k=0,\,1,\,\ldots,\,K-1 \tag{34}$$

are simultaneous confidence intervals with approximate overall confidence level $1-\alpha$.

The quantile $d_{1-\alpha}$ can be approximated through the sub-sampling procedure described in Section 4.3. Define again $S_i=(Y_i,\,T_i,\,\boldsymbol{X}_i)$, $i=1,\,\ldots,\,n$, let $1\le m\le n$, and consider the $n!/[m!(n-m)!]$ sub-samples of size m of $(S_1,\,\ldots,\,S_n)$. The sub-sampling procedure can be described as follows.

1. Select (either deterministically or randomly) M independent sub-samples of size m from the sample of S_is, $i=1,\,\ldots,\,n$.

2. Denote by $\widehat{F}_{k,m;l}(y)$ the estimates of $F_k(y)$s, $k=0,\,1,\,\ldots,\,K$ obtained from lth sub-sample, let $\widehat{\theta}_{jk,m;l}$ be the corresponding estimates of θ_{jk}, and let $\widehat{\lambda}_{j,m;l}=\sum I_{(T_i=j)}/m$ be the estimate of λ_j obtained by the lth sub-sample.

3. Compute the statistic $V_{m;l}^{0-K}=\max_{0\le k\le K-1}\sqrt{m}\left(\widehat{\theta}_{k\,K,m;l}-\widehat{\theta}_{k\,K}\right),\quad l=1,\,\ldots,\,M.$

4. Compute the empirical d.f. $\widehat{R}_{n,m}(z)=\sum I_{(V_{m;l}^{0-K}\le z)}/M.$

5. Compute the corresponding quantile $\widehat{R}_{n,m}^{-1}(p)=\inf\{z:\;\widehat{R}_{n,m}(z)\ge p\}.$

If $m\to\infty$, $m/n\to 0$ as $n\to\infty$, the following two statements hold.

(*i*) $\sup_z \left|\widehat{R}_{n,m}(z) - P\left(\max_{0\le k\le K-1}\sqrt{n}(\widehat{\theta}_{k\,K} - \theta_{k\,K}) \le z\right)\right|$ tends in probability to 0 as n, m, M tend to infinity.

(*ii*) $\widehat{d}_{1-\alpha}$ tends in probability to $d_{1-\alpha}$, for every fixed $0 < \alpha < 1$.

As a consequence, at a significance level α, the test that rejects H_0 whenever

$$\widehat{\theta}_{k\,K} - \frac{d_{1-\alpha}}{\sqrt{n}} < \frac{1}{2} \text{ for at least one } k = 0,\, 1,\, \ldots,\, K-1$$

has approximate significance level α.

5.3 Testing for ordering of all treatments effect

The goal of the present section is the construction of a test for the hypothesis problem

$$\begin{cases} H_0: & Y_{(0)} \preceq Y_{(1)} \preceq \cdots \preceq Y_{(K)} \\ H_1: & Y_{(k)} \not\preceq Y_{(k+1)} \text{ for some } k \end{cases}. \tag{35}$$

Similarly to the above section, because of the monotonicity of θ_{jk}s, the hypothesis problem (35) can be transformed into the weaker version

$$\begin{cases} H_0: & \theta_{k-1,k} \ge \frac{1}{2} \qquad \forall\, k = 1, \ldots K \\ H_1: & \theta_{k-1,k} < \frac{1}{2} \quad \text{for some } k = 1, \ldots K \end{cases}. \tag{36}$$

Again, a test for the hypothesis (36) at a significance level α is constructed by first taking simultaneous confidence intervals for θ_{01}, θ_{12}, $\ldots$, $\theta_{K-1\,K}$ with overall level $1-\alpha$. The null hypothesis H_0 is rejected if at least one of the intervals contains $1/2$. From the arguments of Section S.2.3.2 in the Supplementary Materials, It is seen that

$$\max_{1\le k\le K}\sqrt{n}(\widehat{\theta}_{k-1\,k} - \theta_{k-1\,k}) \xrightarrow{d} V_2^* \;\text{ as } n \to \infty \tag{37}$$

where V_2^* has again absolutely continuous distribution. As a consequence, there exists a unique $d_{1-\alpha}$ satisfying the equality

$$P\left(V_2^* \leq d_{1-\alpha}\right) = 1 - \alpha.$$

The same reasoning as in Section 5.2 shows that the approximate relationship

$$1 - \alpha \simeq P\left(\theta_{k-1\,k} \geq \widehat{\theta}_{k-1\,k} - \frac{d_{1-\alpha}}{\sqrt{n}}\ \forall\, k = 1, \ldots, K\right)$$

holds.

The quantile $d_{1-\alpha}$ can be approximated through sub-sampling, with a procedure virtually identical to that of Section 5.2. Let $\widehat{R}_{n,m}(z)$ be the empirical d.f. obtained through sub-sampling, and let $\widehat{d}_p = \widehat{R}_{n,m}^{-1}(p) = \inf\{z : \widehat{R}_{n,m}(z) \geq p\}$ be the corresponding quantile of order p. If $m \to \infty$, $m/n \to 0$ as $n \to \infty$, the following two statements hold.

(i) $\sup_z \left|\widehat{R}_{n,m}(z) - P\left(\max_{1\leq k\leq K} \sqrt{n}(\widehat{\theta}_{k-1\,k} - \theta_{k-1\,k}) \leq z\right)\right| \xrightarrow{p} 0.$

(ii) $\widehat{d}_{1-\alpha} \xrightarrow{p} d_{1-\alpha}$, for each fixed $0 < \alpha < 1$.

Finally, at a significance level α, the test consisting in rejecting H_0 if

$$\widehat{\theta}_{k-1\,k} - \frac{d_{1-\alpha}}{\sqrt{n}} < \frac{1}{2} \text{ for at least one } k = 1, \ldots, K$$

has approximate significance level α.

5.4 Testing for the best group of treatments

The goal of this section is to construct a test for the hypothesis that a certain group of treatment's levels is the best one. In other words, each treatment in the group

dominates all treatments outside of the group. Consider an integer $0 \leq K_0 < K$; the hypotheses problem under consideration is

$$\begin{cases} H_0: & Y_{(j)} \preceq Y_{(k)} \; \forall\, j = 0,\, 1,\, \ldots,\, K_0 \text{ and } \forall k = K_0+1, \ldots,\, K \\ H_1: & Y_{(j)} \not\preceq Y_{(k)} \text{ for some } j = 0,\, 1,\, \ldots,\, K_0 \text{ and } \forall k = K_0+1, \ldots,\, K \end{cases}. \tag{38}$$

The starting point is still based on the monotonicity of θ_{jk}s, that allows to transform (38) into the weaker version

$$\begin{cases} H_0: & \theta_{jk} \geq \frac{1}{2} \qquad \forall\, j = 0,\, 1,\, \ldots,\, K_0 \text{ and } \forall k = K_0+1, \ldots,\, K \\ H_1: & \theta_{jk} < \frac{1}{2} \quad \text{for some } j = 0,\, 1,\, \ldots,\, K_0 \text{ and } \forall k = K_0+1, \ldots,\, K \end{cases}. \tag{39}$$

Again, a test for the hypothesis (39) at a significance level α is constructed by first taking simultaneous confidence intervals for θ_{jk}, $j = 0,\, 1,,\, \ldots,\, K_0$, $k = K_{0+1},\, \ldots,\, K$, with overall level $1-\alpha$. The null hypothesis H_0 is rejected if at least one of the intervals contains $1/2$.

Using the results in Section S.2.3.2 in the Supplementary Materials, it is seen that that

$$\max_{\substack{j \leq K_0 \\ k \geq K_0+1}} \sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk}) \xrightarrow{d} V_3^* \;\text{ as } n \to \infty \tag{40}$$

where the r.v. in V_3^* in (40) possesses absolutely continuous distribution. Hence, there exists a unique $d_{1-\alpha}$ satisfying the equality $P\left(V_3^* \leq d_{1-\alpha}\right) = 1-\alpha$.

Again, using the same reasoning as in Section 5.2, the approximate relationship

$$1 - \alpha \simeq P\left(\theta_{jk} \geq \widehat{\theta}_{jk} - \frac{d_{1-\alpha}}{\sqrt{n}} \; \forall\, j \leq K_0,\; k \geq K_0 + 1\right)$$

is obtained.

The quantile $d_{1-\alpha}$ can be approximated through sub-sampling, with a procedure identical to that of Section 5.2. Let $\widehat{R}_{n,m}(z)$ be the empirical d.f. obtained through sub-sampling, and let $\widehat{d}_p = \widehat{R}^{-1}_{n,m}(p) = \inf\{z : \widehat{R}_{n,m}(z) \geq p\}$ be the corresponding quantile of order p. If $m \to \infty$, $m/n \to 0$ as $n \to \infty$, the following two statements hold.

(i) $\sup_z \left| \widehat{R}_{n,m}(z) - P\left(\max_{1 \leq k \leq K} \sqrt{n}(\widehat{\theta}_{k-1\,k} - \theta_{k-1\,k}) \leq z \right) \right| \xrightarrow{p} 0.$

(ii) $\widehat{d}_{1-\alpha} \xrightarrow{p} d_{1-\alpha}$, for each fixed $0 < \alpha < 1$.

Finally, at a significance level α, the test consisting in rejecting H_0 whenever

$$\widehat{\theta}_{jk} - \frac{d_{1-\alpha}}{\sqrt{n}} < \frac{1}{2} \text{ for at least one } j \leq K_0 \text{ and } k \geq K_0 + 1$$

has approximate significance level α.

6. Simulation results

The goal of the present section is to study, by Monte Carlo simulation, the performance, in terms of significance level and power, of the tests introduced so far, that are compared with other tests proposed in the literature.

6.1 Testing for absence of treatment effects: simulation study 1

In this section, the Kruskal-Wallis type test proposed in Section 4 is compared to that based on matching on generalized propensity score GPSM, in view of its good properties shown in (Yang et al., 2016). Such a test is essentially based on pairwise treatment effects. In order to control the overall significance level, Bonferroni rule has been used.

In order to perform a fair comparison, the simulation setting considered here is essentially similar to that in Yang et al. (2016) for zero (*i.e. absence* of) treatment effects, and in Li and Li (2019) for non-zero treatment effects.

Two simulation designs have been considered, the first one with $K+1 = 3$ treatment levels and the second one with $K + 1 = 6$ treatment effects. For the sake of brevity, details are given here only for $K + 1 = 3$ levels.

- Six covariates pre-treatment $\boldsymbol{X}' = (1, X_1, X_2, X_3, X_4, X_5, X_6)$ have been considered, where:
  - X_1, X_2, X_3 have joint multivariate Normal distribution with with zero means zero, variances 2, 1, 1, and covariances 1, -1, -0.5, respectively;
  - X_4 in independent of X_1, X_2, X_3 with Uniform distribution in $(-3,\ 3)$: $X_4 \sim U(-3, 3)$;
  - X_5 in independent of $X_1, \ldots, X_4$ with Chi square distribution with 1 d.o.f.: $X_5 \sim \chi^2_1$;
  - X_6 in independent of $X_1, \ldots, X_5$ with Bernoulli distribution with parameter 0.5: $X_6 \sim Be(0.5)$.
- In case of *zero* effects, three treatment groups ($K + 1 = 3$) are generated through a Multinomial regression model $Multinom(p_0(\boldsymbol{x}), p_1(\boldsymbol{x}), p_2(\boldsymbol{x}))$ with
  - $p_k(\boldsymbol{x}) = \exp(\boldsymbol{x}' \times \beta_k)/\sum exp(\boldsymbol{x}' \times \beta_{k'})$;
  - $\beta'_0 = (0, 0, 0, 0, 0, 0, 0)$, $\beta'_1 = 0.7\,(0, 1, 1, 1, -1, 1, 1)$, $\beta'_2 = 0.4\,(0, 1, 1, 1, 1, 1, 1)$.
  - Potential outcomes are generated as $Y_{(k)} = \boldsymbol{X}'\gamma_k + \eta$, with $\eta \sim N(0,\ 1)$ and $\gamma'_0 = (-1.5, 1, 1, 1, 1, 1, 1)$, $\gamma'_1 = (-3, 2, 3, 1, 2, 2, 2)$, $\gamma'_2 = (1.5, 3, 1, 2, -1, -1, -1)$.

- In case of *non-zero* effects, three treatment groups ($K+1=3$) are generated through a Multinomial regression model $Multinom(p_0(\boldsymbol{x}), p_1(\boldsymbol{x}), p_2(\boldsymbol{x}))$ with

  - $\beta_1' = k_1 \times (1,1,1,-1,-1,1)$, $\beta_2' = k_2 \times (1,1,1,1,1,1)$, where $(k_1, k_2) = (0.2, 0.1)$ to simulate a scenario with adequate covariate overlap, and $(k_1, k_2) = (0.8, 0.4)$ to simulate lack of overlap with strong propensity tails.

  - Potential outcomes are generated from $Y_{(k)} = \boldsymbol{X}'\gamma_k + \eta$ with $\eta \sim N(0,1)$ and $\gamma_0' = (-1.5,1,1,1,1,1,1)$, $\gamma_1' = (-4,2,3,1,2,2,2)$, $\gamma_2' = (3,3,1,2,-1,-1,-1)$. Clearly, $E[Y_{(0)}] = 0.0$, $E[Y_{(1)}] = -1.0$, $E[Y_{(2)}] = 1.5$. The potential outcomes have been reorganized in ascending order of expected values (-1, 0.0, 1.5), so that $T=2$ is the "best" treatment.

As already said, the case of $K+1=6$ treatment levels has been also considered in simulations, under both zero and non-zero treatment effects. Details can be found in Yang et al. (2016) for zero effects, and in Section D of the Supplementary material in Li and Li (2019) for non-zero effects.

The total sample size is $n = 1000, 1500, 2000$ for $K+1=3$ and $n = 1500, 3000, 6000$ for $K+1=6$.

Table 1 summarizes the rejection probabilities of the null hypothesis for different sample sizes for $K+1=3$ and $K+1=6$. Figures 1 and 2 show the estimated distribution functions of the potential outcomes for $n=1500$ and $n=6000$ (cfr. formula (19)).

Table 1 here

The results show that the Kruskal-Wallis type test is better than the test based on matching $GPSM$, in terms of both actual significance level and power, especially for small sample sizes and lack of overlap.

Figure 1 here

Figure 2 here

As an overall remark, Kruskal-Wallis type test seems to offer good performance in terms of both simplicity and power.

6.2 Testing for stochastic dominance: simulation study 2

The goal of this section is to study, again *via* simulation, the performance of the stochastic dominance tests proposed in Section 5, namely

- test for the "best" treatment - Section 5.2;

- test for the best group of treatments - Section 5.4;

- test for ordering of all treatments effect - Section 5.3;

Since no other tests of stochastic dominance are proposed in the literature in the case of multi-level treatment, no comparison is made.

The size and power of the above tests are compared in two different cases:

- I: zero treatment effect;

- II: non-zero treatment effect.

The latter, in particular, involves a shift alternative. Two values of the number of treatment levels are considered: $K+1=3$ and $K+1=4$. In the case $K+1=3$, the simulation scenarios are shortly described below.

- *Case* $K+1=3$ - zero treatment scenario.

  - A single covariate X is considered, with Bernoulli distribution: $X \sim Be(0.5)$.

- Propensity scores are generated according to the model

$$p_k(x) = \begin{cases} \frac{0.80(1-x)+0.20x}{(0.80(1-x)+0.20x)+(0.50(1-x)+0.50x)+(0.20(1-x)+0.80x)} & k=0 \\ \frac{0.50(1-x)+0.50x}{(0.80(1-x)+0.20x)+(0.50(1-x)+0.50x)+(0.20(1-x)+0.80x)} & k=1 \\ \frac{0.20(1-x)+0.80x}{(0.80(1-x)+0.20x)+(0.50(1-x)+0.50x)+(0.20(1-x)+0.80x)} & k=2 \end{cases} .$$

- Potential outcomes are generated according to the model

$$Y_{(k)} = 70 + 10X + U_k,\ k = 0,\ 1,\ 2 \tag{41}$$

where U_k possesses Uniform distribution $U_k \sim U(-10,\ 10)$, $k = 0,\ 1,\ 2$. Clearly, $\theta_{01} = \theta_{02} = \theta_{12} = 0.5$, $E[Y_{(0)}] = E[Y_{(1)}] = E[Y_{(2)}] = 75.0$. Furthermore we have $E[Y_{(0)}|T = 0] = 72.0$, $E[Y_{(1)}|T = 1] = 75.0$ and $E[Y_{(2)}|T = 2] = 78.0$. This is clearly due to the confounding effect of X, and makes it difficult to detect the absence of treatment effect.

- *Case* $K + 1 = 3$ - non-zero treatment scenario.

  - A single covariate X is considered, with Bernoulli distribution: $X \sim Be(0.5)$.

  - Propensity scores are generated according to the model

$$p_k(x) = \begin{cases} \frac{0.45(1-x)+0.55x}{(0.45(1-x)+0.55x)+(0.50(1-x)+0.50x)+(0.55(1-x)+0.45x)} & k=0 \\ \frac{0.50(1-x)+0.50x}{(0.75(1-x)+0.25x)+(0.50(1-x)+0.50x)+(0.25(1-x)+0.75x)} & k=1 \\ \frac{0.55(1-x)+0.45x}{(0.45(1-x)+0.55x)+(0.50(1-x)+0.50x)+(0.55(1-x)+0.45x)} & k=2 \end{cases} .$$

- Potential outcomes are generated according to the model

$$\begin{aligned} Y_{(0)} &= 74 + 20X + U_0 \\ Y_{(1)} &= 75 + 20X + U_1 \\ Y_{(2)} &= 76 + 20X + U_2 \end{aligned}$$

where $U_k \sim U(-10,\ 10)$, $k = 0,\ 1,\ 2$. It is not difficult to see that $\theta_{01} = 0.52$, $\theta_{02} = 0.55$, $\theta_{12} = 0.52$, $E[Y_{(0)}] = 84.0$, $E[Y_{(1)}] = 85.0$, $E[Y_{(2)}] = 86.0$. Furthermore we have $E[Y_{(0)}|T = 0] = E[Y_{(1)}|T = 1] = E[Y_{(2)}|T = 2] = 85.0$. This is clearly due to the confounding effect of X, and makes it difficult to detect the presence of treatment effect.

A summary of Scenarios I, II is provided in Table 2.

Table 2 here

The exact distribution functions of potential outcomes for both scenarios are reported in Section S3 of Supplementary Material.

The case $K + 1 = 4$ is similar. A summary is reported in Table 3.

Table 3 here

Samples sizes $n = 500, 1000, 1500$ for $K + 1 = 3$ and $n = 500, 1000, 2000$ for $K + 1 = 4$ have been used in Monte Carlo simulations, with $N = 1000$ replications per sample size. As far as subsample approximation is concerned, $M = 1000$ subsamples of size $m = n^{0.8}$ have been drawn by simple random sampling from each of the $N = 1000$ original samples.

Confidence intervals of the test procedures of Sections 5.2, 5.4, 5.3 have been studied under Scenarios I, II, with $Y_{(2)}$ as the "best" treatment. Results are reported in Tables

4, 5 below.

Table 4 here

Table 5 here

All confidence intervals exhibit coverage probabilities close to the nominal values, although this is less evident for the test of ordering of all treatment effects when $n \leq 1000$. The tests for "best" treatment and for the best group of treatments show the same results as the potential outcomes are ordered with respect to stochastic dominance, and in the test for the best group of treatments the groups (T=0, T=1) and (T=2) are considered.

For the sake of completeness, the Kruskal-Wallis type test and the test based on matching *GPSM* have been compared also for this second simulation study. Results are in Section S4 of the Supplementary Material.

7. Empirical study

7.1 Causal Understanding of Fake News Dissemination on Social Media

In the framework of propagation-based methods for fake news detection, Cheng et al. (2021) consider a bipartite social network between users on one side, and fake news on the other side. A user u is connected to fake news i if the user spreads the fake news. Attributes are associated to users. The outcome is the user susceptibility $B_u \in (0, 1]$ which is formally defined as $B_u = n^u_{fake}/(n^u_{fake} + n^u_{true})$, where n^u_{fake} is the number of fake news user u has shared, under the assumption that the larger the portion of fake news a user has shared, the more susceptible the user to share fake news. In order to understand which attributes potentially cause users to share fake news, Cheng et al. (2021) try to identify confounders, *i.e.* variables that cause spurious associations between treatments (e.g., user attributes) and outcome (e.g., user susceptibility). In fake

news dissemination, confounders can be characterized by fake news sharing behavior that inherently relates to user attributes and online activities. Learning such user behavior is subject to selection bias in users who are susceptible to share news on social media.

The benchmark dataset for fake news detection is *GossipCop* (Shu et al., 2018). In *GossipCop*, entertainment stories were collected from various media outlets. The fact-checking evaluation results came from the rating scores on the *GossipCop* website. Ratings range from 0 to 10, with 0 indicating fake and 10 real. The dataset consists of 16.817 real stories and 5.323 fake stories. Table 6 presents the descriptive statistics of the variables (attributes).

Table 6 here

A three lever treatment has been considered, namely

- $T = 0.$ corresponding to unverified users;

- $T = 1$, corresponding to verified users with register time smaller than the median;

- $T = 2$, corresponding to verified users with register time higher than the median.

As confounders, beside *status count, favourites count*, *followers count*, *friend count*, the two dimensional embedding (x_0, x_1) of fake news spreading behaviour is considered, as motivated in Cheng et al. (2021). A number of methods for embedding a graph into a metric space is proposed in the literature; cfr. Grover and Leskovec (2016); Khosla et al. (2019); Mikolov et al. (2013); Perozzi et al. (2014). The embedding model we have used here is based on a well known Deep-learning technique: the *Skip-Gram* model by Mikolov et al. (2013), originally developed in the field of Natural Language Processing (NLP). In particular, we have used the implementation of the model made available by the `gensim` Python library. The input of the model is represented by a collection of

sequences of adjacent nodes collected by performing random walks on the graph. The basic idea is that, just like the co-occurrence of words in sentences in the NLP context can be used to try to quantify the semantic similarity between words in a language, the co-occurrence of nodes in random walks is used to try to quantify structural similarities between the nodes of the network. We stress that, because the only input of the model is a collection of sequences of random walks on the network, the embedding produced by the model is only dependent on the topology of the graph, encoded in its adjacency matrix. As we are interested in a characterization of the properties of the sharing users, we first deduce a monopartite graph from the original bipartite graph, where each node represents one fake news sharing user. The process of collecting random walks for the training of the model is performed over this monopartite graph. In the present case, this is obtained *via* a simple projection of the bipartite graph on the users layer, by which each couple of users (u_i, u_j) is connected by an undirected weighted link, whose weight is equal to the number of common fake-news shared by both users. If the new monopartite weighted undirected graph is denoted as as G, the random walks collection process is implemented as follows.

1. Fix the length of the random walk L.

2. Choose one starting node $u_i \in V(G)$.

3. Randomly choose one neighbor node u_j of the current node u_i according to the probability distribution $P_{u_i}(u_j) = \frac{w_{ij}}{w_i}$, where w_{ij} is the weight of the link connecting u_i with u_j and w_i is the total weight of all links attached to u_i.

4. Go to step 3, and repeat until the length of the walk is equal to L.

The random walks are collected by starting a single walk from each node of the network taken in some arbitrary order, the nodes' order is then randomly reshuffled

and walks are started again from each node. This procedure is repeated γ times. The number γ is an hyperparameter of the model that fixes the number of walks per node and consequently the size of the ensemble used for training the model, which at the end of the collection procedure is composed of $N \cdot \gamma$ walks, N being the number of nodes in the network. Once the ensemble of random walks is appropriately collected, it is then fed as the input into the Skip-Gram model, so that a 2-dimensional embedding of the graph's nodes can be produced as an output.

The estimated θ_{kl} are $\widehat{\theta}_{01} = 0.57$, $\widehat{\theta}_{02} = 0.58$, $\widehat{\theta}_{12} = 0.51$. The Kruskal-Wallis type test rejects the null hypothesis of zero treatment effect at nominal significance level 0.05. Estimated pairwise ATE_{kl} are $ATE_{01} = 0.10$ $ATE_{02} = 0.15$ $ATE_{12} = 0.05$.

8. Conclusions

In this paper, new techniques for assessing causal treatment effects using observational data in scenarios involving more than two treatment options are proposed. The problem of comparing the effects of a treatment with several different levels is less studied in the literature than the comparison of a treatment vs. a control. The "usual" approach seems to consist in separately focusing on pairwise ATE comparisons. Combination of pairwise comparison, for instance through the Bonferroni rule, is not considered.

In this article, a test for the presence of a treatment effect in case of a treatment with multiple levels is proposed, based on a suitable extension of the Kruskal-Wallis statistic. Tests for stochastic dominance of treatment effects are also considered. Two simulation studies and an application illustrate the advantages of the method.

Supplementary Material

- Estimation of generalized propensity scores.

- Proofs of main results

- Simulation Study 2: exact distribution of potential outcomes.

- Simulation Study 2: Comparison of Kruskal-Wallis type test and matching *GPSM* test.

Acknowledgments

The research of Livia De Giovanni and Ayoub Mounim is supported by the IDMO program of the Italian Digital Media Observatory (project 101158697 of the European Digital Executive Agency HADEA Grant Agreement).

Tables and Figures

Table 1: Simulation study 1 - rejection probabilities (nominal significance level 0.05), $m = n^{0.8}$

Scenario - $K+1=3$	n=1000	n=1500	n=2000
Kruskal-Wallis			
H_0 true	0.10	0.07	0.05
H_1 true - adequate overlap	1.00	1.00	1.00
H_1 true - lack of overlap	0.90	0.95	0.98
GPSM			
H_0 true	0.16	0.12	0.11
H_1 true - adequate overlap	1.00	1.00	1.00
H_1 true - lack of overlap	0.83	0.90	0.95
Scenario - $K+1=6$	n=1500	n=3000	n=6000
Kruskal-Wallis			
H_0 true	0.22	0.20	0.15
H_1 true - adequate overlap	1.00	1.00	1.00
H_1 true - lack of overlap	0.91	0.93	0.96
GPSM			
H_0 true	0.24	0.20	0.16
H_1 true - adequate overlap	0.98	1.00	1.00
H_1 true - lack of overlap	0.80	0.84	0.88

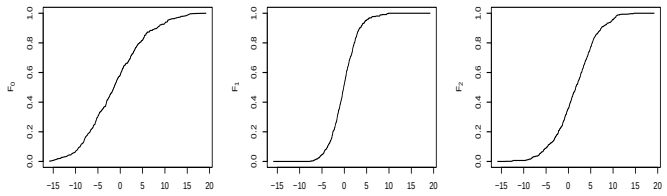

Figure 1: Estimated distribution functions of the potential outcomes, $K+1=3$, H_1 - overlap, $n=1500$

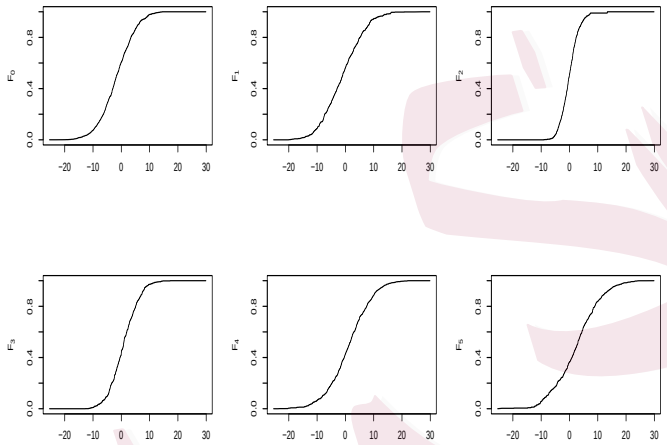

Figure 2: Estimated average treatment effects of of the potential outcomes, $K+1=6$, H_1 - overlap, $n=6000$

Table 2: Simulation Study 2, scenarios I, II - $K+1$=3

Scenario	θ_{ij}	$E[Y_{(j)}]$	$E[Y_{(j)}\|T]$
I	$\theta_{01}=0.50$	$E[Y_{(0)}]=75.0$	$E[Y_{(0)}\|T=0]=72.0$
	$\theta_{02}=0.50$	$E[Y_{(1)}]=75.0$	$E[Y_{(1)}\|T=1]=75.0$
	$\theta_{12}=0.50$	$E[Y_{(2)}]=75.0$	$E[Y_{(2)}\|T=2]=78.0$
II	$\theta_{01}=0.52$	$E[Y_{(0)}]=84.0$	$E[Y_{(0)}\|T=0]=85$
	$\theta_{02}=0.55$	$E[Y_{(1)}]=85.0$	$E[Y_{(1)}\|T=1]=85$
	$\theta_{12}=0.52$	$E[Y_{(2)}]=86.0$	$E[Y_{(2)}\|T=2]=85$

Table 3: Simulation Study 2, scenarios I, II - $K+1$=4

Scenario	$E[Y_{(j)}]$	$E[Y_{(j)}\|T]$
I	$E[Y_{(0)}]=80$	$E[Y_{(0)}\|T=0]=74.0$
	$E[Y_{(1)}]=80$	$E[Y_{(1)}\|T=1]=78.0$
	$E[Y_{(2)}]=80$	$E[Y_{(2)}\|T=2]=82.0$
	$E[Y_{(3)}]=80$	$E[Y_{(3)}\|T=2]=86.0$
II	$E[Y_{(0)}]=94$	$E[Y_{(0)}\|T=0]=97.5$
	$E[Y_{(1)}]=97$	$E[Y_{(1)}\|T=1]=97.5$
	$E[Y_{(2)}]=99$	$E[Y_{(2)}\|T=2]=97.5$
	$E[Y_{(3)}]=101$	$E[Y_{(3)}\|T=2]=97.5$

Table 4: Coverage probabilities (nominal significance level 0.05) - $K+1$=3

	Testing stochastic dominance	$n=500$	$n=1000$	$n=1500$
I	*best treatment*	98.0	99.0	100.0
	best group treatments	98.0	99.0	100.0
	Ordering of all treatments	100.0	100.0	100.0
II	*best treatment*	78.0	82.0	92.0
	best group treatments	78.0	82.0	92.0
	ordering of all treatments	72.0	76.0	87.0

Table 5: Coverage probabilities (nominal significance level 0.05) - $K+1$=4

	Testing stochastic dominance	$n=500$	$n=1000$	$n=2000$
I	*best treatment*	99.0	100.0	100.0
	best group treatments	99.0	100.0	100.0
	Ordering of all treatments	100.0	100.0	100.0
II	*best treatment*	95.0	100.0	100.0
	best group treatments	95.0	100.0	100.0
	Ordering of all treatments	81.0	86.0	90.0

Table 6: Descriptive statistics *Gossipcop*

	1*st quartile*	*Median*	3*rd quartile*	*Mean*
Outcome	1.00	1.00	1.00	0.95
Verified	0	0	0	0.015
x_0	-0.38	-0.25	0.09	0.00
x_1	-1.77	-0.43	-0.24	0.00
Register time	1520	2522	3320	2365
Status count	2809	9585	29702	33003
Favourites count	85	1344	8927	10933
Followers count	81	401	1662	10853
Friend count	145	552	1952	2200

References

Abadie, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American Statistical Association 97*, 284–292.

Anderson, G. (1996). Nonparametric Tests of Stochastic Dominance in Income Distribution. *Econometrica 64*, 1183–1193.

Cheng, L., R. Guo, K. Shu, and H. Liu (2021). Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM.

Conti, P. L. and L. De Giovanni (2022). Testing for the presence of treatment effect under selection on observables. *AStA Advances in Statistical Analysis*, DOI https://doi.org/10.1007/s10182–022–00454–8.

Crump, R., V. J. Hotz, G. W. Imbens, and O. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*, 187–199.

Ding, P. (2017). A Paradox from Randomization-Based Causal Inference. *Statistical Science 32*, 331–345.

Donald, S. G. and Y. C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics 178*, 383–397.

Donald, S. G. and Y. C. Hsu (2016). Improving the power of tests of stochastic dominance. *Econometric Reviews 35*, 553–585.

Firpo, S. (2007). Efficient Semiparametric Estimation of Quantile Treatment Effect. *Econometrica 75*, 259–276.

Grover, A. and J. Leskovec (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 855–864. Association for Computing Machinery.

Hettmansperger, T. P. and J. W. McKean (2011). *Robust Nonparametric Statistical Methods* II Ed. Boca Raton: CRC Press.

Khosla, M., J. Leonhardt, W. Nejdl, and A. Anand (2019). Node representation learning for directed graphs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 395–411. Springer.

Kim, K. (2013). An Alternative Efficient Estimation of Average Treatment Effects. *Journal of Market Economy 42*, 1–41.

Li, F. and F. Li (2019). Propensity Score Weighting for Causal Inference with Multiple Treatments. *The Annals of Applied Statistics 13*(4), pp. 2389–2415.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113*, 390–400.

Li, F., L. E. Thomas, and F. Li (2019). Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology 188*, 250–257.

Lopez, M. J. and R. Gutman (2017). Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas. *Statistical Science 32*(3), 432 – 454.

McFadden, D. (1989). Testing for stochastic dominance. In *Studies in the economics of uncertainty: In honor of Josef Hadar*, pp. 113–134. Springer.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Perozzi, B., R. Al-Rfou, and S. Skiena (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, New York, NY, USA, pp. 701–710. Association for Computing Machinery.

Politis, D. N. and J. P. Romano (1994). Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *The Annals of Statistics 22*, 2031–2050.

Shu, K., D. Mahudeswaran, S. Wang, D. Lee, and H. Liu (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint*

arXiv:1809.01286.

Stürmer, T., K. J. Rothman, and J. A. R. J. Glynn (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American Journal of Epidemiology 172*, 843–854.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wu, J. and P. Ding (2018). Randomization tests for weak null hypotheses in randomized experiments. arXiv:1809.07419 [stat.ME].

Yang, S., G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics 72*(4), 1055–1065.