Journal of Multivariate Analysis 132 (2014) 39–57

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A permutation approach for ranking of multivariate populations

Rosa Arboretti [a], Stefano Bonnini [b], Livio Corain [c,*], Luigi Salmaso [c]

[a] Department of Land, Environment, Agriculture and Forestry, University of Padova, Italy
[b] Department of Economics and Management, University of Ferrara, Italy
[c] Department of Management and Engineering, University of Padova, Italy

ARTICLE INFO

Article history:
Received 13 October 2013
Available online 1 August 2014

AMS subject classifications:
62G09

Keywords:
Permutation tests
Multivariate tests
Nonparametric combination
NPC tests

ABSTRACT

The need to establish the relative superiority of each treatment/group when compared to all the others, that is ordering the effects with respect to the underlying populations, often occurs in many multivariate studies especially in the bio-medical field. Within the framework of multivariate stochastic ordering, the purpose of this work is to propose a nonparametric permutation-based solution for the problem of ranking of multivariate populations, i.e. estimating an ordering related to the possible stochastic dominance among several unknown multivariate distributions. The method is metric-free in the sense that it can be applied to any kind of response variables, i.e. continuous/binary or ordered categorical or mixed (some continuous/binary univariate components and some other ordered categorical), and it is valid also in case the sample sizes are lower than the number of responses. It will be theoretically argued and numerically proved that our method controls the risk of false ranking classification under the hypothesis of population homogeneity while under the alternatives we expect that the true rank can be estimated with satisfactory accuracy, especially for the 'best' populations. Finally, an application to a morphological analysis of primary bovine cerebellum cell cultures is proposed to highlight the practical relevance of the proposed methodology.

1. Introduction and motivation

The necessity and usefulness of defining an appropriate ranking of several populations of interest, e.g. diseases, dosages of a treatment, processes, products/services, is very common within many areas of applied research such as Life Sciences, Pharmacology, Engineering, etc. The idea of ranking in fact occurs more or less explicitly any time when in a study the goal is to determine an ordering among several input conditions/treatments with respect to one or more outputs of interest when there might be a "natural ordering". We remark that the "natural ordering" should be referred to the way the response is interpreted and not to any kind of 'a priori' knowledge on ordering of populations which is not assumed at all within the framework of our proposed methodology. This happens very often in the context of bio-medical problems where population elements can be patients, cell cultures, tissue samples, etc. and the conditions/treatments to be ranked are for example diagnosis groups or different levels of exposure/dosage which are put in relation with some suitable bio-medical endpoints such as survival data, gene expression or proteomic data. Some inferential techniques such as multiple

* Corresponding author.
E-mail address: livio.corain@unipd.it (L. Corain).

http://dx.doi.org/10.1016/j.jmva.2014.07.009
0047-259X/

comparison procedures [41], ranking and selection [21], order restricted inference [40] and ranking models [24], more or less directly or indirectly partially address the issue of population ranking but only under some additional assumptions and in specific situations.

Several times the populations of interest are multivariate in nature, and when the underlying population distributions are not specified we are actually considering the ranking problem from a nonparametric point of view. Similarly to what has been proposed by several authors within the nonparametric ranking and selection framework [20], in this paper we consider a functional of the empirical distribution function $\boldsymbol{F}$ of the population distribution, specifically a combination of the univariate directional permutation p-values which can be viewed as a non-metric "distance measure" among multivariate distributions. Therefore, the combination methodology [32] is a useful tool since it allows us to reduce the dimensionality of the multivariate problem in order to compare and rank the populations under investigation. Given two multivariate random variables $\mathbf{Y}_j$ and $\mathbf{Y}_h$, if $\mathbf{Y}_j$ dominates $\mathbf{Y}_h$ then the significance level function related to the combined test statistic T''_ψ suitable for testing the null hypothesis of equality in distribution against the alternative $\mathbf{Y}_j >^d \mathbf{Y}_h$ is stochastically larger under the alternative that under the null hypothesis of equality in distribution. This idea comes from Tukey's underlining representation of pairwise comparison results [26] and since then, the problem of ranking has been addressed with respect to many different points of view. In the next section we review some procedures proposed in the literature, classifying them within the main reference field where they have been developed. Sections 3 and 4 are devoted to the formalization of the ranking problem, the presentation of the proposed solution along with the theoretical properties of the ranking estimator, and the construction steps to solve the multivariate ranking problem. A simulation study for small sample sizes and ordered categorical responses is then presented in Section 5 while in Section 6 we illustrate an application to morphological analysis of cell cultures. Finally, Section 7 draws conclusions, final remarks and future perspectives.

2. Literature review

There are many situations where we are facing with inferential problems of comparing several – more than two – populations and the goal is not just to accept or reject the so-called homogeneity hypothesis, i.e. the equality of all populations, but an effort is provided to try to rank the populations according to some suitable criterion. Multiple Comparison Procedures—MCPs occurs when one considers a set of statistical inferences simultaneously e.g. when a set, or family, of testing procedures is considered simultaneously, in particular when we wish to compare more than two populations (treatments, groups, etc.) in order to find out possible significant differences among them within the C-samples location testing problem [41]. Since incorrect rejection of the null hypothesis is more likely when the family as a whole is considered, the main issue and goal of MCPs is to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. Some contributions proposed in the field of MCPs have more or less to do with the ranking problem. Hsu and Peruggia [26] critically reviewed the graphical representations of Tukey's multiple comparison method behind which we can clearly see the Tukey's attempt to rank the populations from the 'best' to the 'worst'. The popular Tukey's underlining representation prescribes that after ordering the populations according to the increasing values of their estimated means, all subgroups of populations that cannot be declared different are emphasized by a common line segment. After that, one can infer at least as many groups are strictly not the best and in this way arguing which population can be overall considered as the first, the second, etc. Since the set of all pairwise orderings is equivalent to a set of rankings, from pairwise significant differences and from the specific direction in which each significance occurs, it is possible to specify the subset of rankings selected from the set of all possible rankings (for details see [7,25]). Referring to the so global performance indexes and with the goal of ordering several multivariate populations, Arboretti Giancristofaro et al. [1] proposed a permutation-based method using simultaneous pairwise confidence intervals. In this connection, Arboretti Giancristofaro et al. [2] compared two ranking parameters in a simulation study that highlighted some differences between a parametric and a nonparametric approach.

The selection and ranking approach, also known as multiple decision procedures, arose from the need of answering natural questions regarding the selection of the 'best' population within the framework of a C-sample testing problem [21]. A few selection and ranking proposals are concerned with ranking of several multivariate populations: under assumption of multivariate normal distributions, several real-valued functions of population parameters have been adopted to rank the populations [21]. In case the population distribution functions are not specified, some nonparametric solutions have been proposed. Those procedures are based on general ranking parameters such as the rank correlation coefficient and the probability of concordance [20].

Prior information regarding a statistical model frequently constrains the shape of the parameter set and can often be quantified by placing inequality constraints on the parameters. The use of such ordering information increases the efficiency of procedures developed for statistical inference [13]. On the one hand, such constraints make the statistical inference procedures more complicated, but on the other hand, such constraints contain statistical information as well, so that if properly incorporated they would be more efficient than their counterparts wherein such constraints are ignored [40]. Davidov and Peddada [10] extended the order restricted inference paradigm to the case of multivariate binary response data under two or more naturally ordered experimental conditions.

The ranking problem has been addressed in the literature also from the point of view of investigating and modeling the variability of sampling statistics used to rank populations, that is the empirical estimators whose rank transformation