A Nonparametric Multivariate Scatter-Based Ranking Method with Applications to Biomedical Research and Industrial Quality Management

Stefano Bonnini*
Department of Economics and Management, University of Ferrara, Ferrara, Italy – bnnsfn@unife.it

Rosa Arboretti
Department of Land, Environment, Agriculture and Forestry, Legnaro (PD), Italy – rosa.arboretti@unipd.it

Livio Corain
Department of Management and Engineering, University of Padova, Vicenza, Italy – livio.corain@unipd.it

Bruno Cozzi
Department of Comparative Biomedicine and Food Science, University of Padova, Legnaro (PD), Italy – bruno.cozzi@unipd.it

Stefano Montelli
Department of Comparative Biomedicine and Food Science, University of Padova, Legnaro (PD), Italy – stefano.montelli@unipd.it

Antonella Peruffo
Department of Comparative Biomedicine and Food Science, University of Padova, Legnaro (PD), Italy – antonella.peruffo@unipd.it

Luigi Salmaso
Department of Management and Engineering, University of Padova, Vicenza, Italy – luigi.salmaso@unipd.it

Abstract

As an extension of the multivariate location-based ranking approach proposed by Arboretti et al. (2014), we present in this work a novel nonparametric and permutation-based method for ranking of multivariate populations concerning with the scatter aspect. Besides the methodological novelty of the approach, it has a practical relevance given that there are many real problems where the need of ranking several multivariate treatments/conditions/etc. regarding an overall variability criterion is the natural goal. Finally, two real case studies in the fields of biomedical research and industrial quality management are introduced, i.e. a cytoarchitectonic study of the cerebral cortex and a search for the best storing condition in the leather industry.

Keywords: multivariate ranking problem; nonparametric combination; permutation tests.

1. Introduction

The need 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 a post-hoc ordering among several input conditions/treatments with respect to one or more outputs of interest when there might be a "natural ordering". From a location-oriented point of view, 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.
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. Following the multivariate location-based ranking approach proposed by Arboretti et al. (2014), we present a novel nonparametric and permutation-based method for ranking of multivariate populations with respects to the scatter aspect. In this work we consider a scatter-oriented functional of the empirical distribution function $\boldsymbol{F}$ of the population distribution, specifically a combination of the univariate directional permutation *p*-values using the squares of the original observed values which can be viewed as a scatter non-metric "distance measure" among multivariate distributions. Therefore, the combination methodology (Pesarin and Salmaso, 2010) 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 $\mathbf{Y}_h$ from the dispersion point of view 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 $\mathbf{Y}_j <^d_{scatter} \mathbf{Y}_h$ is stochastically larger under the alternative that under the null hypothesis of equality in distribution.

2. Formalization of the problem and permutation solution

Let us assume that data are drawn from *C* multivariate populations $\Pi_1,..., \Pi_C$ (i.e. items/groups/treatments), $C>2$, by means of a sampling procedure, so that to make inference on their possible equality and in case of rejection of this hypothesis to classify those populations in order to obtain two relative rankings from the 'best' to the 'worst' according to two pre-specified meaningful criteria, one related to the location aspect and the other to the scatter aspect. We use the term *relative ranking* because we want to underline that it is not an absolute ranking but a kind of ordering that is only refereed to the *C* populations at hand.
With reference to the so-called one-way MANOVA layout, let us formalize the problem within a nonparametric framework: let $\mathbf{Y}$ be the continuous *p*-dimensional response variable represented as a *p*-vector of the observed data from population Π and let us assume, without loss of generality, that large values of each univariate aspect *Y* correspond to a better marginal location-performance, so that when comparing two populations the possible marginal stochastic dominance of one population over the other should result in a high ranking position, in other words, we are assuming the location criterion “the larger the better”. As usual in the most of real applications regarding the scatter-aspect, we assume also without loss of generality that more scattered values of each univariate aspect *Y* correspond to a worst marginal scatter-performance, so that when comparing two populations the possible marginal stochastic dispersion dominance of one population over the other should result in a low ranking position, in other words, we are assuming the dispersion criterion “the lower the better”.
We recall that our goal is to classify and ranking $\Pi_1,..., \Pi_C$ multivariate populations with respect to *p* marginal variables when *C* samples $\mathbf{Y}_1,..,\mathbf{Y}_C$ are drawn from *C* populations, where n_j is the number of observations, $j=1,...,C$. We are looking for two estimates ${}_{location}\hat{r}(\Pi_j)$ and ${}_{scatter}\hat{r}(\Pi_j)$ of the rank ${}_{location}r(\Pi_j)$ and ${}_{scatter}r(\Pi_j)$, i.e. the relative location and scatter stochastic orderings of each population when compared among all other populations, i.e. more formally

$${}_{location}r_j = {}_{location}r(\Pi_j) = 1 + \sum_{j\neq h} I(\mathbf{Y}_j <^d \mathbf{Y}_h) = 1 + \{\#\ \mathbf{Y}_j <^d \mathbf{Y}_h,\ h=1,...,C,\ j\neq h\},\ j=1,...,C, \quad (1)$$

$${}_{scatter}r_j = {}_{scatter}r(\Pi_j) = 1 + \sum_{j\neq h} I(\mathbf{Y}_j \overset{d}{>}_{scatter} \mathbf{Y}_h) = 1 + \{\#\ \mathbf{Y}_j \overset{d}{>}_{scatter} \mathbf{Y}_h,\ h=1,...,C,\ j\neq h\},\ j=1,...,C, \quad (2)$$

where $I(\cdot)$ is the indicator function and # means the number of times (see also Gupta and Panchepakesan, 2002 and Hall and Miller, 2009). Note that definitions (1) and (2) are derived by using the concept of stochastic dominance and pairwise counting how many populations are stochastically larger (for location) and smaller (for scatter) than that a specific population.
Let us consider an alternative definition of population ranking,

$$_{location}r_j = 1 + \{\#\,(C - \textstyle\sum_{j\neq h} I(\mathbf{Y}_j \overset{d}{>} \mathbf{Y}_h)) > (C - \textstyle\sum_{j'\neq h} I(\mathbf{Y}_{j'} \overset{d}{>} \mathbf{Y}_h)),\ j'=1,...,C,\ j'\neq j\},\ j=1,...,C. \quad (3)$$

$$_{scatter}r_j = 1 + \{\#\,(C - \textstyle\sum_{j\neq h} I(\mathbf{Y}_j \underset{scatter}{\overset{d}{<}} \mathbf{Y}_h)) > (C - \textstyle\sum_{j'\neq h} I(\mathbf{Y}_j \underset{scatter}{\overset{d}{<}} \mathbf{Y}_h)),\ j'=1,...,C,\ j'\neq j\},\ j=1,...,C. \quad (4)$$

Definitions (3) and (4) are derived by using the concept of stochastic dominance and simply pairwise counting how many populations are stochastically smaller (for location) and larger (for scatter) than a given population. Note that both definitions do provide exactly the same ranking, i.e. (3)-(4) are upward ranking procedure from the worst to the best and (1)-(2) from best to worst in a downward fashion. This is because starting from either the first or last position and then moving to the lower or higher positions respectively, necessarily provide the same ordering (for more details see Arboretti et al., 2014).

It is worth underlying that in definition (1) and (2) no a priori knowledge or assumption on the true ordering is considered at all since r_j is simply obtained by counting how many populations are stochastically location and/or scatter larger or smaller than the j-th population. Accordingly, for the existence of the multivariate ranking $r=\{r_1,...r_j....,r_C\}$ we need that inequalities $\mathbf{Y}_j >^d \mathbf{Y}_h$, $\mathbf{Y}_j <^d \mathbf{Y}_h$ and $\mathbf{Y}_j \underset{scatter}{\overset{d}{<}} \mathbf{Y}_h$, $\mathbf{Y}_j \underset{scatter}{\overset{d}{>}} \mathbf{Y}_h$ are consistently defined. To this end, it is worth noting that in the literature there are several formal definitions of multivariate stochastic ordering which are usually extensions from the univariate concepts of location-based stochastic dominance and stochastic dispersion ordering (Shaked and Shanthikumar, 2007).

Under the hypothesis of distributional equality of the C populations, all true ranks would necessarily be equal to one, hence they would be in a full ex-aequo situation, that is

$$_{location}\mathrm{r}(\Pi_j|H_0) = \{1+\#\,\mathbf{Y}_j \overset{d}{<} \mathbf{Y}_h,\ h=1,...,C,\ j\neq h\} = {}_{scatter}\mathrm{r}(\Pi_j|H_0) = \{1+\#\,\mathbf{Y}_j \underset{scatter}{\overset{d}{>}} \mathbf{Y}_h,\ h=1,...,C,\ j\neq h\} = 1,\ \forall\, j.$$

This situation of equal ranking where all populations belong to just one ranking class may be formally represented in a hypotheses testing framework where the hypotheses of interest are:

$$\begin{cases} H_0 : \mathbf{Y}_1 \overset{d}{=} \mathbf{Y}_2 \overset{d}{=} ... \overset{d}{=} \mathbf{Y}_C \\ H_1 : \exists \mathbf{Y}_j \overset{d}{\neq} \mathbf{Y}_h,\ j,h = 1,...,C,\ j \neq h \end{cases}. \quad (5)$$

In case of rejection of the global multivariate hypothesis H_0, that is when data are evidence that at least one population behaves differently from the others, it is of interest to perform inferences on pairwise comparisons between populations, i.e.

$$\begin{cases} H_{0(jh)} : \mathbf{Y}_j \overset{d}{=} \mathbf{Y}_h \\ H_{1(jh)} : \mathbf{Y}_j \overset{d}{\neq} \mathbf{Y}_h,\ j,h = 1,...,C,\ j \neq h \end{cases}. \quad (6)$$

Note that a rejection of at least one hypothesis $H_{0(jh)}$ implies that we are not in an equal ranking situation, that is at least one multivariate population has a greater ranking position than some other, more formally

$$\exists\ {}_{location}\mathrm{r}(\Pi_j) \neq {}_{location}\mathrm{r}(\Pi_h) \text{ and/or } \exists\ {}_{scatter}\mathrm{r}(\Pi_j) \neq {}_{scatter}\mathrm{r}(\Pi_h),\ \ j,h=1,...,C,\ j\neq h.$$

As usual in the framework of C-sample inference, the rejection of the global null hypothesis is not informative on the specific alternative has caused the rejection so that post-hoc analysis is needed to look for which alternative is more likely. In order to make inference on which marginal variable(s) that inequality is mostly due to, it is useful considering inferences on univariate pairwise comparisons between populations, defined as:

$$\begin{cases} H_{0k(jh)} : Y_{jk} \overset{d}{=} Y_{hk} \\ H_{1k(jh)} : \left(Y_{jk} \overset{d}{<} Y_{hk} \right) \bigcup \left(Y_{jk} \overset{d}{>} Y_{hk} \right) \cup \left(Y_{jk} \underset{scatter}{\overset{d}{<}} Y_{hk} \right) \bigcup \left(Y_{jk} \underset{scatter}{\overset{d}{>}} Y_{hk} \right), \end{cases} \quad (7)$$
$$j,h = 1,...,C, j \neq h, k = 1,.., p.$$

because when $Y_{jk} \overset{d}{\neq} Y_{hk}$ is true, then one and only one between $Y_{jk} \overset{d}{<} Y_{hk}$ and $Y_{jk} \overset{d}{>} Y_{hk}$ and $Y_{jk} <_{scatter}^{d} Y_{hk}$ and $Y_{jk} >_{scatter}^{d} Y_{hk}$ is true, i.e. they cannot be jointly true.

Looking at the univariate alternative hypothesis $H_{1k(jh)}$, we are mostly interested in deciding whether a population is either location and/or scatter greater or smaller than another one (not only establishing if they are different). Hence, we can take into account separately of the directional type alternatives, namely those that are suitable for testing both one-sided alternatives (see Pesarin and Salmaso, 2010, p. 163; Bertoluzzo et. all, 2013). Then expression (7) can be reformulated as

$$\begin{cases} H_{0(jh)} : \mathbf{Y}_j \overset{d}{=} \mathbf{Y}_h \\ H_{1(jh)} : \left(\mathbf{Y}_j \overset{d}{<} \mathbf{Y}_h \right) \bigcup \left(\mathbf{Y}_j \overset{d}{>} \mathbf{Y}_h \right) \cup \left(\mathbf{Y}_j \underset{scatter}{\overset{d}{<}} \mathbf{Y}_h \right) \bigcup \left(\mathbf{Y}_j \underset{scatter}{\overset{d}{>}} \mathbf{Y}_h \right), j,h = 1,...,C, j \neq h \end{cases} . \quad (7bis)$$

Let us focus on the scatter ranking aspect and let $p^{+}_{k(j,h)}$ and $p^{-}_{k(j,h)}$ be the marginal permutation-based directional *p*-value statistics related to the scatter stochastic inferiority or superiority alternatives $H^{+}_{1k(jh)} : Y_{jk} \underset{scatter}{\overset{d}{>}} Y_{hk}$ and $H^{-}_{1k(jh)} : Y_{jk} \underset{scatter}{\overset{d}{<}} Y_{hk}$, respectively (for more details on the permutation-based scatter inference see the references to the so-called multi-aspect problem in Pesarin and Salmaso, 2010). Since by definition $p^{+}_{k(jh)} = 1 - p^{-}_{k(j,h)} = p^{-}_{k(h,j)}$, note that all one-sided inferential results related to the hypotheses (5) can be represented as follows:

$$P^{+} = \begin{bmatrix} - & p^{+}_{1(1,2)} & p^{+}_{1(1,3)} & \cdots & p^{+}_{1(1,C)} \\ p^{+}_{1(2,1)} & - & p^{+}_{1(2,3)} & \cdots & p^{+}_{1(2,C)} \\ \cdots & \cdots & - & \cdots & \cdots \\ p^{+}_{1(C-1,1)} & p^{+}_{1(C-1,2)} & \cdots & - & p^{+}_{1(C-1,C)} \\ p^{+}_{1(C,1)} & p^{+}_{1(C,2)} & \cdots & p^{+}_{1(C,C-1)} & - \end{bmatrix}, \ldots, \begin{bmatrix} - & p^{+}_{p(1,2)} & p^{+}_{p(1,3)} & \cdots & p^{+}_{p(1,C)} \\ p^{+}_{p(2,1)} & - & p^{+}_{p(2,3)} & \cdots & p^{+}_{p(2,C)} \\ \cdots & \cdots & - & \cdots & \cdots \\ p^{+}_{p(C-1,1)} & p^{+}_{p(C-1,2)} & \cdots & - & p^{+}_{p(C-1,C)} \\ p^{+}_{p(C,1)} & p^{+}_{p(C,2)} & \cdots & p^{+}_{p(C,C-1)} & - \end{bmatrix} . \quad (8)$$

Finally, let be $p^{+}_{\bullet(j,h)}$ the directional *p*-value statistic calculated via nonparametric combination methodology (see Pesarin and Salmaso, 2010a) and related to the multivariate scatter stochastic superiority alternatives $H^{+}_{1(jh)} : \mathbf{Y}_j \overset{d}{>} \mathbf{Y}_h$ in (7bis). All the $C \times (C-1)$ $p^{+}_{\bullet(j,h)}$ can be represented as follows:

$$P^{+}_{\bullet} = \begin{bmatrix} - & p^{+}_{\bullet(1,2)} & p^{+}_{\bullet(1,3)} & \cdots & p^{+}_{\bullet(1,C)} \\ p^{+}_{\bullet(2,1)} & - & p^{+}_{\bullet(2,3)} & \cdots & p^{+}_{\bullet(2,C)} \\ \cdots & \cdots & - & \cdots & \cdots \\ p^{+}_{\bullet(C-1,1)} & p^{+}_{\bullet(C-1,2)} & \cdots & - & p^{+}_{\bullet(C-1,C)} \\ p^{+}_{\bullet(C,1)} & p^{+}_{\bullet(C,2)} & \cdots & p^{+}_{\bullet(C,C-1)} & - \end{bmatrix} . \quad (9)$$

Note that *p*-value statistics in expression (9) indicate either if there is significant global scatter dominance between each pairs of populations and in which global direction this dominance can actually exist. It is worth noting that, on the contrary to what happens for the marginal directional *p*-value statistics, the constraint of summing up to one does not hold in this case, i.e. $p^{+}_{\bullet(j,h)} \neq 1 - p^{-}_{\bullet(j,h)}$.

Now, let α be the chosen significance α–level and let S be the $C \times C$ matrix which transforms the adjusted (by multiplicity) p-values $p^{+}_{\bullet(j,h)adj}$ into 0-and-1 scores where each element $s_{(j,h)}$ takes the value of 0 if $p^{+}_{\bullet(j,h)adj} > \alpha/2$, and 1 otherwise (if $p^{+}_{\bullet(j,h)adj} \leq \alpha/2$), i.e.

$$S = \begin{bmatrix} - & s_{(1,2)} & s_{(1,3)} & \cdots & s_{(1,C)} \\ s_{(2,1)} & - & s_{(2,3)} & \cdots & s_{(2,C)} \\ s_{(3,1)} & s_{(3,2)} & - & \cdots & s_{(3,C)} \\ \cdots & \cdots & \cdots & - & \cdots \\ s_{(C,1)} & s_{(C,2)} & \cdots & s_{(C,C-1)} & - \end{bmatrix}. \tag{10}$$

In practice, S is nothing more than a synthetic representation of results from all multivariate scatter directional pairwise comparisons suitable for estimating the possible pairwise dominances. If we consider the sum of the $s_{(j,h)}$ 0-1 scores along the h-th column or the j-th row, then we are respectively counting the number of populations which, at the chosen significance α–level, are considered to be stochastically larger or smaller. Hence, we are defining an estimate ${}_{scatter}\hat{r}(\Pi_h)$ and ${}_{scatter}\hat{r}(\Pi_j)$ of the rank ${}_{scatter}r(\Pi_h)$ or ${}_{scatter}r(\Pi_j)$, i.e. the relative stochastic ordering of each population when compared with all other populations by referring to the ranking definitions (1) or (2), i.e. more formally

$${}_{scatter}\hat{r}^{D}_{h} = 1 + \sum_{j=1}^{C} s_{(j,h)}, \; h=1,...,C, \tag{11}$$

$${}_{scatter}\hat{r}^{U}_{j} = 1 + \{\#\,(C - \sum_{h=1}^{C} s_{(j,h)}) > (C - \sum_{h=1}^{C} s_{(j',h)}),\; j'=1,...,C,\; j' \neq j\},\; j=1,...,C, \tag{12}$$

where D and U stands for downward and upward rank estimates respectively. According to the ranking definitions (1) and (2), we note that the ranking estimators defined in (9) and (10) are deriving by counting, on the basis of empirical evidence, of how many populations are significantly scatter stochastically larger/smaller than the h-th/j-th population at the chosen significance α–level. The two estimated rankings ${}_{scatter}\hat{r}^{D}$ and ${}_{scatter}\hat{r}^{U}$ of the true rank ${}_{scatter}r$ are intentionally denoted with a different notation in order to highlight that sometimes they could provide different rank estimates for the same population because of the intransitivity issue (for details see Arboretti et al., 2014).
The same arguments from (8) to (12) can be applied to the location-based multivariate ranking problem and can be found in Arboretti et al. (2014). It is worth noting that the estimation process of population ranks is performed by means of multivariate directional pairwise comparisons and it could be affected by the so-called intransitivity problem (Dayton, 2003), i.e. a possible inconsistency arising from pairwise results (for more details see Arboretti et al., 2014).

3. Applications to biomedical research and industrial quality management

The idea of ranking occurs more or less explicitly any time when in a study the goal is to determine a post-hoc ordering among several input conditions/treatments with respect to one or more outputs of interest when there might be a "natural ordering". In this section we introduce two real case studies in the fields of biomedical research and industrial quality management, i.e. a cytoarchitectonic study of the cerebral cortex and a search for the best storing condition in the leather industry.
As about the first real case study, with the goal of increase insight into the effect of estradiol - E2 on differentiation of neural growing *in vitro*, the effects of estrogens on neuritic development were evaluated with or without 17-β estradiol 100nM added to the medium components. Cerebellar neurones from sexually segregate bovine fetuses were cultured following an established laboratory protocol (Peruffo et al. 2008). This study is focused on understanding the trophic actions of E2 on the growth of neurons in primary cultures obtained from fetal bovine cerebellum. On a total of 829 identified neuron, in our analysis the following morphological endpoints have been measured: the whole area and the whole perimeter of

neuronal and glial cells (somata); the possible presence of cells with primary and secondary branches; the total number of each order of dendritic branches per cell; the total branch length (i.e. the sum of all dendritic segments) per hierarchic order per neuron; the total branch diameter (i.e. the sum of all dendritic diameters) per order per neuron. The main ranking problem was to find out if both male and female E2-exposed primary cell cultures show a location and/or scatter significant trophic effect when compared with the corresponding control group.
The second real case study is focused on an histological analysis of the skin dermal components in bovine hides stored under different conditions (Montelli et al., 2015). Leather industries are interested to avoid post-mortem alterations of the skin components, since degeneration of the dermal structures composing raw hides decreases the quality of leather so that a scatter-based multivariate ranking can be useful to choose the desired curing and timing conditions to employ during refrigeration or salt-based treatment of the skins.

4. Conclusions

In this paper we proposed a novel nonparametric permutation and combination-based approach aimed at ranking of multivariate populations from the scatter point of view. The proposed solution requires a key element: a general hypothesis testing procedure for directional multivariate alternatives by means of the nonparametric combination of dependent permutation tests using suitable dispersion-sensitive test statistics (Pesarin and Salmaso, 2010). This approach provides an exact solution for whatever sample size, is very low demanding in terms of assumptions and finally is quite flexible and it may be extended in the future either for the ordered categorical response variables either for the mixed case, i.e. the jointly presence of mixed response variables, i.e. numeric, binary and ordered categorical even in the presence of any non-informative or informative missing data (missing completely at random or not at random). Furthermore, thanks to a family-wise error rate control by closed testing methods (Pesarin and Salmaso, 2010, we can easily and effectively control for multiplicity without the need to refer to traditional but very conservative methods such as the Bonferroni correction.

References

Arboretti Giancristofaro R., Bonnini S., Corain L., Salmaso L. (2014). A Permutation Approach for Ranking of Multivariate Populations, *Journal of Multivariate Analysis*, 132, 39–57.
Bertoluzzo F., Pesarin F., Salmaso L. (2013). On Multi-Sided Permutation Tests, *Communications in Statistics: Simulation and Computation*, 46, 6, 1380-1390.
Dayton M.C. (2003). Model Comparisons Using Information Measures, *Journal of Modern Applied Statistical Methods*, 2, 2, 281-292.
Gupta S.S., Panchapakesan S. (2002). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. SIAM –Society for Industrial and Applied Mathematics, Philadelphia, USA.
Hall P., Miller H. (2009). Using the Bootstrap to Quantify the Authority of an Empirical Ranking, *The Annals of Statistics*, 37, 6B, 3929–3959.
Montelli S., Corain L., Cozzi B., Peruffo A. (2015). Histological analysis of the skin dermal components in bovine hides stored under different conditions, *Journal of the American Leather Chemists Association*, 110, 54-61.
Peruffo A., Buson G., Cozzi B., Ballarin C. (2008). Primary cell cultures from fetal bovine hypothalamus and cerebral cortex: a reliable model to study P450Arom and alpha and beta estrogen receptors in vitro. *Neuroscience Letters*, 434, 1, 83-7.
Pesarin F., Salmaso L. (2010). *Permutation tests for complex data: theory, applications and software*. Wiley, Chichester, UK.
Shaked M., Shanthikumar J.G. (2007). *Stochastic Orders*. Springer Series in Statistics, New York.