doi.org/10.26398/asaproc.0002

Multivariate permutation-based ranking: An application to college basketball data

Rosa Arboretti [a], Nicolò Biasetton [b], Riccardo Ceccato [b], Livio Corain [b], Giacomo Vezzosi [b]

[a] Department of Civil, Environmental and Architectural Engineering, University of Padua, Padua, Italy

[b] Department of Management and Engineering, University of Padua, Vicenza, Italy

1. Introduction

Players in a basketball team can perform quite differently, even when playing in the same position. It is therefore fundamental for team managers to be supported in their technical and tactical decisions with insights from objective statistics. Data collected during matches and analyzed using appropriate techniques assist managers in choosing the best team for each match. The aim of this paper is to introduce a multivariate nonparametric procedure, based on the work by Arboretti et al. (2014), to enhance decision-making in this context.

The ranking procedure proposed by Arboretti et al. (2014) is indeed a valuable and versatile solution for ranking players according to their performances on the court. This approach exploits multiple pairwise comparisons and uses the achieved p-values to compute a ranking of C multivariate (or univariate) populations $\mathbf{X}_c$. The user is firstly required to identify appropriate statistical tests to assess multiple systems of hypotheses in the form:

$$\begin{cases} H_0: \mathbf{X}_c \overset{d}{=} \mathbf{X}_k \\ H_1: \mathbf{X}_c \overset{d}{>} \mathbf{X}_k, \end{cases} c, k = 1, \dots, C, c \neq k. \tag{1}$$

In the context of basketball, we have multiple performance measures for each player that we want to take into account in our comparisons. Given the multivariate nature of our data and the absence of information on data distribution, we relied on the NonParametric Combination (NPC) technique (Pesarin and Salmaso, 2010) to perform comparisons between pairs of players. NPC is in fact a highly flexible permutation-based methodology which proved to be quite successful in multivariate scenarios.

Let us assume that in a generic comparison (P_c,P_k) we want to compare the performance of two players, namely P_c and P_k. For each player we have information on V variables (i.e., performance indexes). The NPC methodology initially decomposes system of hypotheses (1) into V sub-systems (i.e., one for each performance index). The generic v^{th} sub-system of hypotheses (v = 1, ..., V) is as follows:

$$\begin{cases} H_{0v}: X_{cv} \overset{d}{=} X_{kv} \\ H_{1v}: X_{cv} \overset{d}{>} X_{kv} \end{cases} c, k = 1, \dots, C, c \neq k. \tag{2}$$

The user now needs to choose an appropriate test statistic, such as the difference in mean or the Anderson-Darling test statistic proposed in Pesarin and Salmaso (2010). For the problem at hand we choose the difference in medians to take into account the presence of possible outliers and the numerical nature of the data: $T_{c,k}^{v} = \text{median}_{cv} - \text{median}_{kv}$.

The NPC technique then addresses each sub-system individually and returns V partial p-values achieved through permutation. In doing so, the same permutation mechanism is applied

for the computation of each partial *p*-value (Pesarin and Salmaso, 2010) so that we implicitly take into account the existing dependence between variables. This allows us to compare players in terms of individual performance indexes.

A combination step is then required to achieve an overall evaluation of player performances. The user must firstly choose an appropriate combining function, a choice driven mainly by the number of available sub-problems in which the null hypothesis is expected to be rejected and by the correlation between variables (Langthaler et al., 2022). In this study, we rely on Fisher's combining function because it has been shown to be the best solution in many scenarios. A combination step is therefore undertaken and a combined *p*-value is achieved. This p-value merges the insights provided by the partial *p*-values and can be used to address the multivariate problem of interest (see system 1).

After using NPC to perform all possible directional pairwise comparisons (P_c, P_k), $c,k = 1,...,C$, $c \neq k$ and retrieve the related $C\times(C-1)$ p-values, we move on to the second step of the ranking procedure. Let us suppose that Λ denotes the $C\times C$ matrix containing all the combined p-values $\lambda^{(c,k)}$, $c,k = 1,...,C$. A multiplicity correction is applied to the matrix of p-values Λ and the matrix of adjusted p-values Λ_{adj} is achieved using commonly accepted corrections such as the Bonferroni-Holm-Shaffer (Shaffer, 1986), the Bonferroni-Holm (Holm, 1979), or the Benjamini-Hochberg (Benjamini and Hochberg, 1995). A matrix Z is then created to keep track of the significant comparisons, with $Z_{c,k} = 1$ if $\lambda_{adj}^{(c,k)} \leq \alpha/2$ and $Z_{c,k} = 0$ if $\lambda_{adj}^{(c,k)} > \alpha/2$, where α is the desired significance level. A vector of downward rank estimates is firstly computed as:

$$r_d^k = 1 + \sum_{c=1}^{C} Z_{c,k}\,, k = 1, \dots, C,$$

counting how many populations are stochastically larger than the k^{th} population. A vector of upward rank estimates is calculated as:

$$r_u^c = 1 + \#\left[(C - \sum_{k=1}^{C} Z_{c,k}) > (C - \sum_{k=1}^{C} Z_{c',k})\right], c' = 1, \dots, C, c' \neq c, c = 1, \dots, C,$$

where # means number of times. In this case, we are counting how many populations are stochastically smaller than the c^{th} population. The final vector of ranking estimates is computed using both the downward and upward estimates as:

$$r^c = 1 + \#\left[\frac{(r_u^c + r_d^c)}{2} > \frac{(r_u^k + r_d^k)}{2}\right],$$

$k = 1, ..., C$, $c \neq k$, $c = 1, ..., C$, and can be used to identify the best population according to the considered hypotheses.

2. Case study description

The case study analyzed in this research involves a total of 22 matches (14 from the regular season and 8 from the playoff round) played by the women's basketball team of the Padua University Sports Centre (CUS) during the 2022-2023 season. For each match and each player who took part in the match (both as a starter and a substitute), several pieces of information were collected, and in particular: shooting statistics (points scored; field goals attempted and made, distinguishing between 2 and 3-point shots; free throws attempted and made; throws attempted and made from different areas of the court), assists, offensive and defensive rebounds, blocks, steals, fouls, turnovers and minutes played.

As described in Corain et al. (2019) and Metulini and Gnecco (2023), several indexes were computed to evaluate player performance and allow more systematic comparisons. Based on data collected during the matches, the following indexes were computed:

- Floor impact counter (FIC) (Ferrario, 2021). This index represents a player's evaluation rating with greater importance given to the construction of offensive actions, in particular to offensive rebounds and assists. Other statistics are considered in the computation, such as points, defensive rebounds, steals and blocks (positive), but also missed field goals, missed free throws, turnovers and fouls committed (negative).
- Performance index rating (PIR) (Cene et al., 2018). This metric is useful to assess the efficiency of players in a match and is computed by summing up all the positive actions (i.e., points, assists, total rebounds, steals, blocks, received fouls) and subtracting all the negative actions (i.e., missed field goals, missed free throws, turnovers, fouls committed) performed by the player in the match.
- Adjusted field goal (AFG) (Cene et al., 2018). This index allows us to measure the shooting ability of each player by considering points scored as well as free throws made and attempted field goals.
- Player impact estimate (PIE) (Senatore et al., 2022). This index is useful to compare players and teams and is computed for each player, each team, and each match. It provides a measure of the overall contribution of the player and the team to each match.
- Offensive efficiency (OE) (Lee and Page, 2021). This index measures the quality of a player's offensive contribution and is computed as the ratio between the number of profitable offensive possessions in which the player was involved (sum of field goals made and assists) and the player's total number of potential end-of-possession situations (sum of attempted field goals, assists and turnovers, subtracting offensive rebounds).
- Efficient offensive production (EOP) (Lee and Page, 2021). This index considers the contribution of an assist to the final points scored, estimated using a tuning parameter equal to 0.75, and is computed as: $(0.75 \times \text{assists} + \text{points}) \times \text{OE}$, thus derived from the previous measure.

The conducted analysis, which aims to extend the NPC technique to the analysis of c samples with a dedicated ranking procedure, looked at three players from the same team (identified as Player 1, Player 2 and Player 3) and the previously described performance indexes were computed. The number of matches and minutes played by the three players were comparable therefore this was not taken into consideration. To apply the ranking procedure, only data from matches in which all three players took part at the same time were considered (13 matches out of 14 for the regular season, and all 8 matches for the playoff round).

Figure 1 shows the boxplot of the distributions of the performance indexes computed using data from the regular season and the playoffs. In terms of median value, Player 3 is seen to have performed better overall than her teammates during the regular season. Similarly, during the playoffs, Player 3 seems to outperform the others in terms of median value for almost all computed indexes.

3. Findings and conclusion

The results of the permutation-based ranking procedure for both regular season and playoff matches are reported in Table 1. As anticipated in the descriptive analysis, Player 3 has the best ranking within the group, both in the regular season and in the playoffs, for all indexes (in terms of the median value) as well as globally.

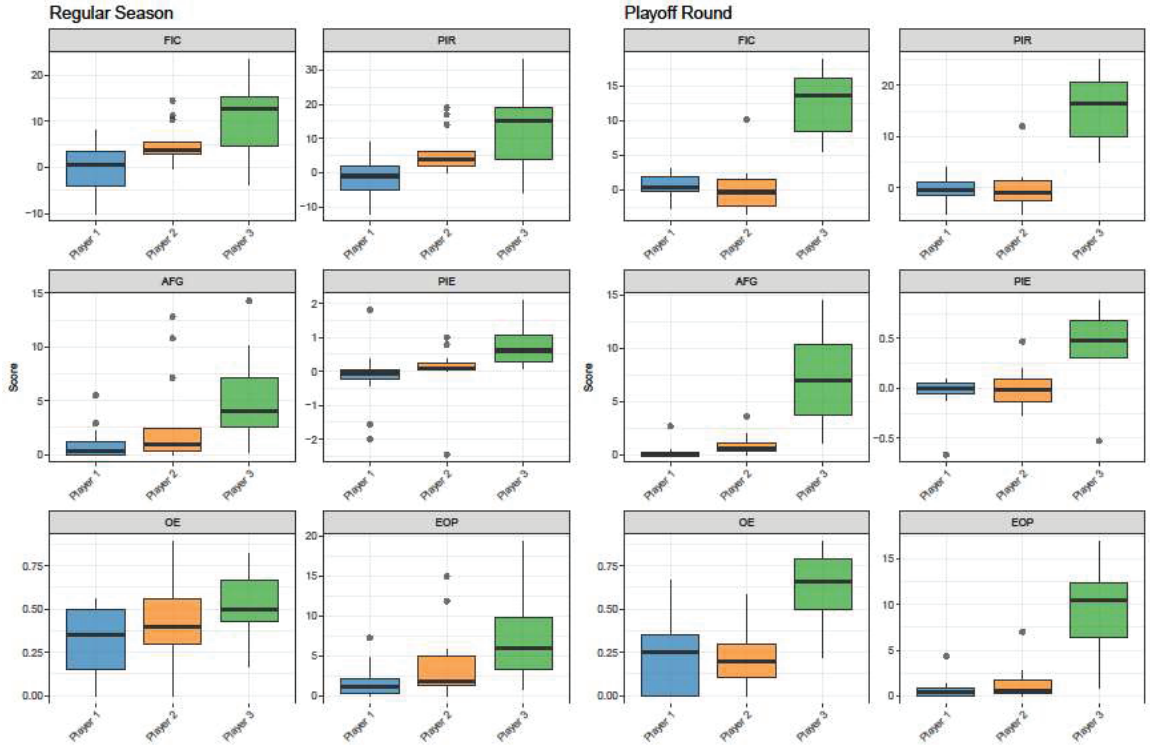

Figure 1: Boxplot of the distributions of the performance indexes per player for both regular season and playoff matches.

Players 1 and 2 have the same global ranking therefore their overall performance can be considered quite similar, even if the results obtained for individual indexes show a better ranking for Player 2 in the regular season. Additionally, in this period, the values obtained for the OE index show substantial performance similarities, therefore all three players have the same ranking if only this index is considered.

Table 1: Ranking of players for each index and globally, computed both on regular season and playoff matches.

	Regular Season			Playoff Round		
Index	Rank Player 1	Rank Player 2	Rank Player 3	Rank Player 1	Rank Player 2	Rank Player 3
FIC	3	2	1	2	2	1
PIR	3	2	1	2	2	1
AFG	3	2	1	2	2	1
PIE	3	2	1	2	2	1
OE	1	1	1	2	2	1
EOP	3	2	1	2	2	1
Global	3	2	1	2	2	1

This study shows the usefulness of the ranking procedure by Arboretti et al. (2014) for the analysis of sports data. In particular, it serves as a valuable tool for ranking players according to their performances, aiding team managers in selecting the best possible team. The proposed procedure can easily be extended to the analysis of larger groups of players and performance indexes.

Acknowledgement

This study was carried out within the framework of the MICS (Made in Italy - Circular and Sustainable) extended partnership and received funding from the European Union NextGeneration EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 - D.D. 1551.11-10-2022, PE00000004). This article reflects only the authors' views and opinions, and neither the European Union nor the European Commission can be considered responsible for them. This study was also supported by the following projects: BIRD 2022 project entitled "Fuzzy theory in Unsupervised Machine Learning algorithm and Sentiment Analysis"; PNRR - MOST - Centro Nazionale per la MObilità soSTenibile, Spoke8: Maas & Innovative services.

References

Arboretti, R., Bonnini, S., Corain, L., Salmaso, L. (2014). A permutation approach for ranking of multivariate populations. *Journal of Multivariate Analysis*, **132**, pp. 39–57.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**(1), pp. 289–300.

Cene, E., Parim, C., Ozkan, B. (2018). Comparing the performance of basketball players¨ with decision trees and TOPSIS. *Data Science and Applications*, **1**(1), pp. 21–28.

Corain, L., Arboretti, R., Ceccato, R., Ronchi, F., Salmaso, L. (2019). Testing and ranking on round-robin design for data sport analytics with application to basketball. *Statistical Modelling*, **19**(1), pp. 5–27.

Ferrario, A. (2021). *Basketball Analytics: The Use of Data Science to Describe and Predict the Performance of an NBA Team*. Master's Thesis, Università degli Studi di Milano Bicocca.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**(2), pp. 65–70.

Langthaler, P.B., Ceccato, R., Salmaso, L., Arboretti, R., Bathke, A.C. (2022). Permutation testing for thick data when the number of variables is much greater than the sample size: Recent developments and some recommendations. *Computational Statistics*, **38**(1), pp. 1–32.

Lee, D.-J., Page, G.L. (2021). Big data in sports: Predictive models for basketball player's performance. *Mathematics in Industry Reports* (https://www.academia.edu/70747795/Big_Data_in_Sports_Predictive_Models_for_Basketball_Players_Performance)

Metulini, R., Gnecco, G. (2023). Measuring players' importance in basketball using the generalized Shapley value. *Annals of Operations Research*, 325(1), pp. 441–465.

Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. John Wiley & Sons.

Senatore, J.V., Fellingham, G., Lamas, L. (2022). Efficiency and productivity evaluation of basketball players' performance. *Motriz: Revista de Educação Fısica*, **28**.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**(395), pp. 826–831.