



Dyadic analysis for multi-block data in sport surveys analytics

Maria Iannario¹ · Rosaria Romano² · Domenico Vistocco¹

Accepted: 7 June 2022 / Published online: 7 July 2022
© The Author(s) 2022

Abstract

Analyzing sports data has become a challenging issue as it involves not standard data structures coming from several sources and with different formats, being often high dimensional and complex. This paper deals with a dyadic structure (athletes/coaches), characterized by a large number of manifest and latent variables. Data were collected in a survey administered within a joint project of University of Naples Federico II and Italian Swimmer Federation. The survey gathers information about psychosocial aspects influencing swimmers' performance. The paper introduces a data processing method for dyadic data by presenting an alternative approach with respect to the current used models and provides an analysis of psychological factors affecting the actor/partner interdependence by means of a quantile regression. The obtained results could be an asset to design strategies and actions both for coaches and swimmers establishing an original use of statistical methods for analysing athletes psychological behaviour.

Keywords Athletes'/coaches' perception · Dyadic analysis · Complex data structures · Quantile regression

1 Introduction

The main goal of competitive sports is to yield a higher ranking sporting performance getting promotions or winning relevant competitions. Data concerning athletes' performance and their behavior are the essential core for competitive sports (Du & Yuan, 2021). Technological enhancements concerning remote sensor technology or motion capture systems have been developed together with advanced statistical approaches to model, infer or predict perfor-

✉ Maria Iannario
maria.iannario@unina.it

Rosaria Romano
rosaroma@unina.it

Domenico Vistocco
domenico.vistocco@unina.it

¹ Department of Political Sciences, University of Naples Federico II, Via L. Rodinó, 22, 80138 Naples, Italy

² Department of Economics and Statistics, University of Naples Federico II, Via Cintia, 21, 80126 Naples, Italy

mance outcomes in sport [e.g. (Albert et al., 2016; Nevill et al., 2008)]. An analysis of sports statistics can effectively identify athletes' behavioral patterns (Legg et al., 2012; Losada et al., 2016), including their individual contribution and degree of activity. However, sports data contain multiple dimensions, and relying purely on numbers cannot fully represent the data analysis results. Apart from the value of the performance indicators used to categorise athletes' and teams' achievements, there is still a lack of understanding regarding why and how some behaviours emerge in performance contexts (McGarry, 2009). It provides a basis for the study of the law of athletes' life and the habits of athletes' beings (Du & Yuan, 2021).

Recently, an increasing interest has been devoted to understanding the psychological behaviour of some athletes and how personality traits influence their performance [see (Aidman & Schofield, 2004; Laborde et al., 2020), among others], also by means of complex statistical models (Fabbriatore et al., 2021; Fabbriatore & Iannario, 2022). This data analysis work would be helpful for professional analytics, allowing effective behavior-based decision-making during games, improving the effects of teams' training and performance in competitions (Janetzko et al., 2014; Legg et al., 2012, 2013; Rusu et al., 2010). Some contributions focused on interpersonal relationships in athlete-athlete and coach-athlete dyads studying their interdependence to improve results (Bell, 2007; Jowett & Nezek, 2012; Rhind & Jowett, 2011). Others collect Likert-type responses as psychometric item scoring schemes for attempting to quantify athletes' opinions, interests, or perceived efficacy of an intervention relaying on multi-block data. The statistical techniques that can be used to analyse them are called multi-block methods (Smilde et al., 2000). The essential requirement is that these blocks have one dimension or *mode* in common, i.e. different groups of variables are observed on the same statistical units, or the same set of variables is observed on different groups of statistical units. Our contribution fits these areas aiming to understand the *feeling* between coach and athlete, their possible (*dis*)*agreement* and the psychological factors/reasons that motivate/influence their relationship. The coach-athlete dyad is probably one of the most relevant within athletic communities because it is a relationship whereby the athlete expresses needs (e.g., autonomy, competence) and goals (e.g., skill development, performance success) (Côté & Gilbert, 2009; Lyle, 2002). The analysis concerns several observed variables collected on the dyads, which are grouped into homogeneous blocks measuring partial aspects of the phenomenon under investigation. Data come from a survey collected in 2019 for the Statistical Modelling and Data Analytics for Sports project, which involved the University of Naples Federico II and the Italian Swimmer Federation (Campania Regional Committee). They gather information about psychosocial aspects influencing swimmers' performance. Data concern 100 elite swimmers and their coaches; the latter were sampled by the Italian Swimmer Federation (Campania Regional Committee) picking among their lists with a random selection. Each coach randomly selected one of her/his athlete and asked her/him to fill out a questionnaire. Interviewed people answered several questions concerning their mental strategies and skills based on one of the main theoretical framework for analysing personality and coping behaviour in sport: the five-factor model (McCrae & Costa, 2008), also named Big Five (BigF5). The complex structure of data derives both from the dyads and the large number of observed and corresponding latent variables with respect to the sample size. The degree of complexity is also related to the mixed-type data, categorical and quantitative ones. This enhances practical and theoretical challenges requiring a specific treatment and a "statistical learning" from data (Hastie et al., 2013).

The approach pursued in our proposal exploits quantile regression (Koenker & Bassett, 1978), in line with a previous proposal (Davino et al., 2020) where such method has been used as kernel of a strategy to assess heterogeneity in a different multi-block type data structure. The study models the actor/partner interdependence in the case of dyadic data by presenting

an alternative approach with respect to the current used methods (Kenny et al., 2006). After a preliminary analysis of the athletes/coaches matrix of responses, aiming to evaluate the consistency of perceived assessments on some topics related to the performance, a quantile regression has been implemented to disclose how disagreement is connected to athletes' psychological aspects. The remainder of the paper is structured as follows. Section 2 provides a detailed description of the survey, the complex data and the latent variables considered in the research. Section 3 illustrates the statistical method used for the analysis and the main results. Conclusions and discussion of further research developments to be explored are included in Sect. 4.

2 Survey description

In the last few years, it has been common practice to collect complex data sets composed of different groups of variables observed on the same units. Such costume has been largely adopted in the contexts of *sports analytics* (Lebed, 2017), where surveys and in-depth interviews are collected for both description, and prediction aims (Davenport & Harris, 2007). In this study, the complexity is threefold. On the one hand, there are dyads, consisting of scores/evaluations provided by coaches and athletes on the athletes' performance. On the other hand, there are the variables on athletes: the athletic profile, personal data, habits and demographics. Finally, psychometric scales are used to measure elite swimmers' latent traits, i.e. groups of items that aim to measure the corresponding latent constructs. Analyzing such complex datasets requires a multi-step strategy investigating the relationships between different groups of variables using supervised and unsupervised methods. In some cases, the groups of variables are synthesized through dimensionality reduction techniques (Hastie et al., 2013), specifically principal component analysis (Jolliffe, 1986). At other times, as in the case of psychometric scales, an appropriate synthesis can be the sum or the average of the items of each block (McNeish & Wolf, 2020).

A sample of 100 elite swimmers (from now on, simply swimmers) enrolled in professional-level registers was examined. They were randomly selected by coaches who, in turn, were sampled by the Italian Swimming Federation (Campania unit) list. Leading details on the sample are reported in Table 1.

Personality was assessed by using a list of 25 adjectives representative of the five-factor model (Big Five–BigF5) in the Italian lexical context (Barbaranelli et al., 2007; Caprara & Perugini, 1994). The list consists of 5 adjectives for each of the five personality dimensions: Extraversion, Emotional stability, Openness, Agreeableness, and Conscientiousness. Swimmers were required to fill out the questionnaire indicating how appropriate each adjective was for describing themselves on a 5-point scale. Furthermore, athletes and coaches answered nine questions assessed by visual analogue scales about their perceived assessments on some topics related to the performance.

The strategy proposed in this study, graphically outlined in Fig. 1 and detailed in the next section, consists of the following steps, each requiring specific statistical learning techniques:

1. *Analysis of the dyads* (see step 1. in Fig. 1):
 - a. Analysis of the disagreement between the two groups of variables on the athletes' sport performance: here the disagreement matrix \mathbf{Y} was derived resorting to absolute values of the differences in scores of coaches ($\mathbf{Z1}$) and athletes ($\mathbf{Z2}$) (see step 1.a in Fig. 1).

Table 1 Summary statistics concerning the characteristics of the 100 athletes of the survey

Gender
 Female (43)–male (57)

Age
 Min (10)–max (30)–mean (14.81)–standard deviation (3.74)

Style
 Dolphin crawl (9), freestyle stroke (56),
 Backstroke (11), breaststroke (24)

Years of competition
 < 5 (38), 5–10 (33), 10–15 (21), > 15 (8)

Level of competition
 International (9), national (22), regional (69)

Hours of training (per week)
 < 10 (19), 10–15 (56), > 15 (25)

Numbers of podium
 0 (49), < 5 (24), 5–10 (9), 11–15 (8), > 15 (10)

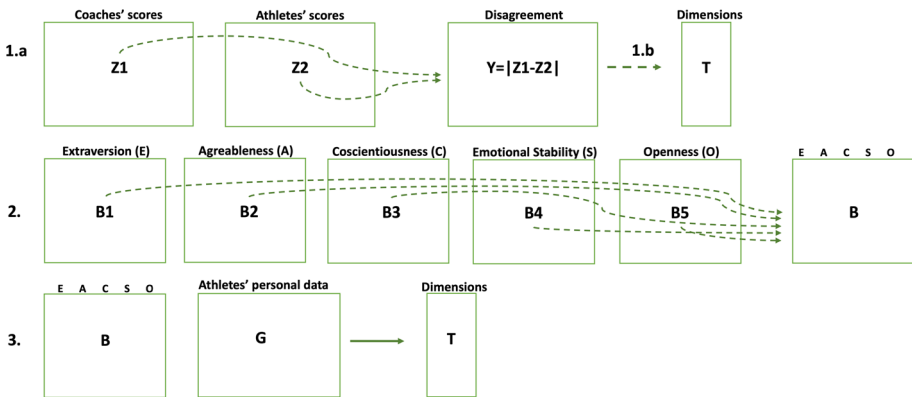


Fig. 1 Scheme of the data analysis strategy: (1.a) the dyad ($Z1 - Z2$) is studied using the deviation of the respective values (Y), (1.b) from which a small number of relevant dimensions (T) are extracted; (2.) the BigF5 psychometric scale is summarized in block B , including the syntheses of the 5 dimensions (E, A, C, S, O); (3.) the synthesis of the BigF5 scale and personal data on the athletes (G) are regressed on the dimensions of the disagreement (T)

b. Synthesis of the disagreement through a reduced number of dimensions embedded in the T matrix: here we exploited principal component analysis (PCA) (Jolliffe, 1986; Mardia et al., 1979) on the disagreement matrix Y (see step 1.b in Fig. 1).

2. *Summary of the psychometric scale:* different options could be exploited for this step, among them: sum/average/weighted average/principal components by dimension. Regardless of the specific method used to synthesise the five dimensions of the BigF5 scale ($B1 - B5$), the final result will be the new block of variables denoted B , which contains the five syntheses (E, A, C, S, O) (see step 2 in Fig. 1).
3. *Analysis of the influence of psychological factors and athletes' personal data on the dimensions of the disagreement:* we exploited quantile regression (QR) (Koenker &

Bassett, 1978) both for its distribution-free nature and for modelling the dependence structure at different locations of the response. Specifically, the dependence relationship between the block of predictors \mathbf{X} , composed of the variables in \mathbf{B} (obtained in step 2) and \mathbf{G} (personal data on the athletes), and the block of dependent variables \mathbf{T} , including the disagreement dimensions obtained in step 1.a, is analyzed through QR (see step 3 in Fig. 1).

3 Data analytics

The following subsections details the steps of the strategy outlined in Fig. 1, briefly recalling the involved statistical methods.

3.1 Analysis of the dyads through dimension reduction methods

In this study, available data follow the dyad structure (Kenny et al., 2006), where each dyad corresponds to a statistical unit on which two levels of each variable are observed. The two levels correspond to *athletes* and *coaches*, and the variables are nine measures of the athletes' sports performance (see Fig. 1 for details). Formally, data are organised into two datasets, each consisting of J variables measured on I objects/dyads. The two datasets, of orders $(I \times J)$, are denoted here by $\mathbf{Z1}$ and $\mathbf{Z2}$: $\mathbf{Z1}$ contains coaches' scores, $\mathbf{Z2}$ athletes' scores. Coaches' and athletes' variables are shown in Fig. 2 where divergent stacked bar charts are reported. Each panel refers to a given variable related to coaches and athletes. The bars in each panel are located with reference to the neutral point scale expressed on a discretized version of a 7-point scale of a visual analogue scale [see (Cox, 1980), for the discretization]. Therefore, if the bar for a given item tends to lie in the right part of the plot, this denotes a percentage of respondents with points in the upper part of the correspondent scale. Inversely, in case the major part of the bar is located in the left part of the plot. Segments of the same level in the same color are comparable across items and panels.

Figure 2 shows that the expressed values for coaches are lower than those of athletes for the variables *anxiety*, *recovery* and *get nervous*, while being higher for *talent*, *work out*, *pressure* and *sacrifice*. The visual inspection indicates the variables with the most outstanding disagreement between coach and athletes. Here, disagreement stands for under(over) evaluation of the athlete with respect to the coach (and vice versa), that is both a level of feeling and compliance from one side and a different perception of ability/skill from the other side.

Different types of multivariate analysis techniques can be used to analyse the relationships between two datasets depending on the relationship hypothesised between the two sets of variables. For example *asymmetric* methods try to predict one dataset from another, thus treating the two datasets differently. Principal Component Regression (Martens & Næs, 1992), Partial Least Squares Regression (Abdi, 2010; Geladi & Kowalski, 1986), Redundancy Analysis (Van Den Wollenberg, 1977) are among these methods. *Symmetric* methods treat the two datasets similarly. Here, the goal is to study relationships between the two sets rather than predict one from the other. Examples of these methods are Canonical Correlations (Hotelling, 1936) and Procrustes Analysis (Schönemann, 1966). The relationship analysed in dyadic data analysis is usually of symmetrical type. Specifically, the main interest in this research is to study the disagreement between coaches' and athletes' scores. Therefore, examining the deviation between the two matrices seems more appropriate. To this end, the

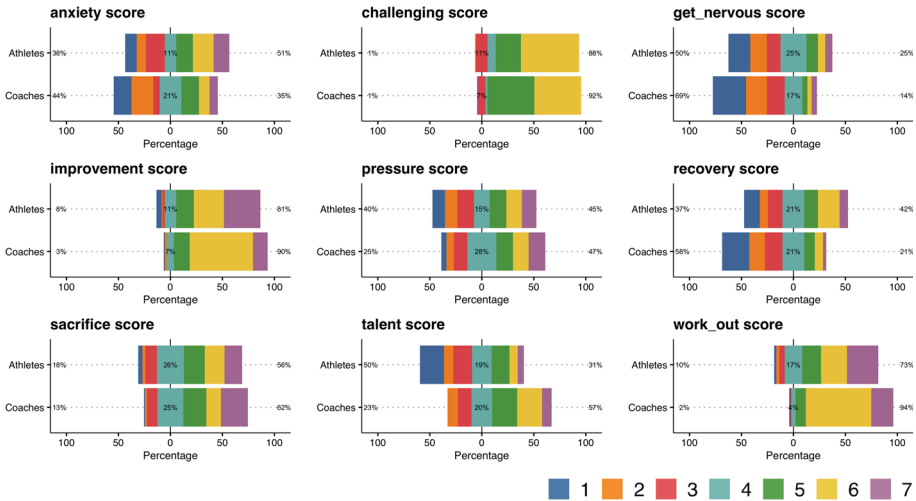


Fig. 2 Divergent stacked bar chart for the nine measures of the athletes’ sport performance

strategy of analysis proposed here consists of calculating the matrix of absolute deviations $\mathbf{Y} = |\mathbf{Z1} - \mathbf{Z2}|$ and applying a PCA to identify the dimensions of maximum disagreement.

PCA is a well-established, and long-standing multivariate statistical technique (Hotelling, 1933; Pearson, 1901) that has made a comeback as an unsupervised machine learning technique (Hastie et al., 2013) due to its information synthesis capabilities. Let us consider a matrix \mathbf{Y} with I rows ($i = 1, \dots, I$, usually samples/objects) and J columns ($j = 1, \dots, J$, usually variables). Formally, PCA aims to obtain a small number of new variables, called principal components, which are linear combinations of the variables in \mathbf{Y} and contain as much as possible variation present in \mathbf{Y} . Given the first linear combination $\mathbf{t} = \mathbf{Y}\mathbf{w}$ and the corresponding variance $var(\mathbf{t})$, the problem turns into maximizing this variance, choosing the optimal \mathbf{w} of length one

$$\underset{\|\mathbf{w}\|}{\operatorname{argmax}} \operatorname{var}(\mathbf{t}), \tag{1}$$

where the matrix \mathbf{Y} is mean-centered. The restriction on the length of \mathbf{w} is needed to obtain a unique solution. The problem in equation (1) is a standard problem in linear algebra and the selected \mathbf{w} is the first eigenvector of the covariance matrix $\mathbf{Y}^T\mathbf{Y}/(I - 1)$.

The following components are obtained in the same way, but with the additional constraint that each is uncorrelated to the component that precedes it. The maximum number A of possible components equals the minimum of $I - 1$ and J . The eigenvalue λ corresponding to each eigenvector provides a measure of the variability conveyed by the component. Therefore, it is possible to calculate the percentage of explained variation of \mathbf{t}_a as

$$\frac{\lambda_a}{\sum_{a=1}^A \lambda_a} 100\%. \tag{2}$$

Figure 3 reports the scree plot of the explained variability (left-hand side), and the biplot (right-hand side) resulting from the PCA on the \mathbf{Y} matrix. The scree plot highlights two relevant dimensions of disagreement. The biplot suggests interpreting the first dimension as a disagreement concerning attitudinal and psychological factors that influence the athletes’

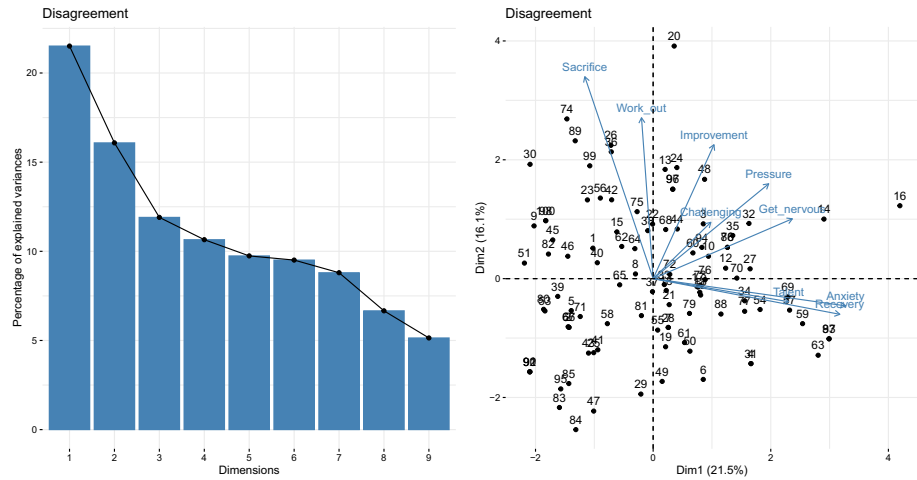


Fig. 3 Percentage of explained variance (left-hand side) and corresponding biplot (right-hand side) carried out by PCA on the disagreement matrix

performance (psycho-attitudinal disagreement). In contrast, the second dimension describes a disagreement with the athletes’ commitment to the discipline (commitment disagreement).

A question of relevant importance for interpreting the results of a PCA concerns the selection of the statistically significant principal components. Several procedures have been proposed for this purpose: resampling methods (Peres-Neto et al., 2005), cross-validation methods (Bro et al., 2008; Josse & Husson, 2012), statistical tests (Choi et al., 2017), and parallel analysis (Horn, 1965). The present work applies the recent bootstrap approach proposed by Forkman et al. (2019). The null hypotheses are $H_0: m = K$, where m is the unknown true number of principal components. Tests at significance level $\alpha = 0.05$ are carried out sequentially, for $k = 0, 1, 2, \dots$, up to $K = M - 2$, where $M = \min(I - 1, J)$, or a non-significant ($p > 0.05$) result is obtained. The corresponding alternative hypotheses are $H_1: m > K$, where K is the candidate order of the model. For the disagreement block, only the first principal component was highly significant ($p = 0.000$), while the second component was not significant ($p = 0.095$).

3.2 Extraction of the latent dimensions of the psychometric scale

Frequently psychologists need to measure abstract concepts (constructs) that are not directly measurable. One way to obtain a more reliable measure of the concept is to use a series of highly correlated items (scale) corresponding to different measurements of the same phenomenon. The construct is defined as one-dimensional if correlations between items emerge from a single underlying dimension. On the other hand, a multidimensional construct has two or more underlying dimensions that appear from groups of variables with high intra-group correlations and less relevant inter-group correlations. A detailed description of the psychometric scale employed in this study and the analysed traits/constructs is reported in Table 2. The five-factor model (Big Five–BigF5) is multidimensional and includes five dimensions, each measured by five items corresponding to adjectives, each measuring one of the five personality dimensions.

Two are the two main methods of combining the different items of a one-dimensional or a multidimensional construct into a single measurement: a) the sum of the scores of the

Table 2 Description of the scales and related constructs used in the analysis*Big Five assessment—BigF5*

Extraversion (determined, energetic, enterprising, dominant, decided)

Agreeableness (unselfish, loving, warm, friendly, fair)

Conscientiousness (responsible, scrupulous, accurate, conscientious, careful)

Emotional stability (relaxed, patient, optimistic, calm, serene)

Openness (imaginative, innovative, creative, original, modern)

individual items (Gleser & Dubois, 1951); b) the factor analysis (Fabrigar et al., 1999). The latter, in particular, is used when the aim is to propose a new scale or to validate a scale on a specific sample. Since the BigF5 is a widely known and utilised scale, and the paper's objective is not to validate it, the sum of the items is adopted to obtain the score of each of the five dimensions. These, as also shown graphically in the Fig. 1, are used in the next section as predictors of the disagreement.

3.3 Analysis of the relations between disagreement and psychosocial factors

This section deals with quantile regression, introducing it in Sect. 3.3.1, then using it on the analyzed data in Sect. 3.3.2 to relate the dimensions of disagreement and the psychosocial variables. The analysis concerns only the first dimension as the second dimension is resulted to be not significant, as described in Sect. 3.1.

3.3.1 A short briefing on quantile regression

Quantile regression (QR) extends classical regression to a set of quantile functions of a response variable \mathbf{y} , conditional on a set of covariates \mathbf{X} . Initially proposed by Koenker and Basset (Koenker & Bassett, 1978), QR is a regression approach completely distribution-free, since it does not pose any parametric assumption on the error (and hence response) distribution. It aims to estimate the effects of a set of regressors on the quantiles of a response variable. In particular, QR estimates separate models for different asymmetries $\tau \in (0, 1)$, where τ denotes the particular conditional quantile of interest. Unlike the classical regression model, where the conditional mean of the error $E(\epsilon_{i\tau}) = 0$, in QR the τ -quantile of the error term is 0, namely $P(\epsilon_{i\tau} \leq 0) = \tau$. The separate models provided by QR, one for each quantile of interest, are interpretable in terms of regression models for the associated conditional quantiles of the response. A dense set of quantiles completely characterizes the conditional distribution of the response: the use of “enough” quantiles makes it possible to virtually analyze any property of the response distribution (Davino et al., 2013). Although it is theoretically possible to extract infinite quantiles, a finite number is numerically distinct in practice, which is known as the quantile process (Furno & Vistocco, 2018). A good tension between the wealth of information and interpretation issues leads to using a selected set of quantiles, typically the three quartiles, along with two extreme quantiles, to model the tails.

As stated above, the effect of covariates acts not only on the conditional mean but on the complete conditional distribution of the response given the covariates. Moreover, not including the classical restrictive assumptions of the mean regression model, QR is well suited to deal with heteroscedasticity, but even more important, to model the higher-order characteristics of the response in terms of covariates. One should also bear in mind that

in many studies, as the assessment reported in our analysis, there is a genuine interest in conditional quantiles, and hence the QR ability to describe extreme observations in terms of covariates provides an incomparable added value. This can be attractive in sports applications, since the focus on the distribution's tails helps inspect dependence models for athletes with low/high performance, low/high level of stress, low/high motivation, and in our case, with low/high disagreement with coaches judgment.

The QR model, linking the response of each single unit to the regressors, is estimated for different quantiles

$$\mathbf{y} = \mathbf{X}\beta(\tau) + \epsilon.$$

The conditional quantile estimator (Koenker, 2005) is

$$\hat{\beta}(\tau) = \underset{\beta(\tau)}{\operatorname{argmin}} \rho_{\tau}(\mathbf{y} - \mathbf{X}\beta(\tau)), \quad (3)$$

where $\rho_{\tau}(\cdot)$ is the check function which asymmetrically weights positive and negative residuals

$$\rho_{\tau}(u) = \begin{cases} \tau u, & \text{if } u > 0 \\ (\tau - 1)u, & \text{if } u \leq 0. \end{cases}$$

Equation (3) provides a quantile regression line for each conditional quantile τ of interest.

The bootstrap procedure is typically used for inference to avoid the assumptions required by finite sample or asymptotic theory. Bootstrap offers the flexibility to obtain the standard error and confidence interval for any estimates and combinations of estimates, whilst also keeping the distribution-free nature of QR. Finally, it is worth recalling that QR estimates are not sensitive to outliers in \mathbf{y} . Specifically, any change in the value of the response variable for a data point lying above (or below) the fitted QR lines does not affect the estimates when the data point does not change its previous position concerning the specific line. Instead, QR estimator can be very sensitive to outliers in the explanatory variables, even if several proposals in the literature attempt to attain more robust estimators (Furno & Vistocco, 2018).

The machinery for solving the quantile regression problem initially exploited linear programming, and in particular, the simplex algorithm (Furno & Vistocco, 2018). However, the least absolute deviation criterion is even more ancient than the most popular least squares counterpart (Stigler, 1986). Wagner (1959) presented the linear programming techniques in the mainstream statistical literature. A few years later, the least absolute deviations were approached using linear programming and later adopted for quantile regression. Indeed, the original algorithm for solving the quantile regression problem (Koenker & D'Orey, 1987) extended the Barrodale and Roberts algorithm (Barrodale & Roberts, 1973) to conditional quantiles. The Barrodale and Roberts algorithm was initially introduced for solving the median regression problem. Koenker and Bassett (1978) slightly modified the original L_1 problem placing asymmetric weights on positive and negative residuals, introducing quantile regression. The simplex approach is not the only available approach for quantile regression. Indeed, while the simplex approach exploits the movement along the corner of the feasible region (exterior-point methods), barrier methods (interior-point methods) start from an initial point inside the feasible region and, at each iteration, move to a better feasible solution. A detailed treatment of interior-point methods is available in Koenker (2000).

See the literature mentioned in Furno and Vistocco (2018) for recent alternative QR estimators. In our analysis, even if analyzed data have a complex structure, their dimension is not huge. Therefore there is no relevant difference in the solutions provided by the different algorithms.

3.3.2 Quantile regression results

As outlined in Sect. 2, QR has been exploited to analyse the influence of psychosocial factors on the first dimension of disagreement. In particular, the five personality dimensions (Extraversion, Emotional stability, Openness, Agreeableness, and Conscientiousness) representative of the BigF5 scale were used as regressors in a QR model: the first dimension of the disagreement obtained using PCA is the response variable. Figure 4 depicts the QR coefficients for the regressors. In particular, the horizontal axis displays the different quantiles, while the effect of each feature holding the others constant (QR estimate) is represented on the vertical axis. The shaded area depicts the confidence intervals, the horizontal solid lines placed at 0 helping to reveal significant effects. Each panel refers to a different personality dimension. The aim is to graphically catch the coefficient trends moving from lower to upper quantiles. Coefficients have been estimated for a sequence of quantiles from 0.1 to 0.9 with a step of 0.05.

Figure 4 shows a significant positive effect of Agreeableness for high levels of disagreement ($\tau = 0.75$). Emotional stability has a significant negative impact on low levels of disagreement ($\tau = 0.25$) and shows a slight upward trend. Extraversion shows a positive and significant effect for almost all levels of disagreement, except the extremes. Finally, Openness always has a negative impact, significant in some extreme parts of the distribution.

Some additional information has been tested in the analysis. Among these, the age of the athlete, the gender, and some measures related to the objective performance, because they may influence the personality and the possible *dis(agreement)*. However they are not significant from a statistical point of view signaling a sort of homogeneity in the sample of respondents.

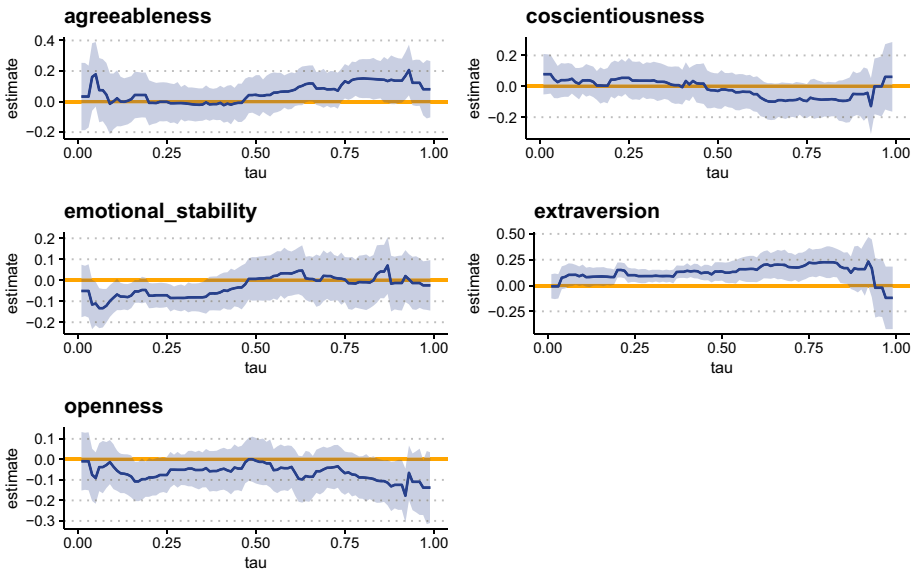


Fig. 4 QR coefficients for significant regressors. Each panel refers to a personality dimension, the horizontal axis depicts the different quantiles, and the vertical axis the QR coefficients. A solid horizontal line is placed at 0. The shaded area in each panel shows the 95% confidence intervals for the QR coefficients

4 Conclusions and discussion

The study of personality in sports psychology is focused on assessing the associations between personality, participation and athletic achievement (Aidman & Schofield, 2004; Allen et al., 2013; Allen & Laborde, 2014). When the BigF5 scale is considered, the main findings concerning organized sports suggest higher athletes' score on Extraversion (Egloff & Gruhn, 1996), Conscientiousness (Kajtna et al., 2004), Emotional stability (Kajtna et al., 2004; Mckelvie et al., 2003), and Openness (Kajtna et al., 2004) when compared with non-athletes. Individual sport athletes, instead, demonstrated higher Conscientiousness, Openness and Emotional stability, as well as lower levels of Extraversion than team-sport athletes (Allen et al., 2011; Eagleton et al., 2007). The analysis we reported opens a new perspective. It shows the relationship in the coach/athlete dyad. The disagreement explained as the score distance between the subjective evaluation of the couple, is a measure of understanding and synergy. Higher levels of disagreement point out an under(over) evaluation of the athlete with respect to the coach (and vice versa), which is both a low level of feeling and compliance from one side and a different perception of ability/skill from the other side. The study on elite swimmers underlines high levels of disagreement for agreeable athletes, that is, polite, trusting, and cooperative, over competitive subjects. These athletes, evaluated as straightforward, altruistic, compliant and modest, may underestimate their value with respect to coaches' perception. Emotionally stable swimmers, a characteristic trait of high-level athletes (e.g., athletes competing at a national or international level) (Allen et al., 2011; Kirkcaldy, 1982) reduce the distance of their evaluation concerning their coaches (Allen et al., 2011; Kirkcaldy, 1982). This result is marked for the highest level of compliance, possibly related to swimmers with high levels of self-confidence. Openness athletes, more aware of their feelings and more likely to engage in risky behaviour, reveal a high distance from their coach's assessment.

A multi-step data analysis strategy gave these results, making it possible to extract information from multiple sources. Despite the remarkable development of statistical methods for multi-block data, when blocks are linked together in a complex structure of relationships, there is no simultaneous analysis method. Therefore, a strategy that sequentially combines information sources (blocks) is needed. In the approach adopted in this study, we used some well-founded methods such as PCA and others more recently introduced, such as QR. We faced the dyadic structure starting with an analysis of the differences in the scores of the athletes and their coaches. Then, relevant dimensions were derived from these differences. Such dimensions can be interpreted as dimensions of disagreement. Next, we summarized the blocks relating to the different dimensions of the psychological scale into a single block by adding the scores of the individual items. Finally, we connected the two previous syntheses through a linear quantile regression model. The use of quantile regression was crucial to highlight the impact of psychological variables on the whole distribution of disagreement. For example, it was possible to highlight a significant effect of agreeableness on high levels of disagreement ($\tau = 0.75$), effect not detectable using the classical linear regression model. The innovative contribution of this paper is, therefore, twofold. On the one hand, the dataset's structure combines the dyad with the psychometric scale and other demographic variables of the athletes. On the other hand, the proposed data analysis strategy combines dimensionality reduction techniques with a quantile regression model. The study signals the personality traits on which the dyadic relationship needs a more efficient dialogue to eliminate possible gaps. It also deals with a homogeneous sample in which young athletes' age stands for a limitation, especially for assessing the questions related to their personality. It remains unclear whether the results are cyclically related and caused by such factors as sociocultural variables

and other factors related to the coaches' relationship. Further investigations on these topics represent a possible future subject of research analysis.

Acknowledgements This research was carried out in the context of the project “Statistical Modelling and Data Analytics for Sports. Psychosocial aspects to assess the performance: the case of swimmers” (University of Naples Federico II—Italian Swimming Federation, Campania Regional Committee) and partially supported by Osservatorio Regionale delle Politiche Giovanili 2—POR CAMPANIA FSE 2014-2020-Cup project: E64119002390005.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 97–106.
- Aidman, E., & Schofield, G. (2004). Personality and individual differences in sport. In T. Morris, & J. Summers (Eds.), *Sport psychology: Theory, applications and issues*, 2nd edn (pp. 22–47). Wiley.
- Albert, J., Glickman, M. E., Swartz, T. B., & Koning, R. H. (2016). *Handbook of statistical methods and analyses in sports*, 1st edn. Chapman and Hall/CRC.
- Allen, M. S., Greenlees, I., & Jones, M. (2011). An investigation of the five-factor model of personality and coping behaviour in sport. *Journal of Sports Sciences*, 29(8), 841–850.
- Allen, M. S., Greenlees, I., & Jones, M. (2013). Personality in sport: A comprehensive review. *International Review of Sport and Exercise Psychology*, 6(1), 184–208.
- Allen, M. S., & Laborde, S. (2014). The role of personality in sport and physical activity. *Current Directions, Psychological Science*, 23(6), 460–465.
- Barbaranelli, C., Caprara, G. V., Vecchione, M., & Fraley, C. R. (2007). Voters' personality traits in presidential elections. *Personality and Individual Differences*, 42(7), 1199–1208.
- Barrodale, I., & Roberts, F. (1973). An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10, 839–848.
- Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta analysis. *Journal of Applied Psychology*, 92(3), 595–615.
- Bro, R., Kjeldahl, K., Smilde, A. K., & Kiers, H. A. L. (2008). Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390(5), 1241–1251.
- Caprara, G. V., & Perugini, M. (1994). Personality described by adjectives: The generalizability of the Big Five to the Italian lexical context. *European Journal of Personality*, 8(5), 357–369.
- Choi, Y., Taylor, J., & Tibshirani, R. (2017). Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, 45, 2590–2617.
- Côté, J., & Gilbert, W. (2009). An integrative definition of coaching effectiveness and expertise. *International Journal of Sports Science and Coaching*, 4, 307–323.
- Cox, E. P., III. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4), 407–422.
- Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business Press.
- Davino, C., Furno, M., & Vistocco, D. (2013). *Quantile regression: Theory and applications*. Wiley.
- Davino, C., Romano, R., & Vistocco, D. (2020). On the use of quantile regression to deal with heterogeneity: The case of multi-block data. *Advances in Data Analysis and Classification*, 14, 771–784.
- Du, M., & Yuan, X. (2021). A survey of competitive sports data visualization and visual analysis. *Journal of Visualization*, 24, 47–67.

- Eagleton, J. R., McKelvie, S. J., & De Man, A. (2007). Extraversion and neuroticism in team sport participants, individual sport participants, and nonparticipants. *Perceptual and Motor Skills*, 105(1), 265–275.
- Egloff, B., & Gruhn, A. J. (1996). Personality and endurance sports. *Personality and Individual Differences*, 21(2), 223–229.
- Fabbricatore, R., & Iannario, M. (2022). *Uncertainty and response style in latent trait models to assess emotional intelligence of elite swimmers*. Technical report.
- Fabbricatore, R., Iannario, M., Romano, R., & Vistocco, D. (2021). Component-based structural equation modelling for the assessment of psycho-social aspects and performance of athletes. *ASTA advances in statistical analysis*, 1–25.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Forkman, J., Josse, J., & Piepho, H. P. (2019). Hypothesis tests for principal component analysis when variables are standardized. *Journal of Agricultural, Biological and Environmental Statistics*, 24(2), 289–308.
- Furno, M., & Vistocco, D. (2018). *Quantile regression: Estimation and simulation*. Wiley.
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17.
- Gleser, G. C., & Dubois, P. H. (1951). A successive approximation method of maximizing test validity. *Psychometrika*, 16(1), 129–139.
- Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1), 33–51.
- Hastie, T., Tibshirani, R., & Friedman, J. (2013). *Elements of statistical learning: Data mining, inference, and prediction*, 2nd edn. Springer.
- Horn, J. L. (1965). A rational and technique for estimating the number of factors in factor analysis. *Psychometrika*, 30, 179–185.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321–377.
- Janetzko, H., Sacha, D., Stein, M., Schreck, T., Keim, D. A., Deussen, O. (2014). Feature-driven visual analytics of soccer data. In *2014 IEEE conference on visual analytics science and technology (VAST)* (pp. 13–22). IEEE.
- Jolliffe, I. T. (1986). *Principal component analysis*. Springer.
- Josse, J., & Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6), 1869–1879.
- Jowett, S., & Nezlek, J. (2012). Relationship interdependence and satisfaction with important outcomes in coach–athlete dyads. *Journal of Social and Personal Relationships*, 29, 287–301.
- Kajtna, T., Tušák, M., Barić, R., & Burnik, S. (2004). Personality in high-risk sports athletes. *Kineziologija*, 36(1), 24–34.
- Kenny, D. A., Kashy, D. A., & Cook, W. (2006). *Dyadic data analysis*. Guilford.
- Kirkcaldy, B. D. (1982). Personality profiles at various levels of athletic participation. *Personality and Individual Differences*, 3(3), 321–326.
- Koenker, R. (2000). Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *Journal of Econometrics*, 95, 347–374.
- Koenker, R. (2005). *Quantile regression, Econometric society monograph* (Vol. 38). Cambridge University Press.
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50.
- Koenker, R., & D’Orey, V. (1987). Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 36(3), 383–393.
- Laborde, S., Allen, M. S., Katschak, K., Mattonet, K., & Lachner, N. (2020). Trait personality in sport and exercise psychology: A mapping review and research agenda. *International Journal of Sport and Exercise Psychology*, 18(6), 701–716.
- Lebed, F. (2017). *Complex sport analytics*. Taylor & Francis.
- Legg, P. A., Chung, D. H., Parry, M. L., Bown, R., Jones, M. W., Griffiths, I. W., & Chen, M. (2013). Transformation of an uncertain video search pipeline to a sketch-based visual analytics loop. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2109–2118.
- Legg, P. A., Chung, D. H., Parry, M. L., Jones, M. W., Long, R., Griffiths, I. W., & Chen, M. (2012). Matchpad: Interactive glyph-based visualization for real-time sports performance analysis. *Computer Graphics Forum, Wiley Online Library*, 31, 1255–1264.
- Losada, A. G., Therón, R., & Benito, A. (2016). Bkviz: A basketball visual analysis tool. *IEEE Computer Graphics and Applications*, 36(6), 58–68.
- Lyle, J. (2002). *Sports coaching concepts: A framework for coaches’ behaviour*. Routledge.
- Mardia, K. V., Kent, J. T., & Bibby, I. M. (1979). *Multivariate analysis*. Academic Press Inc, Ltd.

- Martens, H., & Næs, T. (1992). *Multivariate calibration*. Wiley.
- McCrae, R. R., & Costa, P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research*, 3rd edn (pp. 159–181). Guilford Press.
- McGarry, T. (2009). Applied and theoretical perspectives of performance analysis in sport: Scientific issues and challenges. *International Journal of Performance Analysis in Sport*, 9(1), 128–140.
- Mckelvie, S. J., Lemieux, P., & Stout, D. (2003). Extraversion and neuroticism in contact athletes, no contact athletes and non-athletes: A research note. *Athletic Insight The Online Journal of Sport Psychology*, 5(3), 19–27.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287–2305.
- Nevill, A., Atkinson, G., & Hughes, M. (2008). Twenty-five years of sport performance research in the Journal of Sports Sciences. *Journal of Sports Sciences*, 26(4), 413–426.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.
- Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4), 974–997.
- Rhind, D. J. A., & Jowett, S. (2011). Working with coach–athlete relationships: Their quality and maintenance. In S. Mellalieu, & S. Hanton (Eds.), *Professional practice in sport psychology: A review* (pp. 219–248). Routledge.
- Rusu, A., Stoica, D., Burns, E., Hample, B., McGarry, K., & Russell, R. (2010). Dynamic visualizations for soccer statistical analysis. In *2010 14th international conference on information visualisation (IV)* (pp. 207–212). IEEE.
- Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1), 1–10.
- Smilde, A. K., Westerhuis, J. A., & Boque, R. (2000). Multiway multiblock component and covariates regression models. *Journal of Chemometrics*, 14(3), 301–331.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.
- Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2), 207–219.
- Wagner, H. M. (1959). Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, 54(285), 206–212.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.