Comparison of Individual Playing Styles in Football

Tianyu Guan, Sumit Sarkar and Tim B. Swartz [*]

Abstract

This paper attempts to identify football players who have a similar style to a player of interest. Playing style is not adequately quantified with traditional statistics, and therefore style statistics are created using tracking data. Tracking data allow us to monitor players throughout a match, and therefore include both "on-the-ball" and "off-the-ball" observations. Having developed style features, tractable discrepancy measures are introduced that are based on Kullback-Leibler divergence in the context of multivariate normal distributions. Examples are provided where a pool of players from the Chinese Super League are identified as having a playing style that is similar to players of interest.

Keywords : big data, Bayesian analyses, divergence measures, player tracking data, spatio-temporal analyses.

[*] T. Guan is Assistant Professor, Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario M3J1P3, S. Sarkar is Professor, XLRI Xavier School of Management, Jamshedpur - 831001, Jarkhand, India, and T. Swartz (corresponding author) is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6, email: tim@stat.sfu.ca. Guan and Swartz have been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work has been carried out with support from the CANSSI (Canadian Statistical Sciences Institute) Collaborative Research Team (CRT) in Sports Analytics. The authors thank Daniel Stenz, former Technical Director of Shandong Luneng Taishan FC who provided the data used in this paper. The authors thank the Associate Editor and three anonymous Reviewers whose comments helped improve the paper.

1 INTRODUCTION

In association football (i.e. soccer), most managers have a sense of how they might improve their side. For example, they may feel the need to acquire a holding midfielder, a centre back or a "number 9". In professional football, determining a team's needs is an important problem, and filling these needs may be accomplished through various means including trades, transfers, player development and drafting.

However, the identification of potential players is a challenging problem as sometimes players do not meet expectations. In this paper, we pose a simple problem - can we identify players who are similar in style to a given player? For example, all teams would welcome a star player like Kylian Mbappe on their team. But there is only one Kylian Mbappe, and he is not likely a realistic acquisition. Therefore, it may be of interest to identify a pool of players who have stylistic characteristics that are similar to Mbappe.

Playing style is an important component of success. Yet, historically, it has been difficult to analyze playing style in invasion sports (e.g., football, hockey, rugby, basketball) since these sports are fluid, with players in continual motion. Motion on the field can be both complex and subtle. The landscape for studying playing style has changed in recent years with the advent of player tracking data. With player tracking data, the location coordinates for every player on the field are recorded frequently (e.g. 10 times per second in football). Player tracking data are particularly important in football because traditional statistics only consider "on-the-ball" activities where players typically have the ball at their feet for less than three minutes per match (https://www.pastemagazine.com/soccer/football/25-johan-cruyff-quotes/). Tracking data also allow us to investigate what players are doing when they are off-the-ball, and this is a component of style.

Of course, player style and player role are slightly different concepts. For example, a manager who has more of a defensive focus may require that a particular midfielder does not play too high up the pitch. This would limit the player's intrinsic attacking style. Our methods consider an individual's playing style subject to the system in which they currently play. It is understood that managers have an influence on player roles. In the characterization of playing style, we have attempted to select playing features that are more personal and are less influenced by player role.

With detailed tracking data, the opportunity to explore novel questions in sport has never been greater. The massive datasets associated with player tracking also introduce data management issues and the need to adopt and develop modern data science methods beyond traditional statistical analyses. Gudmundsson and Horton (2017) provide a review of spatio-temporal analyses that have been used in invasion sports where player tracking data are available. For a review of statistical contributions that have been made across major sports, see the text by Albert et al. (2017). Specifically in football, player tracking data have been successfully utilized by McHale and Relton (2018) to identify key players in a team using network analysis. In this paper, we use tracking data to develop statistics related to the individual playing styles of outfield players.

Most of the literature that involves playing style in football is concerned with style at the team level rather than at the player level. For example, Shen (2022) and Epasinghege Dona and Swartz (2023) examined pace of play in football. However, the most investigated aspect of team playing style in football concerns formations. For example, the book "Inverting the Pyramid" (Wilson 2013) considers the history of football tactics with an emphasis on positional play and player roles. It is now common for television broadcasts to provide graphical statistics that depict the average location of each player during a match. Such information is useful in determining match strategy as it can highlight features such as gaps in player alignment. There are also technical papers that investigate player formation. For example, Shaw and Glickman (2019) used tracking data and clustering methods to determine a team's offensive and defensive formations. This is useful as the fluidity of the sport and changing tactics sometimes make it difficult to distinguish between formations (e.g. 4-4-2 versus 3-5-2). Goes et al. (2021) also identified formations using tracking data and related attacking success to formations. Other papers that address team playing style in football include Hewitt, Greenham and Norton (2016), Lagos-Peñas, Gómez-Ruano and Yang (2017) and Gómez et al. (2018).

Although the literature on team-level playing style in football is abundant, little work has been done on the style of individual players. Decroos and Davis (2020) developed 18-dimensional player vectors and utilized machine learning techniques such as clustering and nearest neighbor analyses to characterize playing style. However, they only considered offensive actions. Decroos, Van Roy and Davis (2021) proposed a mixture-model based soft clustering approach to analyze playing style. A major difference between their approaches and the methods developed in this paper is that the football actions that are considered by Decroos and Davis (2020) and Decroos, Van Roy and Davis (2021) are determined via event data. Event data consist of on-the-ball activities, and hence, omit off-the-ball player movements which comprise the vast majority of play in football. In addition, this paper takes a different approach and introduces statistical discrepancy measures and stochastic models to discriminate players. In the sport of basketball, individual playing style has been investigated by Skinner and Guy (2015) who used tracking data to learn player skills.

Our methods provide a data-driven and objective approach for the stylistic comparison of football players. With scouting, comparisons are typically subjective. With the increasing availability of tracking data, our methods may result in significant time savings when searching the universe of football players. Earlier, we had motivated the problem by considering the search for the next Mbappe. In fact, football is widely considered to be a weak-link sport. That is, a team's success is largely predicated on the strength of its weakest players rather than the strength of its strongest players (a strong-link sport). If this is true, then searching replacements for a team's weakest link does not necessarily involve finding a great replacement player. Rather, a team is simply looking for improvements involving the weakest links. Therefore, it is possible that there are many potential replacements that will help a team improve, and the identification of such players is an exercise of considerable value. Gill and Swartz (2019) characterize the degree of weak and strong links in doubles sports.

In Section 2, we describe the player tracking data that are used in our analyses. We then use the tracking data to develop features that are related to individual playing style. In Section 3, we develop methods from multivariate statistics to identify players who are similar in style to a player of interest. The player characteristics form a multivariate distribution and we use distributional distance measures (i.e. Kullback-Leibler divergence) to assess similarity. Two estimation schemes of player parameters are introduced; a basic approach and a Bayesian approach based on a more complex model. Section 4 identifies players who are similar to a specific player of interest, Marouane Fellaini, and a second player, Graziano Pellè. A reliability analysis and a prior sensitivity analysis are also provided. We conclude with a short discussion in Section 5.

2 DATA

Our data consist of matches from the 2019 season of the Chinese Super League (CSL). The league involved 16 teams where each team played every opponent twice, once at home and once on the road. From these potential 240 matches, we have three missing matches.

Event data and tracking data are collected independently where event data consist of on-the-ball actions such as tackles and passes, and these are recorded along with auxiliary information whenever an "event" takes place. The events are manually recorded by technicians who view film. Both event data and tracking data have timestamps so that the two files can be compared for internal consistency. There are various ways in which tracking data are collected. One approach involves the use of Radio Frequency Identification (RFID) technology where each player and the ball have tags that allow for the accurate tracking of objects. In the CSL dataset, tracking data are obtained from video and the use of optical recognition software. The tracking data consist of roughly one million rows per match measured on 7 variables where the data are recorded every 1/10th of a second. Each row corresponds to either the ball or a particular player at a given instant in time. Of particular interest, a row corresponding to a player contains the (x, y) coordinate for the player which provides their position on the field. With 237 matches of roughly one million rows, the dataset may be thought of as big data.

2.1 Features Involving Style

Football is a team game. Consequently, common player statistics not only reflect their play but also the play of teammates and the opposition. For example, a striker who does not receive service from the midfield is unlikely to score.

In this subsection, we attempt to define features/statistics that are intrinsic to a player's individual style and rely less on the actions of teammates and the opposition. Values of the style statistics that we propose can neither be assessed as good nor bad; they describe qualities such as positioning and role, and are relative to the amount of possession by the player's team. To illustrate the point, there are great players who play further up the pitch and there are great players whose positioning is deeper. Our methods can be applied to all playing positions excluding the goalkeeper.

In Table 1 (see back of manuscript), we propose variables that describe playing style. The calculation of these features is facilitated through the use of tracking data since many of the variables relate to "off-the-ball" activities which can only be measured by knowing the whereabouts of the player at all times during a match. To our knowledge, the development of individual style features in football is novel.

In Table 1, we have divided the variables into three broad categories, dividing the field into an offensive third, a middle third and a defensive third. It is well-known that player responsibilities change according to these categories. It is also well-known that responsibilities change according to possession (team of interest versus the opponent), and hence, the variables are also divided according to possession.

We have developed the features in Table 1 by assessing the options that are available to players. This has been done using domain knowledge of football. For example, when a player is in possession of the ball, the player has three options: shoot, pass or dribble. When a teammate has the ball, the player has choices concerning movement. When the opponent has possession, we have developed features by considering movement, interceptions and tackles. For the positional variables, $x_1, x_8, x_{12}, x_{18}, x_{22}$ and x_{26}, we recognize the symmetry of football. That is, the responsibilities and actions on the right side of the field tend to mirror those on the left side of the field. Therefore, we have flipped right-sided locations to the left side. With this adjustment, a left fullback will not look too different from a right fullback, for example, in terms of positioning. Specifically, to develop positional features involving styles, we create the following coordinate system. We first set the middle of the pitch as the origin. The x-axis and y-axis are parallel to the touch lines and end lines, respectively. Each team has its own coordinate system with the x-axis direction being the team's attacking direction. To flip right-sided locations to the left side, negative y-axis values are adjusted to be positive.

An important feature of the statistics proposed in Table 1 is that they have been carefully developed so that stylistic comparisons are sensible. For example, the use of percentages for some of the statistics allows us to sensibly compare players from weaker teams with those from stronger teams. For example, a stronger team will typically have more shots, and therefore, a player's percentage of a team shots is more indicative of style than the player's total number of shots. Also, the stylistic variables permit the comparison of players from different teams and leagues since the features are defined relative to team play and are not absolute measures. Now, a case can be made that a player's style may vary according to the demands imposed by their manager. However, we take the view that the statistics reflect a player's current style, and not an alternative style that a player may or may not be able to adopt. The ability to compare the style of players from different leagues is potentially of great benefit since football is a world game, and it is impossible to have comprehensive scouting throughout the world.

We wish to emphasize the utility of the proposed variables in Table 1 that are based on expected possession value EPV (i.e. x_2, x_4, x_{13}, x_{15}). The idea behind EPV is that there are positions on the field that are more threatening from a goal-scoring perspective. For example, having possession six feet directly in front of goal provides a better goal-scoring opportunity than 60 feet from goal near the side of the pitch. Whereas different factors are utilized in the various EPV formulae, the most simple derivation of EPV at a particular region of the field involves the ratio of the number of goals scored from that location by the number of shots taken from that location. The EPV statistics in Table 1 are intended to quantify the attacking quality of actions. The calculation of EPV was made publicly available by Shaw (2019). When making a pass, we consider the increase in EPV from where the pass was initiated to its final destination. We modify the EPV covariate of a player by setting it equal to zero if the player is in an offside position. EPV provides us with meaningful information concerning player movement.

All of the variables in Table 1 have been standardized such that they have zero mean and unit variance. This is a common pre-processing step in multivariate statistics such that large variables do not have undue influence in regression procedures. The variables proposed in Table 1 may obviously be modified. However, our main objective is the development of a measure of player discrepancy (Section 3) which is an amalgam of the features in Table 1.

The data management issues and computational tasks associated with the collection of the features in Table 1 are considerable. For a given player, we need to step through all of the frames of the tracking data for all matches throughout the season. Possession needs to be noted and statistics accumulated in each frame. The computational time associated with this task is roughly two hours on a laptop computer.

Summary statistics for the features in Table 1 are presented in Table 2. These statistics are calculated across all players and all matches, and are presented prior to standardization. We observe that, players generally position themselves further back on the pitch when the ball is in the defensive and middle thirds, compared to the offensive third. This pattern is illustrated by the smaller values of x coordinates for x_{22} and x_{12} compared to x_1, as well as the decreased x coordinates values for x_{26} and x_{18} relative to x_8. Furthermore, it is observed that x values of player's positional variables are smaller when the opponent has possession, with x values for x_8, x_{18} and x_{26} smaller than those of x_1, x_{12} and x_{22}. This suggests that when the opponent is in possession of the ball, players transition between offensive and defensive roles, moving back to support defense. As previously discussed in Section 2.1, locations on the right side have been mirrored to the left, leading to positive values of all positional y values. In addition, we observe a decreasing trend in x_9, x_{19} and x_{27}. This shows a tactical response where players get closer to their nearest opponent when the ball gets closer to the player's own goal.

3 METHODS

Having developed the playing style statistics in Table 1, the methods proposed here for assessing player similarity are straightforward. Recall that the statistics listed in Table 1 are obtained for each match.

We refer to the foundational player of interest as player zero. For this player, define the k-th match vector $x_k^{(0)} = (x_{k1}^{(0)}, \ldots, x_{kn}^{(0)})'$ for matches $k = 1, \ldots, m^{(0)}$ as multivariate normal with dimension $n = 39$ whose component statistics are described in Table 1. Note that the dimension does not match the number of table entries; this is explained by noting that the 9 positional variables $x_1, x_3, x_8, x_{12}, x_{14}, x_{18}, x_{22}, x_{23}$ and x_{26} have both an x and y coordinates. We write

$$x_k^{(0)} \sim \text{Normal}_n(\mu^{(0)}, \Sigma^{(0)}) \tag{1}$$

where independence is assumed across the matches. Multivariate normal distributions are particularly attractive for this application since they provide interpretable correlation parameters via Σ that allow us to model the relationships between the variables in Table 1. For example, when making a decision between passing or shooting, only one option can be chosen. Therefore, passing and shooting are negatively correlated.

3.1 Discrepancy Measures based on Kullback-Leibler Divergence

The mean vector $\mu^{(0)} = (\mu_1^{(0)}, \ldots, \mu_n^{(0)})'$ of the foundational player is estimated via $\mu_i^{(0)} = \sum_{k=1}^{m^{(0)}} x_{ki}^{(0)}/m^{(0)}$ and the covariance matrix $\Sigma^{(0)} = (\sigma_{ij}^{(0)})$ is estimated via $\sigma_{ij}^{(0)} = \sum_{k=1}^{m^{(0)}} (x_{ki}^{(0)} - \mu_i^{(0)})(x_{kj}^{(0)} - \mu_j^{(0)})/m^{(0)}$.

We also have N additional players whom we would like to compare against player zero. For the j-th player, we use similar notation, and for the k-th match, we write

$$x_k^{(j)} \sim \text{Normal}_n(\mu^{(j)}, \Sigma^{(j)}) \tag{2}$$

where independence across matches is again assumed. For these comparison players, the mean vectors and covariance matrices are estimated similarly.

Our goal in comparing players is based on the recognition that there is variability in performance. Therefore, rather than directly compare observed statistics, it seems sensible to compare distributions from which the statistics arise. Accordingly, we use the Kullback-

Leibler divergence measure (Kullback and Leibler 1951) which describes the discrepancy in the distribution of $x_k^{(j)}$ from the distribution of $x_k^{(0)}$. In general, the Kullback-Leibler divergence of a distribution with density f relative to a distribution with density g is given by $\mathrm{KL}(f \mid g) = \int f(x) \ln(f(x)/g(x)) \, dx$. In the context of the multivariate normal distributions given by (1) and (2), the measure is given by

$$\mathrm{KL}(j \mid 0) \;=\; \frac{1}{2}\left[(\mu^{(j)} - \mu^{(0)})'(\Sigma^{(0)})^{-1}(\mu^{(j)} - \mu^{(0)}) + \mathrm{tr}((\Sigma^{(0)})^{-1}\Sigma^{(j)}) + \ln\frac{|\Sigma^{(0)}|}{|\Sigma^{(j)}|} - n\right] \quad (3)$$

Proof: Using $\langle\rangle$ to denote the integration operator with respect to f and using properties of the trace function, we have

$$\begin{aligned}
\mathrm{KL}(j \mid 0) &= \left\langle \ln \frac{|\Sigma^{(j)}|^{-1/2}}{|\Sigma^{(0)}|^{-1/2}} \; \frac{\exp\left[-\frac{1}{2}(x-\mu^{(j)})'(\Sigma^{(j)})^{-1}(x-\mu^{(j)})\right]}{\exp\left[-\frac{1}{2}(x-\mu^{(0)})'(\Sigma^{(0)})^{-1}(x-\mu^{(0)})\right]} \right\rangle \\
&= \frac{1}{2}\left\langle \ln\frac{|\Sigma^{(0)}|}{|\Sigma^{(j)}|} \; - (x-\mu^{(j)})'(\Sigma^{(j)})^{-1}(x-\mu^{(j)}) + (x-\mu^{(0)})'(\Sigma^{(0)})^{-1}(x-\mu^{(0)}) \right\rangle \\
&= \frac{1}{2}\left[\ln\frac{|\Sigma^{(0)}|}{|\Sigma^{(j)}|} \; - \left\langle \mathrm{tr}((\Sigma^{(\mathrm{j})})^{-1}(\mathrm{x}-\mu^{(\mathrm{j})})(\mathrm{x}-\mu^{(\mathrm{j})})')\right\rangle + \left\langle \mathrm{tr}((\Sigma^{(0)})^{-1}(\mathrm{x}-\mu^{(0)})(\mathrm{x}-\mu^{(0)})')\right\rangle\right] \\
&= \frac{1}{2}\left[\ln\frac{|\Sigma^{(0)}|}{|\Sigma^{(j)}|} \; - \mathrm{tr}((\Sigma^{(\mathrm{j})})^{-1}(\Sigma^{(\mathrm{j})})) + \left\langle \mathrm{tr}(\Sigma^{(0)})^{-1}(\mathrm{x}-\mu^{(\mathrm{j})}+\mu^{(\mathrm{j})}-\mu^{(0)})(\mathrm{x}-\mu^{(\mathrm{j})}+\mu^{(\mathrm{j})}-\mu^{(0)})'\right\rangle\right] \\
&= \frac{1}{2}\left[\ln\frac{|\Sigma^{(0)}|}{|\Sigma^{(j)}|} \; - n + \mathrm{tr}(\Sigma^{(0)})^{-1}(\Sigma^{(\mathrm{j})} + (\mu^{(\mathrm{j})}-\mu^{(0)})(\mu^{(\mathrm{j})}-\mu^{(0)})')\right] \\
&= \frac{1}{2}\left[\ln\frac{|\Sigma^{(0)}|}{|\Sigma^{(j)}|} \; - n + \mathrm{tr}((\Sigma^{(0)})^{-1}\Sigma^{(\mathrm{j})}) + (\mu^{(\mathrm{j})}-\mu^{(0)})'(\Sigma^{(0)})^{-1}(\mu^{(\mathrm{j})}-\mu^{(0)})\right] \; . \quad Q.E.D.
\end{aligned}$$

Therefore, our procedure is conceptually and computationally simple. We consider a player of interest (player zero) and we specify players j whom we wish to compare against player zero. Then, following (3), $\mathrm{KL}(j \mid 0)$ is calculated for all potential players $j = 1, \ldots, N$. We then rank the players where small values of $\mathrm{KL}(j \mid 0)$ indicate greater similarity involving player j to player zero.

Now, consider player zero whom we believe to be excellent, and players j_1 and j_2. Suppose further that $\mathrm{KL}(j_1 \mid 0) < \mathrm{KL}(j_2 \mid 0)$. We emphasize that this does not imply that player j_1 is a better player than j_2. If we look at the descriptions of the stylistic variables in Table 1, we see that none of these statistics definitively describe quality. There is no directional meaning in the statistics in terms of excellence. For example, one player could play further up the pitch than another player. Instead, $\mathrm{KL}(j_1 \mid 0) < \mathrm{KL}(j_2 \mid 0)$ implies only that player j_1 is more similar in style to player zero than is player j_2. The quality of players is a separate issue.

3.2 Discrepancy Measures based on a Bayesian Model

In the construction of the multivariate normal distributions (1) and (2), it may be sensible to assume that the covariance matrices $\Sigma^{(j)},\ j = 0, \ldots, N$ have an underlying similarity. For example, it is reasonable to assume that players who attack and play further up the field are likely to take more of a team's shots, and we would expect this pattern to hold across all players.

To implement this idea, we consider a hierarchical model where we retain the distributional assumptions (1) and (2), but further assume that the covariance matrices $\Sigma^{(0)}, \ldots, \Sigma^{(N)}$ form a sample from an underlying distribution. That is, the Σ's may be different but they are all related in the sense that they arise from a common distribution.

Hierarchical models are conveniently handled in a Bayesian framework. Therefore, we complete the Bayesian specification by introducing the prior distributions

$$\Sigma^{(j)} \sim \text{Wishart}^{-1}(V, v) \tag{4}$$

and

$$\mu^{(j)} \propto 1$$

for $j = 0, \ldots, N$.

The n by n matrix V and the degrees of freedom v are hyperparameters in the Bayesian framework. We specify the hyperparameters in Section 3.3 and we provide a prior sensitivity analysis in Section 4.6.

To analyze the proposed Bayesian model, we note that posterior estimates for $\mu^{(j)}$ and $\Sigma^{(j)}$, $j = 0, \ldots, N$ cannot be obtained analytically. Alternatively, a sampling approach is used based on a Markov chain Monte Carlo (MCMC) implementation, where we estimate the posterior means of $\mu^{(j)}$ and $\Sigma^{(j)}$ for $j = 0, \ldots, N$ by averaging the MCMC output. Here, MCMC is carried out using a Gibbs sampler. We derived the corresponding full conditional distributions and implemented the Gibbs sampler in the R programming language.

To summarize the approach, posterior samples are generated using MCMC. The samples are then used to obtain parameter estimates which in turn allows us to calculate the discrepancy measures given by equation (3). The discrepancy measures allow us to compare the playing styles between player zero and player j, $j = 1, \ldots, N$.

3.3 Specification of the Hyperparameters in Section 3.2

We now consider the specification of the hyperparameters V and v introduced in the prior distribution (4).

A property of the inverse Wishart distribution is that $\mathrm{E}(\Sigma^{(j)}) = V/(v - n - 1)$ for $j = 0, \ldots, N$. Therefore, we let $V = (v - n - 1)S$ where S is the sample covariance matrix based on data from all players. It follows that $\mathrm{E}(\Sigma^{(j)}) = S$ such that all players have a common expected covariance matrix. This specification may be considered empirical Bayes since the hyperparameter V is a function of the data (Carlin and Louis 2000).

To complete the hyperparameter specification, we need to set the degrees of freedom parameter v. We prefer a prior distribution that is diffuse such that it does not overinfluence the data. With this in mind, it is a property of the inverse Wishart distribution that $\mathrm{Var}(\Sigma^{(j)})$ is finite and decreasing for $v > n + 3$. We therefore set $v = n + 4$ since this is the integer that leads to minimum but finite variance. Larger values of v would lead to posterior estimates of $\Sigma^{(j)}$ that are increasingly similar.

4 EXAMPLES

We first consider Marouane Fellaini of Shandong Luneng Taishan as a player of interest (i.e., player zero). Although the CSL is not one of the top leagues in the world, Fellaini is well-known as a former Belgium international, and he is also known for his seven seasons (2013-2019) playing with Manchester United of the English Premier League. He is a visible player on the pitch, standing 6' 5". His physical characteristics suggest that he may have an unusual playing style.

As the CSL allows only a fixed numbers of internationals, players such as Fellaini have commanded high salaries which adds to their notoriety. Should a player like Fellaini leave Shandong Luneng, there would be considerable interest in his replacement. Note that Fellaini has a playing style that contributes to his uniqueness; for example, it is believed that Fellaini is aggressive (see https://theexecutionersbong.wordpress.com/2012/03/01/is-marouane-fellaini-the-top-flights-most-aggressive-midfielder/).

In seeking potential replacements for Fellaini, we consider the $N = 28$ midfielders in the CSL who each played at least 1500 minutes and 20 matches during the 2019 season.

4.1 Consideration of the Features in Table 1

The discrepancy measures introduced in Section 3.1 and Section 3.2 are omnibus statistics that take into account the totality of the features listed in Table 1.

It may also be interesting to investigate the features on an individual basis. We therefore refer to Fellaini as the foundational player (player zero) with average feature vector $\bar{x}^{(0)} = (\bar{x}_1^{(0)}, \ldots, \bar{x}_n^{(0)})'$ averaged over Fellaini's matches. Likewise, the average feature vector for player j is given by $\bar{x}^{(j)} = (\bar{x}_1^{(j)}, \ldots, \bar{x}_n^{(j)})'$, $j = 1, \ldots, N$. To get a sense of the degree to which Fellaini differs from the other players with respect to feature l, $l = 1, \ldots, n$, we define

$$b_l = \sum_{j=1}^{N} (\bar{x}_l^{(0)} - \bar{x}_l^{(j)})^2 \ . \tag{5}$$

We identify the five largest values b_l in (5), and for feature l, we calculate the extent

$$\bar{x}_l^{(0)} - \bar{x}_l^{(j)} \tag{6}$$

to which player j differs from Fellaini with respect to feature l.

In Figure 1, we provide boxplots using the statistics given in (6) for the five features of interest. The five features with largest b_l values are x_2 and y coordinates of x_{18}, x_{26}, x_8, and x_1. Fellaini played most of his matches as a central midfielder who operates in the middle of the field, focusing on both defense and attack. Among the five features, x_2 and the y coordinate of x_1 characterize a midfielder's offensive style, while x_{18}, x_{26}, and x_8 relate more closely to defensive positioning. The most prominent statistic x_2 is the average EPV when the ball is in the offensive third when the team is in possession. Fellaini's x_2 statistic is much larger than almost all other midfielders; this suggests that he gets himself into dangerous positions more often. Therefore, Fellaini makes great contributtions to his team's offensive dynamics. The only player who has a larger average x_2 is Paulinho, another well-known player in the CSL. This indicates a similarity in Fellaini and Paulinho's aggressive midfield roles. We note that all prominent positional features are expressed via the y coordinates, the extent to which players move downfield. Fellaini played most matches as a central midfielder, and therefore, his y coordinates shown in Figure 1 are smaller than those of most players.

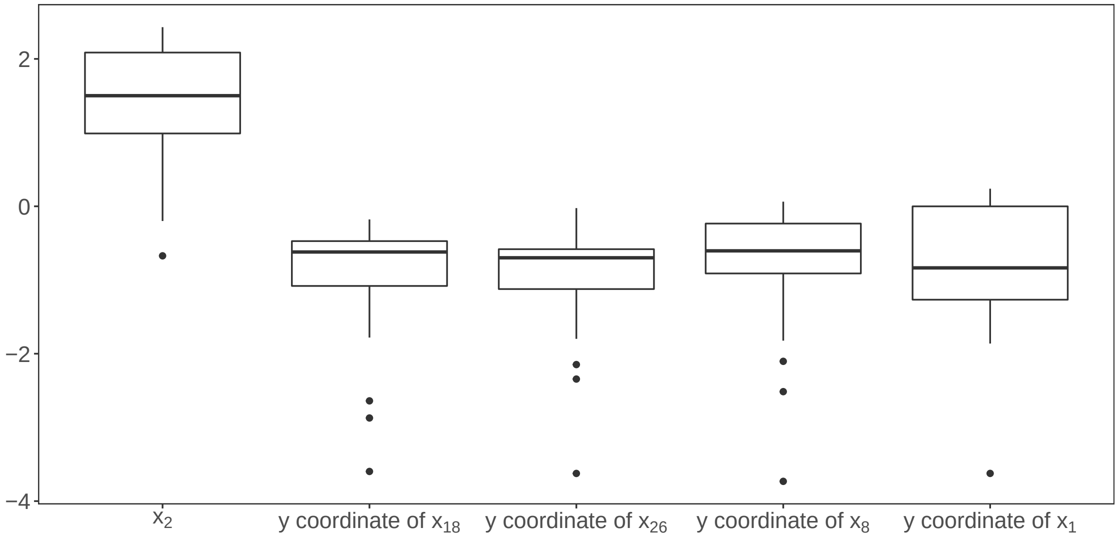

Figure 1: Boxplots of the five most prominent features where players differ from Fellaini.

4.2 The Multivariate Normal Framework

In Section 3.1, we proposed a statistic to assess the similarity of playing style between a player of interest and candidate players. The idea was extended in Section 3.2 where a Bayesian model was proposed that recognized similarity in the covariance matrices $\Sigma^{(j)}$ across players.

An assumption in both approaches is that match vectors follow multivariate normal distributions. The multivariate normal distribution is appealing as we expect the distribution of features to have a concave shape; i.e., as we move away from the mean value of a feature, fewer players will exhibit extreme values of the feature. Also, the multivariate normal is a tractable distribution which leads to a closed form expression (3) for the Kullback-Leibler discrepancy measure. In addition, the multivariate normal distribution has a covariance structure which leads to interpretability when comparing features.

We use the Henze–Zirkler test (Henze and Zirkler 1990) to justify the validity of the multivariate normality assumptions for Fellaini (player 0) and the 28 comparison midfielders. The Henze-Zirkler test statistic is dependent on the sample covariance matrix. However, when the sample size is small or close to the variable dimension ($m^{(j)} \leq n$), the sample covariance matrix tends to singularity, and hence, the Henze–Zirkler test cannot be directly used. In the case of Fellaini, the sample size is $20 \leq m^{(j)} \leq 30$ and the variable dimension is $n = 39$. One possible approach to solving this issue is to divide matches into halves. This results in 7 players having at least 40 complete halves of data (see Section 4.3 for more details). The Henze–Zirkler test yields p-values that are greater than 0.12 for each of these 7 players, which implies that the normality assumption for match vectors is reasonable.

As a further check on the adequacy of normality, we have produced histograms to investigate modality and skewness of the underlying features. Recall that normal distributions are unimodal and symmetric. In Figure 2, we provide histograms for a sample of standardized features taken from Table 1. Figure 2 provides no evidence of such violations from normality.

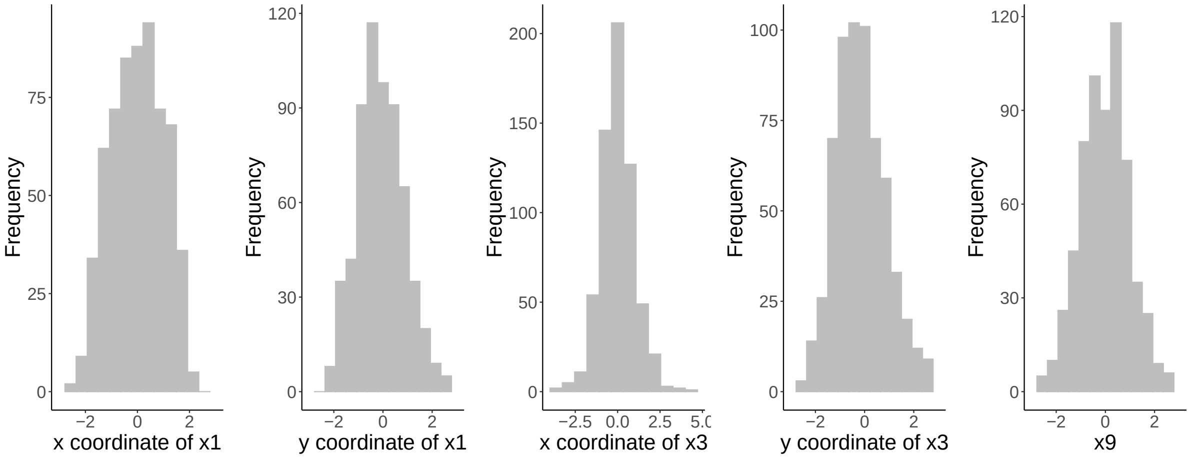

Figure 2: Histograms for a subset of five standardized features corresponding to midfielders.

4.3 Analysis using the Basic Approach

We first consider the approach presented in Section 3.1 which does not specify any relationship between the covariance matrices $\Sigma^{(j)}$. Recall, we have matrices $\Sigma^{(j)}$ that are n by n (where $n = 39$) corresponding to the style statistics presented in Table 1. In this application, the midfielders of interest played a minimum of $m^{(j)} = 20$ matches. Now, it is a fact of linear algebra that the estimated matrix $\Sigma^{(j)}$ is not full rank (i.e., not positive definite) if $m^{(j)} \leq n$. This is problematic for the methods of Section 3.1 since this leads to a zero determinant for $\Sigma^{(j)}$ in equation (3). Consequently, the style statistic $\mathrm{KL}(j \mid 0)$ is undefined. A potential solution to this difficulty is to divide the matches into halves and collect data for each half. This effectively doubles the number of observations for each player such that $m^{(j)} \leq n$ is avoided.

Although the "fix" that divides matches into halves is appealing, we have a remaining problem regarding the estimation of $\Sigma^{(j)}$. Some players have incomplete data, which can also result in covariance matrices that are not positive-definite. For example, Zhongguo Chi is a midfielder from the Beijing Sinobo Guoan Football Club. He played 23 matches as a midfielder. For two matches, he didn't play during the first half. Also, in 8 halves, he didn't have any successful passes in the offensive third, which led to NAs for the variables x_3 and x_4. As a result, Zhongguo Chi had at most $m = 2(23) - 2 - 8 = 36$ complete halves of data, and since $m = 36 \leq n = 39$, Chi's covariance matrix is not positive-definite. In fact, among the 28 midfielders, only 7 players had at least $n = 40$ complete halves of data. The results for these 7 players are provided in Table 3.

Player	Team	$\mathrm{KL}(j \mid 0)$
Augusto Renato	Beijing Sinobo Guoan	1.75
Paulinho	Guangzhou Evergrande Taobao	2.11
Marek Hamšík	Dalian Yifang	3.11
Xi Wu	Jiangsu Suning	4.24
Javier Mascherano	Hebei China Fortune	5.36
Hang Li	Wuhan Zall	6.60
Dingyang Zhou	Henan Construction	6.81

Table 3: CSL midfielders and the discrepancy measure $\mathrm{KL}(j \mid 0)$ with respect to Marouane Fellaini of Shandong Luneng. The methods are based on the approach described in Section 3.1 where we have effectively doubled the number of observations by dividing matches into halves. For ease of presentation, the $\mathrm{KL}(j \mid 0)$ entries in the table ought to be multiplied by 10^3.

Although the approach described in Section 3.1 is simple and is a good starting point, we recognize the limitation of dividing matches into halves to effectively double the number of observations. First, there is less data in halves than in full matches, and this introduces added variability in the Kullback-Leibler statistic. Second, it is possible that second half tactics and style may differ from the first half. Of course, with larger datasets (e.g. multiple years worth of player data), there may be no need to divide matches into halves, and therefore the methods of Section 3.1 can be easily implemented.

4.4 Analysis using the Bayesian Model

4.4.1 Marouane Fellaini

We now consider the preferred approach described in Section 3.2 which introduces similarity between the covariance matrices $\Sigma^{(j)}$. Table 4 provides a ranked order of the Kullback-Leibler statistic (3) which compares CSL midfielders to Fellaini. The table includes the 7 players out of the 28 midfielders who are most similar to Fellaini. We observe that Paulinho from Guangzhou Evergrande Taobao is most similar in style to Fellaini. Given Fellaini's notoriety and excellence, it might be assumed that he does a lot of things "correctly" on the field in terms of style. One might therefore draw the connection that Paulinho also does good things on the field, and this is supported by his selection over many years (2011-2018) to the Brazilian National Team. Like Fellaini, Paulinho is also a marquee foreign player in the CSL. Paulinho has been described as "bringing energy and power" to football (del Rio 2017) which is also reminiscent of Fellaini's aggressive off-the-ball style (Wu and Swartz 2023). Among the 22 matches that Fellaini played during the season, he was characterized as a central midfielder for 18 matches, and a holding central midfielder for 4 matches. As expected, the majority of the 7 players who are most similar to Fellaini in Table 4 often played as central midfielders. For example, Paulinho, Zheng, Hamšík, Wang, and Wu played as central midfielders for 88.46%, 85.71%, 78.26%, 80.00%, and 92.31% of their matches, respectively.

It is interesting to compare Table 3 with Table 4. Three of the top four players who are most similar to Fellaini in Table 3 also appear in the Table 4 list. However, there are some troubling aspects concerning the comparison of Table 3 and Table 4. For example, the most similar player to Fellaini in Table 3 (Renato) does not appear in Table 4. Also, the scale of the KL$(j \mid 0)$ values in Table 3 are roughly two magnitudes larger than those in Table 4. We suggest that the Table 3 values are less trustworthy. The greater variability of dividing game data into halves is one reason for the inflated discrepancy measures. The other reason is that the estimated covariance matrices from Table 3 are much too variable and cause great differences in style statistics between players and Fellaini. In Section 4.5, we demonstrate that the methods of Section 3.2 leading to Table 4 are reliable.

Player	Team	KL$(j \mid 0)$
Paulinho	Guangzhou Evergrande Taobao	32.37
Xiaogang Zhu	Dalian Yifang	38.69
Kaimu Zheng	Tianjin Teda	39.06
Marek Hamšík	Dalian Yifang	39.38
Qiuming Wang	Hebei China Fortune	39.64
Xi Wu	Jiangsu Suning	40.71
Bowen Huang	Guangzhou Evergrande Taobao	43.73

Table 4: CSL midfielders and the discrepancy measure KL$(j \mid 0)$ for the 7 most similar players to Marouane Fellaini. The methods are based on the Bayesian approach described in Section 3.2 where similarity between the covariance matrices $\Sigma^{(j)}$ is implemented based on the chosen hyperparameters of Section 3.3.

4.4.2 Graziano Pellè

As a second example, we consider an analysis involving Graziano Pellè of Shandong Luneng. Unlike Fellaini who is a midfielder, Pellè is a forward. Pellè is also well-known, having played internationally for Italy and also for Southampton of the English Premier League during the years 2014-2016.

For this analysis, Pellè is now the foundational player zero. We consider the preferred Bayesian analysis using the methods described in Section 3.2. To obtain a pool of comparison players for Pellè, we consider forwards in the CSL who each played at least 1500 minutes during the 2019 season. This yields a subset of $N = 21$ players against whom we can compare Pellè.

In Table 5, we provide the results involving the 7 players out of the 21 forwards who are most similar to Pellè. Zahavi from Guangzhou R&F is most similar in style to Pellè. They shared similarities in many aspects. For example, they both excelled at goal scoring and creating spaces for teamates. During the 2019 season, Pellè scored 17 goals, sixth in the league, while Zahavi was the top scorer in the league with a total of 29 goals. Other players in Table 5 also had exceptional scoring ability. Elkeson, Kardec, and Wagner scored 18, 14, and 13 goals, and they ranked third, eighth, and ninth in goalscoring, respectively. Pellè played as a central forward in all his matches. Similarly, every player in Table 5 played at least 90% of their matches as a central forward. Elkeson is listed with two teams due to a transfer during the season.

Player	Team	$\mathrm{KL}(j \mid 0)$
Eran Zahavi	Guangzhou R&F	37.86
Yang Xu	Tianjin Tianhai	38.72
Elkeson	Shanghai SIPG/Guangzhou Evergrande Taobao	43.32
Kardec	Chongqing SWM	44.90
Yuning Zhang	Beijing Sino Guoan	50.34
Makhete Diop	Beijing Renhe	54.36
Dandro Wagner	Tianjin Teda	56.46

Table 5: CSL forwards and the discrepancy measure $\mathrm{KL}(j \mid 0)$ for the 7 most similar players to Graziano Pellè. The methods are based on the Bayesian approach described in Section 3.2 where similarity between the covariance matrices $\Sigma^{(j)}$ is implemented based on the chosen hyperparameters of Section 3.3.

4.5 Reliability Analysis

Reliability is a critical component of analyses. When introducing new methods, it is important to demonstrate that similar results are obtained when data are collected on different occasions.

To investigate the reliability of the discrepancy statistic (3) using the Bayesian methods of Section 3.2, we divide Fellaini's season (22 matches) into 10 home matches and 12 road matches. We then introduce a Fellaini statistic $x^{(F_1)}$ based on five randomly selected home matches and six randomly selected road matches. A complementary Fellaini statistic $x^{(F_2)}$ is then obtained based on the remaining matches.

In Table 6, we calculate $\mathrm{KL}(j \mid F_1)$ for $j = F_2$ and for the four midfielders who are most similar to F_1. We observe that $\mathrm{KL}(F_2 \mid F_1)$ is the smallest entry which demonstrates the reliability of the approach. Note that Fellaini's style should be more similar across his own matches than when he is compared to other players. Note also that the three of the four non-Fellaini players in Table 6 also appear in Table 4.

Player	Team	KL($j \mid F_1$)
$x^{(F_2)}$	Shandong Luneng	24.29
Paulinho	Guangzhou Evergrande Taobao	29.81
Qiuming Wang	Hebei China Fortune	30.30
Xi Wu	Jiangsu Suning	33.42
Huikang Cai	Shanghai SIPG	33.67

Table 6: CSL midfielders and the discrepancy measure KL($j \mid F_1$) corresponding to the reliability analysis of Section 4.5. The players listed are Fellaini F_2 and the four other players who are most similar to Fellaini F_1.

Besides Fellaini, we also conducted a reliability check for the 28 midfielders. For each player, we treated him as player 0 and then divided his season into home and road matches. We then randomly selected half of the home and away matches to produce the statistic $x^{(F_1)}$ with the other half of the matches providing the statistic $x^{(F_2)}$. We calculated the Kullback-Leibler divergence to rank the players. Our results show that in 82.76% of the cases, the player's style is most similar to his own. Decroos and Davis (2020) and Decroos, Van Roy and Davis (2021) did similar experiments. Whereas we split the matches in one season for the reliability check, they compared a player's style across two seasons in the top European soccer leagues. In Decroos and Davis (2020), they found that a player's style matched his own 38.2% of the time. The subsequent study by Decroos, Van Roy and Davis (2021) increased the percentage to 48.2%.

4.6 Prior Sensitivity Analysis

It is interesting to investigate the degree to which the discrepancy measure developed in Section 3.2 depends on the choice of the hyperparameters selected in Section 3.3. Recall that the hyperparameters V and v in (4) were chosen with the intention that the data impact the posterior more than the prior.

In Table 7, we expand the results from Section 4.4.1 based on the preferred prior developed in Section 3.3. The expansion now includes the 10 players out of the 28 CSL midfielders who are most similar to Fellaini based on the preferred prior. The results are given in the first data column of Table 7. We concentrate on the ranks of the players (in terms of similarity to Fellaini) as we recognize that the Kullback Liebler statistic (3) is based on samples from the posterior, and is therefore sensitive to the prior specification.

For comparison with the preferred prior, we first consider the alternative prior $\Sigma^{(j)} \sim \text{Wishart}^{-1}(I, n+2)$ for $j = 0, \ldots, N$ where the covariance matrices are assumed statistically independent. Unlike the prior specification in Section 3.3, this prior does not depend on the data and therefore the resultant procedure is not empirical Bayes. This prior may be viewed as a default prior which is even more diffuse. The results of the Fellaini similarity study based on the diffuse prior are given in the second data column of Table 7.

When comparing the more diffuse prior to the preferred prior, we observe that there is rough agreement in the ranks. That is, players that are deemed similar to Fellaini using the preferred prior are also similar to Fellaini using the more diffuse prior. The agreement in the two analyses is reassuring since both frameworks are intended to rely strongly on the data.

As a contrast, we now consider the use of a more informative prior. Accordingly, we remove the variability associated with the prior $\Sigma^{(j)}$ and instead set $\Sigma^{(j)} = S$ for $j = 1, \ldots, N$ where S is the sample covariance matrix based on data from all players. This strict specification may be reasonable as it simply states that the stylistic variables from Table 1 have the same covariance structure for all players. In this case, only $\mu^{(0)}, \ldots, \mu^{(N)}$ are generated from the posterior, and the discrepancy statistic (3) reduces to

$$\text{KL}(j \mid 0) = \frac{1}{2}\left[(\mu^{(0)} - \mu^{(j)})' S^{-1} (\mu^{(0)} - \mu^{(j)})\right] .$$

The third data column of Table 7 provides the results associated with the more informative prior. Again, we see that the top few players deemed similar to Fellaini using the preferred prior are also similar to Fellaini using the more informative prior. As we look beyond the top-ranked players, variations in the rankings in the lists become evident, which can be attributed to the influence of the prior. The greates discrepancy in the analyses concerns Qiuming Wang; he is ranked 5th, 7th and 14th in similarity to Fellaini using the preferred prior, the more diffuse prior and the more informative prior, respectively.

Therefore, the sensitivity analysis provides comparisons to the preferred prior using both a diffuse and an informative prior. The comparisons show that while the top few players' rankings are robust to changes in the prior, the rankings of other players are more sensitive to these assumptions. The general agreement in the ranks for the top few players provides us with confidence in the approach. It is these players (the most similar ones) who are the primary focus of our investigation.

Again, we believe that the prior suggested in Section 3.3 is the preferred prior as it imposes greater similarity in the covariance matrices than the diffuse prior and assumes less than the informative prior.

Player	Preferred Prior	More Diffuse Prior	More Informative Prior
	Rank (KL stat)	Rank (KL stat)	Rank (KL stat)
Paulinho	1 (32.4)	1 (82.8)	2 (5.5)
Xiaogang Zhu	2 (38.7)	4 (99.5)	3 (5.6)
Kaimu Zheng	3 (39.1)	3 (97.2)	1 (4.5)
Marek Hamšík	4 (39.4)	2 (96.0)	5 (5.8)
Qiuming Wang	5 (39.6)	7 (103.4)	14 (6.9)
Xi Wu	6 (40.7)	5 (102.2)	8 (6.2)
Bowen Huang	7 (43.7)	6 (102.6)	9 (6.4)
Huikang Cai	8 (44.6)	9 (112.1)	12 (6.7)
Zhuoyi Feng	9 (47.5)	14 (122.3)	6 (5.9)
Xinli Peng	10 (47.9)	10 (113.5)	15 (7.2)

Table 7: The first data column provides the ranks and the Kullback-Leibler statistics corresponding to the 10 CSL midfielders who are most similar to Fellaini based on the preferred prior of Section 3.3. The second and third data columns provide the corresponding entries based on the more diffuse prior and the more informative prior, respectively, as discussed in Section 4.6.

5 DISCUSSION

This paper develops methods that identify players who have styles that are similar to a specified player of interest. We utilize tracking data to construct player style statistics. These style statistics are assumed to follow multivariate normal distributions. To assess the similarity between players, we utilize the Kullback-Leibler divergence measure. The parameters of multivariate normal distributions are estimated by MCMC methods in a Bayesian framework.

Identifying players who have similar playing style to a specific player is an important problem for teams that are attempting to fill their rosters. Admittedly, the development of the methods proposed in this paper are demanding. However, given the software, teams and data providers (who have access to tracking data) can make stylistic comparisons between players nearly an automatic procedure. The programs can sift through many players in various leagues and develop a pool of players who have similar styles to a player of interest. Consequently, this saves time compared to watching videos or live matches involving potential players.

Our work not only is useful in searching for a player's replacement but may be utilized by clubs who are financially judicious in the transfer market. Professional football clubs in Europe are now constrained by the Financial Fair Play (FFP) rule of the Union of European Football Associations (UEFA), which imposes a cap on a club's spending where the cap is determined by the club's revenues. This regulation requires the clubs to spend judiciously, and consequently, the clubs should seek value for money in the transfer market while recruiting players. Estimating value for money in player transfers, McHale and Holmes (2023) demonstrate that major clubs like FC Barcelona and Manchester United FC performed poorly in the transfer market. To estimate the value for money, they used player ratings like the plus-minus ratings (Kharrat, McHale, and Pena, 2020), and the action value ratings (Liu, Luo, Schulte, and Kharrat, 2020) to capture the contribution of a player to on-the-pitch performances. There have also been attempts to directly estimate transfer fees using basic performance statistics like goals scored and minutes played (e.g., Muller, Simons and Weinmann, 2017; Coates and Parshakov, 2021). Clustering techniques on player performance have been applied (e.g., D'Urso et al., 2022; Carpita et al., 2023). However, none of these papers account for the synergy of a player to a team considering the player's playing style. Given the playing style of a player, she or he may be valued differently by different clubs. In 2009, Barcelona recruited Zlatan Ibrahimović in exchange for Samuel Eto'o and a transfer fee of 70 million Euros. In retrospect, it was considered a bad transfer as Eto'o's performance in Barcelona was superior to that of Ibrahimović. Such imprudent transfers may be avoided if transfer fees are determined after considering the player's playing style by applying the methods described in this paper.

There is an important limitation to our work. In some cases, playing style may not be an intrinsic property of a player. A player may have a particular style that is dictated by the manager. For example, Alphonso Davies plays full-back for Bayern Munich. However, he plays as a forward when competing for the Canadian National Team. Davies will have different characteristics according to Table 1 under the two positional scenarios. Another limitation is that we have only one season of tracking data from the CSL. With additional seasons, playing style may be estimated more reliably.

Having identified a pool of players who are similar to a player of interest, an important future research question is: who are the best players from the pool? A data-based solution is not straightforward since player performance statistics (e.g. goals, passes, etc) are highly dependent on teammates and opponents. For the time being, these assessments are left to the subjective evaluations of subject matter experts (e.g. managers, assistants, scouts, etc).

Another direction of future work is the consideration of alternative discrepancy measures to assess similarity. For example, one could use the Bhattacharyya distance and the Jensen-Shannon divergence measure.

6 REFERENCES

Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Editors (2017). *Handbook of Statistical Methods and Analyses in Sports*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Carlin, B.P. and Louis, T.A. (2000). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452), 1286-1289.

Carpita, M., Pasca, P., Arima, S. and Ciavolino, E. (2023). Clustering of variables methods and measurement models for soccer players' performances. *Annals of Operations Research*, 1-20. https://doi.org/10.1007/s10479-023-05185-w

Coates, D. and Parshakov, P. (2022). The wisdom of crowds and transfer market values. *European Journal of Operational Research*, 301(2), 523-534.

D'Urso, P., De Giovanni, L. and Vitale, V. (2022). A robust method for clustering football players with mixed attributes. *Annals of Operations Research*, 1-28. https://doi.org/10.1007/s10479-022-04558-x

Decroos, T. and Davis, J. (2020). Player vectors: Characterizing soccer players' playing style from match event streams. Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis M. and C. Robardet (Editors), In *ECML PKDD 2019: Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science*, vol 11908, Springer, Cham, 569-584.

Decroos, T., Van Roy, M., and Davis, J. (2021). SoccerMix: Representing soccer actions with mixture models. Dong, Y., Ifrim, G., Mladenić, D., Saunders, C. and Van Hoecke, S. (Editors), In *ECML PKDD 2020: Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science*, vol 12461, Springer, Cham, 459-474.

del Rio, J. (2017). Paulinho brings power and energy to Barcelona. *Marca*, Accessed March 2, 2023 at https://www.marca.com/en/football/barcelona/2017/08/14/5991a6aae2704e8e5d8b45cd.html

Epasinghege Dona, N. and Swartz, T.B. (2023). A causal investigation of pace of play in soccer. *Statistica Applicata - Italian Journal of Applied Statistcs*, 35(1), Article 6.

Gill, P. and Swartz, T.B. (2019). A characterization of the degree of weak and strong links in doubles sports. *Journal of Quantitative Analysis in Sports*, 15, 155-162.

Goes, F.R., Brink, M.S., Elferink-Gemser, M.T., Kempe, M. and Lemmink, K.A.P.M. (2021). The tactics of successful attacks in professional association football: Large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, 39(5), 523-532.

Gómez, M.A., Mitrotasios, M., Armatas, V. and Lago-Peñas. (2018). Analysis of playing styles according to team quality and match location in Greek professional soccer. *International Journal of Performance Analysis in Sport*, 18, 986-987.

Gramacy, R.B. and Pantaleo, E. (2010). Shrinkage regression for multivariate inference with missing data, and application to portfolio balancing. *Bayesian Analysis*, 5(1), 237-262.

Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, 50(2), Article 22.

Henze, N. and Wagner, T. (1997). A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis*, 62, 1-23.

Hewitt, A., Greenham, G. and Norton, K. (2016). Game style in soccer: What is it and can we quantify it? *International Journal of Performance Analysis in Sport*, 16, 355-372.

Kharrat, T., McHale, I. G. and Peña, J.L. (2020). Plus–minus player ratings for soccer. *European Journal of Operational Research*, 283(2), 726-736.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 21, 79-86.

Lago-Peñas, C., Gómez-Ruano, M. and Yang, G. (2017). Styles of play in professional soccer: an approach of the Chinese Super League. *International Journal of Performance Analysis in Sport*, 17, 1073-1084.

Liu, G., Luo, Y., Schulte, O. and Kharrat, T. (2020). Deep soccer analytics: learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery*, 34, 1531-1559.

McHale, I. G. and Holmes, B. (2023). Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. *European Journal of Operational Research*, 306(1), 389-399.

McHale, I. G. and Relton, S.D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, 268(1), 339-347.

Müller, O., Simons, A. and Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, 263(2), 611-624.

Shaw, L. (2019). Friends-of-Tracking-Data-FoTD/LaurieOnTracking Accessed November 20, 2021 at https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

Shaw, L. and Glickman, M. (2019). Dynamic analysis of team strategy in professional football. *Barça Sports Analytics Summit.*

Shen, E. (2022). Analyzing pace-of-play in soccer using spatio-temporal event data. *Journal of Sports Analytics*, 8(2), 127-139.

Skinner, B. and Guy, S.J. (2015). A method for using player tracking data in basketball to learn player skills and predict team performance. *PLoS ONE*, 10: https://doi.org/10.1371/journal .pone.0136393

Wilson, J. (2013). *Inverting the Pyramid*, Nation Books, New York.

Wu, Y. and Swartz, T.B. (2023). Evaluation of off-the-ball actions in soccer. *Statistica Applicata - Italian Journal of Applied Statistics*, 35(2), Article 2.

Feature	Description
x_1	avg player position when ball is in offensive third and team in possession
x_2	avg EPV of player's position when ball is in offensive third and team in possession
x_3	avg vertical distance (downfield distance) and horizontal distance of player's successful passes when ball is in offensive third and team in possession
x_4	avg EPV increase in player's successful passes when ball is in offensive third and team in possession
x_5	pct of time player has the ball when ball is in offensive third and team in possession
x_6	pct of the team's passes by player when ball is in offensive third and team in possession
x_7	pct of the team's shots by player when ball is in offensive third and team in possession
x_8	avg player position when ball is in offensive third and opponent in possession
x_9	avg distance from nearest opponent when ball is in offensive third and opponent in possession
x_{10}	pct of the team's interceptions by player when ball is in offensive third and opponent in possession
x_{11}	pct of the team's tackles by player when ball is in offensive third and opponent in possession
x_{12}	avg player position when ball is in middle third and team in possession
x_{13}	avg EPV of player's position when ball is in middle third and team in possession
x_{14}	avg vertical distance (downfield distance) and horizontal distance of player's successful passes when ball is in middle third and team in possession
x_{15}	avg EPV increase in player's successful passes when ball is in middle third and team in possession
x_{16}	pct of time player has the ball when ball is in middle third and team in possession
x_{17}	pct of the team's passes by player when ball is in middle third and team in possession
x_{18}	avg player position when ball is in middle third and opponent in possession
x_{19}	avg distance from nearest opponent when ball is in middle third and opponent in possession
x_{20}	pct of the team's interceptions by player when ball is in middle third and opponent in possession
x_{21}	pct of the team's tackles by player when ball is in middle third and opponent in possession
x_{22}	avg player position when ball is in defensive third and team in possession
x_{23}	avg vertical distance (downfield distance) and horizontal distance of player's successful passes when ball is in defensive third and team in possession
x_{24}	pct of time player has the ball when ball is in defensive third and team in possession
x_{25}	pct of the team's passes by player when ball is in defensive third and team in possession
x_{26}	avg player position when ball is in defensive third and opponent in possession
x_{27}	avg distance from nearest opponent when ball is in defensive third and opponent in possession
x_{28}	pct of the team's interceptions by player when ball is in penalty box and opponent in possession
x_{29}	pct of the team's interceptions by player when ball is in defensive third outside penalty box and opponent in possession
x_{30}	pct of the team's tackles by player when ball is in defensive third and opponent in possession

Table 1: Player features obtained from tracking data that reflect playing style. The statistics are collected on a per-match basis where the three table categories correspond to the position of the ball on the pitch. Within each category, statistics are defined according to whether the team of interest or the opponent has possession of the ball.

Feature	Mean	Min	Max	StdDev
x_1	(16.60, 11.87)	(-42.31, 0.37)	(45.12, 32.02)	(17.53, 5.18)
x_2	0.04	0.01	0.17	0.02
x_3	(0.34, 10.28)	(-43.30, 0)	(29.90, 57.10)	(5.97, 5.37)
x_4	0.01	-0.25	0.43	0.03
x_5	0.07	0	0.65	0.09
x_6	0.07	0	0.38	0.07
x_7	0.07	0	1	0.11
x_8	(7.40, 11.27)	(-42.63, 0.43)	(49.46, 31.26)	(16.27, 4.67)
x_9	8.40	0.45	44.65	7.33
x_{10}	0.07	0	1	0.16
x_{11}	0.07	0	1	0.17
x_{12}	(-1.16, 13.54)	(-44.69, 0.63)	(28.46, 32.80)	(14.70, 5.80)
x_{13}	0.02	0.01	0.05	5.25×10^{-3}
x_{14}	(2.11, 10.89)	(-43.20, 0.10)	(49.50, 45.10)	(6.04, 4.09)
x_{15}	1.88×10^{-3}	-0.01	0.09	3.67×10^{-3}
x_{16}	0.07	0	0.39	0.06
x_{17}	0.07	0	0.29	0.05
x_{18}	(-10.26, 11.91)	(-47.24, 0.73)	(21.70, 28.83)	(13.08, 4.51)
x_{19}	7.62	1.64	34.07	5.44
x_{20}	0.07	0	0.45	0.08
x_{21}	0.07	0	0.67	0.10
x_{22}	(-18.50, 12.48)	(-49.36, 0.24)	(22.49, 36.82)	(13.13, 5.12)
x_{23}	(6.71, 10.68)	(-25.90, 0)	(59.10, 54.00)	(9.78, 5.12)
x_{24}	0.07	0	0.86	0.11
x_{25}	0.07	0	0.58	0.07
x_{26}	(-28.21, 10.70)	(-50.66, 0.51)	(27.83, 36.23)	(11.15, 4.10)
x_{27}	5.89	0.49	25.52	2.61
x_{28}	0.07	0	1	0.14
x_{29}	0.07	0	0.52	0.09
x_{30}	0.07	0	0.60	0.09

Table 2: Summary statistics for the features presented in Table 1 calculated across all players and matches. The statistics are provided for the raw data prior to standardization. Note that the bivariate statistics correspond to features that have (x, y) coordinates.