The Annals of Applied Statistics
2022, Vol. 16, No. 1, 326–348
https://doi.org/10.1214/21-AOAS1506
© Institute of Mathematical Statistics, 2022

THE ROLE OF INTRINSIC DIMENSION IN HIGH-RESOLUTION PLAYER TRACKING DATA—INSIGHTS IN BASKETBALL

BY EDGAR SANTOS-FERNANDEZ[1,a], FRANCESCO DENTI[2,c], KERRIE MENGERSEN[1,b] AND ANTONIETTA MIRA[3,d]

[1]*School of Mathematical Sciences, Queensland University of Technology,* [a]*santosfe@qut.edu.au,* [b]*k.mengersen@qut.edu.au*
[2]*Department of Statistics, University of California, Irvine,* [c]*fdenti@uci.edu*
[3]*Faculty of Economics, Università della Svizzera italiana,* [d]*antonietta.mira@usi.ch*

Following the introduction of high-resolution player tracking technology, a new range of statistical analysis has emerged in sports, specifically in basketball. However, such high-dimensional data are often challenging for statistical inference and decision making. In this article we employ a state-of-the-art Bayesian mixture model that allows the estimation of heterogeneous intrinsic dimension (ID) within a dataset, and we propose some theoretical enhancements. Informally, the ID can be seen as an indicator of complexity and dependence of the data at hand, and it is usually assumed unique. Our method provides the capacity to reveal valuable insights about the hidden dynamics of sports interactions in space and time which helps to translate complex patterns into more coherent statistics. The application of this technique is illustrated using NBA basketball players' tracking data, allowing effective classification and clustering. In movement data the analysis identified key stages of offensive actions, such as creating space for passing, preparation/shooting, and following through which are relevant for invasion sports. We found that the ID value spikes, reaching a peak between four and eight seconds in the offensive part of the court, after which it declines. In shot charts we obtained groups of shots that produce substantially higher and lower successes. Overall, game-winners tend to have a larger intrinsic dimension, indicative of greater unpredictability and unique shot placements. Similarly, we found higher ID values in plays when the score margin is smaller rather than larger. The exploitation of these results can bring clear strategic advantages in sports games.

1. Introduction. Basketball is a popular and highly dynamic invasion sport. At its most basic level, it is simple to understand. The game takes place on a hard court with round hoops at each end, which are 10 feet off the ground. Two teams, each with five players on the court at a time, go from end to end trying to get the basketball in the opposite team's hoop, or goal, while the other team tries to stop them. At an advanced level, especially in professional basketball, teams use a large variety of trained plays to increase their chances of scoring. Sports teams around the world are increasingly using quantitative methods to produce insights and strategies. Current technology is now able to track player movement on a basketball court. Specifically, SportVU NBA player tracking technology can capture each player's movements at 25 frames per second. These high-resolution data have motivated several spatial and spatiotemporal statistical analyses (e.g., Goldsberry (2012), Shortridge, Goldsberry and Adams (2014), Cervone et al. (2016)). However, modeling the complexity of these high-dimensional data is often challenging and computationally expensive and, therefore, requires more sophisticated statistical techniques.

Received December 2020; revised May 2021.
Key words and phrases. Bayesian clustering, high-dimensional data, intrinsic dimension, plays classification, movement data.

Many factors need to be considered when assessing the success of a play. It is well known that the placement of the players in attack and defense—and particularly the guard to the player taking the shot—is related to the success of a play. Similarly, the success of a team also depends on the versatility of the attacking players: more versatile players will score shots from more unique locations in the court. In other words, versatile players tend to have higher placement variability. Furthermore, increased uncertainty about attackers' positions tends to generate plays that are harder to defend. Finally, successful teams create more shooting opportunities by passing the ball more effectively. All of these factors are deemed to produce an increased success for the attacking team (e.g., Lamas et al. (2014)). We can conclude that, when a team is in attack, a large coordinated unpredictability, combined with uncertainty in the players' movements and passes, and a broad spectrum of shooting locations can lead to more successful plays. However, the impact of this unpredictability in coordinated plays has not received much attention in the literature. Recently suggested measures, like *ball entropy* or *unpredictability*, have been regarded as key performance factors in sports games (Lucey et al. (2012), D'Amour et al. (2015), Skinner and Goldman (2017), Hobbs et al. (2018)). In this regard, Skinner and Goldman (2017) have pointed out that the expected return in a play will decrease as it is used more often.

Nowadays, a large number of individual statistics are collected in basketball games. Teams monitor players' traditional summary statistics, like the number of points (PTS), the defensive rebounds (DREB), the number of assists (AST), the field goals made (FGM), the Three-Point Field Goals Made (3PM), the minutes played (MIN), etc. In addition, the tracking technology estimates several other metrics: distance (Dist. Feet), average speed (Avg. Speed), passes made and received, etc.

Analysis, with this large number of variables for each of the 15 players on the active roster during the 82 games usually played in the NBA season, motivates a multivariate perspective for analysis. Hence, data scientists working in sports analytics are increasingly switching from descriptive statistics to more refined data-analysis techniques, for example, clustering. These more complex methods allow better communication of performance to coaches. Lutz (2012) has used informative statistics, such as field goals, steals, and assist ratio to cluster players with similar features into 10 categories. Franks et al. (2015) used nonnegative matrix factorization and clustered defensive players using field-goal locations. This approach provides a measure of the impact of defensive players on shot frequency and probability of scoring. Other simple clustering algorithms, such as the popular k-means, have been successfully used for analyzing basketball data. For instance, Sampaio et al. (2015), grouped players based on performance, employing attacking, defense, and passing statistics. More recently, Nistala and Guttag (2019) provided a classification of players' movements based on Euclidean distance. They clustered attacking movements into 20 groups, including screen action, movement along each sideline, run along the baseline, etc. Additional clustering applications can be found in Metulini, Manisera and Zuccolotto (2017) and Metulini (2018).

Other methods, including dimensionality reduction techniques, also play a crucial role in sports. Examples of principal components analysis (PCA) in basketball can be found in Sampaio, Drinkwater and Leite (2010) and Teramoto et al. (2018).

Generally, high-dimensional datasets can be projected onto lower-dimensional manifolds without losing much information (Levina and Bickel (2005), Camastra and Staiano (2016)). The dimensionality of these manifolds is called the *intrinsic dimension* (ID from now on) of the data. We will provide a more formal definition in the next section. Estimating the ID of a dataset offers a measure of the amount of information and redundancy among the considered variables. We can employ these statistical techniques with success to modern sports datasets. Positions and movements of players on attack and defense are generally correlated because defensive players guard those in attack during the execution of the play. Also, the critical

parts of plays occur in localized regions of the court, mainly around the hoops. Thus, using ID estimation and dimensionality reduction techniques can provide important insights into sports analytics. However, little attention has been paid to study these players' movement data in invasion sports. Similarly, we found no discussion on the complexity assessment of the players' placements in shot charts and their relation to performance.

In this paper we employ a Bayesian model-based clustering technique to group the data according to a measure of their ID. This approach was first developed by Allegra et al. (2020) which was, in turn, built on the model proposed in Facco et al. (2017). As shown in these papers, this novel methodology applies to a wide variety of datasets and cases, including fMRI data, financial data, and the analysis of protein folding structures. Their method allows the presence of different groups with different IDs in the same dataset. In this paper our data points are the locations, distance traveled, and directions of players on the court. Therefore, the purpose of this paper is to use the mentioned Bayesian model-based clustering based on the ID of the data to: (i) assess players' movement complexities in three-point field-goal, midrange, and close shots, (ii) identify the phases in the execution of a play (e.g., ball handling, creating space for passing, preparation, shooting and following through), (iii) study patterns in shot charts identifying plays that produce worse/better outcomes, and (iv) examine whether unpredictability in attack is linked to better performance. Finally, we aim to describe the complexity of movement data by assessing the progression of the ID over time.

The article proceeds as follows. In Section 2 we introduce the modeling framework and propose some methodological enhancements. Section 3 presents the data and some technical definitions of the game. Then, in Section 4 we show and discuss the main results of the analysis. Finally, Section 5 concludes with a discussion of the findings and limitations.

2. The intrinsic dimension of the data. The concept of intrinsic dimension (ID) is a fundamental tool for unraveling the structure of high-dimensional, complex datasets. The ID of a dataset, observed in a D-dimensional space, can be defined as the dimension d of the latent (possibly nonlinear) manifold in which the statistical units lie. Generally, we expect some degree of dependency among the variables of a dataset, so we can assume that $d \leq D$. More formally, we refer to the ID as *the minimal number of parameters needed to represent the data without significant information loss* (Ansuini et al. (2019), Rozza et al. (2012), Bennett (1969)). In other words, the ID provides a unique indication of the complexity and redundancy of the features in a dataset.

Estimating the ID is an essential step in any manifold learning analysis. For example, relying on a good estimate of the ID is crucial for subsequent dimensionality reduction steps. In the past decades, several methods for ID estimation have been developed. The literature in this field is vast: we refer to Campadelli et al. (2015) for a comprehensive review. In this paper we employ a model-based ID estimator recently introduced by Facco et al. (2017) and extended in Allegra et al. (2020). In the following we briefly provide the theoretical background needed to understand the model. For more details the reader is referred to the articles mentioned above and the references therein.

2.1. *The modeling background.* Consider a dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, with each observation measured over D continuous variables, $\mathbf{x}_i \in \mathbb{R}^D$. We postulate that the points are realizations from a Poisson point process with intensity function lying on a d dimensional manifold, with $d < D$. We define as $\mathbf{x}_{(j,i)}$ the jth nearest neighbor (NN) of the ith observation and consider a distance function $\Delta : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$ (e.g., the Euclidean distance). Finally, we let $r_{ij} = \Delta(\mathbf{x}_i, \mathbf{x}_{(j,i)})$ be the distance between the ith observation and its jth NN. Then, *if the intensity of the Poisson point process is assumed to be constant on the scale of the second NN, then the following distributional results hold*:

$$\mu_i = r_{i2}/r_{i1} \sim \text{Pareto}(1, d), \quad \mu_i \in [1, +\infty], i = 1, \ldots, N. \tag{1}$$

In other words, if the density of the data can be assumed to be locally constant, we can employ a simple transform of the data and estimate the ID as the shape parameter of a Pareto distribution. Recall that a Pareto random variable X with scale parameter 1 and shape parameter d, defined on $x \in [1, +\infty]$, is characterized by the p.d.f. $f_X(x) = \frac{d}{x^{d+1}}$ and the c.d.f. $F_X(x) = 1 - (1/x)^d$. To estimate the ID, Facco et al. (2017) proposed a frequentist estimator based on least squared regression, obtained by linearizing the c.d.f. of the Pareto distribution. The estimate $\hat{d}_{\text{OLS}}$ is obtained as the solution of

$$\log\bigl(1 - \hat{F}(\mu_i)\bigr) = -d \log(\mu_i),$$

where $\hat{F}(\cdot)$ denotes the empirical c.d.f. of the sample. This model provides a reliable model-based method for the estimation of the ID of a dataset.

However, the assumption of a single ID for an entire dataset may be too restrictive, especially when data present complex dependence structures. It is reasonable to assume that a dataset may be composed of different groups of statistical units that are characterized by different IDs. That is, the intensity of the underlying Poisson point process can be seen as a mixture of K distributions defined on K different latent manifolds. This formulation, in turn, induces a mixture of Pareto distributions over the ration μ_i's. Allegra et al. (2020) investigated this framework, defining a Bayesian mixture model,

$$P(\mu_i|\mathbf{d}, \mathbf{p}) \doteq \mathcal{P}(\mu_i) = \sum_{k=1}^{K} p_k d_k \mu_i^{-(d_k+1)}, \quad \mathbf{p} = (p_k)_{k=1}^{K} \sim \text{Dir}(c_1, \ldots, c_K), \tag{2}$$

where $d_k \sim \text{Gamma}(a, b)$, $\forall k = 1, \ldots, K$. These prior choices are motivated by conjugacy which greatly simplifies posterior simulation and inference. This model induces a partition among the observations, with every cluster characterized by its ID value.

As standard practice when dealing with mixture models, a set of auxiliary variables, indicating the cluster membership labels $\mathbf{z} = (z_i)_{i=1}^{N}$, is introduced. In this context, these membership labels can be interpreted as the latent manifold assignment. Consequently, model (2) can be expressed as

$$\mu_i|z, \mathbf{d} \sim \text{Pareto}(1, d_{z_i}), \qquad z_i|\mathbf{p} \sim \sum_{k=1}^{K} \pi_k \delta_k(z_i), \tag{3}$$

where $\delta_x(y)$ is the usual Dirac delta, equal to 1 if $x = y$, and 0 otherwise. The posterior distribution for the parameters cannot be obtained analytically but is approximated with MCMC techniques. Within this modeling setting the best number of mixture components K is chosen ex post. We can compare the average log-posterior estimated over the MCMC iterations. Alternatively, one could use more complete measures of model comparisons, like DIC, BIC, AICm, BICm, or WAIC.

However, the partitions obtained by estimating model (3) are unreliable and so are the estimated IDs. The reason is the strong overlap that characterizes different Pareto densities. Thus, multiple Pareto densities can be a viable modeling choice for the same data point. To solve this issue, Allegra et al. (2020) introduced an additional assumption: *the different manifolds are separated in space, and the neighborhood of a point should be more likely to contain points sampled from the same manifold than points sampled from a different manifold*. They imposed this effect in the model with the addition of an extra term in the likelihood. The novel term explicitly models the neighboring structure of the data via a binary adjacency matrix $\mathcal{N}^{(q)}$, aiming at enforcing local homogeneity. The entry $\mathcal{N}_{ij}^{(q)} = 1$ if $\boldsymbol{x}_j$ is among the q nearest neighbors of $\boldsymbol{x}_i$. This event can happen with probability $\zeta > 0.5$ if $z_i = z_j$, that is, if the two points are assigned to the same latent manifold. Otherwise, if $z_i \neq z_j$, we have

$\mathbb{P}[\mathcal{N}_{ij}^{(q)} = 1] = 1 - \zeta$; see Allegra et al. (2020) for more technical details. The resulting joint likelihood for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\mathcal{N}^{(q)}$ is

$$\mathcal{L}(\boldsymbol{\mu}, \mathcal{N}^{(q)} | \mathbf{d}, \mathbf{z}, \zeta) = \prod_{i=1}^{n} \mathcal{P}(\mu_i | d_{z_i}) \times f(\mathcal{N}_i^{(q)} | \mathbf{z}, \zeta). \tag{4}$$

The mixture model can be sampled with straightforward MCMC techniques, obtaining a posterior sample for the ID parameters and cluster memberships.

In this paper our methodological contribution is two-fold. First, we propose to adopt different prior specifications for $\mathbf{d}$ that may better suit the problem at hand. Second, we introduce a tailored postprocessing procedure to maximize the extraction of information from the MCMC sample.

2.2. *More meaningful priors for the ID.* The default prior specification in Allegra et al. (2020) for each element of the parameter vector $\mathbf{d}$ is a Gamma distribution. However, there are cases where a more thoughtful choice could help.

Truncated prior. Given the unbounded support of the Gamma distribution, it is possible that part of the posterior distribution may fall above the dataset nominal dimension D. This problem occurs more frequently when the nominal dimension D is low, that is, the original dataset contains only a few variables. To address this issue, we propose to substitute the Gamma prior on d_k with a Truncated Gamma over $(0, D)$ for $k = 1, \ldots, K$ we assume $\pi(d_k) = \mathcal{C}_{a,b,D} d_k^{a-1} \exp\{-b d_k\} \mathbb{1}_{(0,D)}(d_k)$, where $\mathcal{C}_{a,b,D}$ is the corresponding normalizing constant.

Alternatively, if we want to include the case where $d = D$, we can employ a mixture model between a truncated distribution and a point mass in D. The density is

$$\pi(d_k) = \hat{\rho}\big[\mathcal{C}_{a,b,D} d_k^{a-1} \exp\{-b d_k\} \mathbb{1}_{(0,D)}\big] + (1 - \hat{\rho}) \delta_D(d_k) \quad \forall k,$$

where $\hat{\rho}$ denotes the mixing proportion. That is, the ID of a specific mixture component can be equal to the upper bound with prior probability $1 - \hat{\rho}$. To draw inference about $\hat{\rho}$, a conjugate Beta prior is specified. Notice that we chose a truncated Gamma distribution to keep exploiting its conjugacy property. However, different priors with support on $(0, D]$ can be adopted, for example, a Uniform or a rescaled Beta distribution.

Repulsive prior. Dealing with mixture models could lead to overfitting, in the sense that the model tends to create more components than the ones that are needed. In some applications, one may observe different clusters of observations characterized by very similar IDs. Instead of reflecting real differences in the latent manifold dimensions or the cluster allocations, this distinction could be due to noise in the observed data or minor curvatures in the latent geometry. To mitigate this issue and avoid the creations of redundant components, one can employ a repulsive density, as in Petralia, Rao and Dunson (2012),

$$\pi(\mathbf{d}) = c_1 \left(\prod_{k=1}^{K} g_0(d_k) \right) h(\mathbf{d}), \quad h(\mathbf{d}) = \min_{\{(s,j) \in A\}} g\big(\Delta(d_s, d_j)\big), \tag{5}$$

where Δ is a suitable distance in $\mathbb{R}^+$, g_0 is a univariate density function for d_k, and $A = \{(s, j) : s = 1, \ldots, K; j < s\}$. Instead of specifying the function g as in the aforementioned paper, that is, $g(\Delta) = \exp[-\tau(\Delta)^{-\nu}]$ with $\tau, \nu > 0$, we adopt the following sigmoidal function:

$$g(\Delta) = \left(1 + \exp\left[-\frac{\Delta - \tau}{\nu}\right]\right)^{-1}, \quad \tau, \nu > 0. \tag{6}$$

This functional form gives us direct control over the amount of separation between realizations. This prior favors more heterogeneous realizations of mixture parameters d_k, therefore, avoiding clusters with almost identical ID values.

2.3. *A postprocessing procedure for posterior inference*. We now present the postprocessing techniques we will employ to estimate the heterogenous IDs and to obtain a partition of the data.

A more considerate estimation of the ID. The main goal of the Bayesian mixture model we described is to recover a meaningful partition of the observations according to their IDs. Directly estimating the IDs from the MCMC chains of the vector $\mathbb{d}$ should be avoided, since the result can be hindered by the label switching issue (Celeux (1998)). We propose to derive the ID estimate for every data point by tracking the ID value assigned to each observation across the MCMC iterations. Specifically, let us consider an MCMC sample of length T. Denote with z_i^t the membership label for observation i at the tth iteration. Let $d_{z_i^t}$, the ID value assigned to the cluster where observation i has been allocated. In other words, we are creating N observation-specific chains monitoring via the cluster assignation. Once the MCMC chains have been postprocessed, the ID can be easily estimated by the ergodic mean or the ergodic median. These estimators automatically correct for the label switching problem. We are well aware that more complete and involved methodologies have been proposed for handling this nonidentifiability issue (Robert (2010), Rodríguez and Walker (2014), Sperrin, Jaki and Wit (2010), Frühwirth-Schnatter (2011)), and we leave the adoption of refinements for future research.

Recovering a meaningful partition. To obtain an estimate of the best partition in the data, we adopt a decision-theoretical rationale, as commonly done in Bayesian mixture models literature. First, we compute the $N \times N$ pairwise coclustering matrix $\widehat{PPC}$. The entries of this symmetric matrix are defined as

$$\widehat{PPC}(i, j) = \frac{\sum_{t=1}^{T} \mathbb{1}_{(z_i^t = z_j^t)}}{T}, \quad i, j = 1, \ldots, N.$$

In other words, the matrix describes the frequency of times that two observation i and j have been clustered together across the MCMC iterations. This matrix is the input for the computation of widely used loss functions defined on the space of the partitions, such as the Binder loss or the Variation of Information (Lau and Green (2007), Wade and Ghahramani (2018)). The optimal clustering solution is the one that minimizes the adopted loss. However, we suggest interpreting this estimate with caution. Despite the correcting term introduced in the likelihood, the model-based clustering induced by the mixture may still suffer from the overlapping among the Pareto distributions and, consequently, might not be reliable. A potential solution to this issue could be to cluster the observations, according to MCMC posterior estimates, by classical clustering algorithms, such as k-means on $\{d_{z_i^t}\}$. The optimal number of groups k can be fixed by studying the behavior of cluster quality indexes such as *Silhouette* (Rosseeuw (1987)) or the *Calinski–Harabasz index* (Caliński and Harabasz (1974)).

3. Data analysis.

3.1. *Description of the dataset*. We used STATS SportVU high-resolution player tracking raw data from the NBA during the 2015–16 season. These spatiotemporal state-of-art

TABLE 1
The 15 *randomly selected games from the season* 2015–16

Away	Home	Date (MM.DD.YYYY)	Result
GSW	LAL	01.05.2016	109–88
MIL	CHI	01.05.2016	106–117
MIA	TOR	01.22.2016	81–101
CLE	GSW	12.25.2015	83–89
HOU	SAS	01.02.2016	103–121
PHI	LAL	01.01.2016	84–93
MEM	OKC	01.06.2016	94–112
UTA	SAS	01.06.2016	98–123
BKN	BOS	01.02.2016	100–97
TOR	CLE	01.04.2016	100–122
MIA	GSW	01.11.2016	103–111
OKC	CHA	01.02.2016	109–90
MIA	WAS	01.03.2016	97–75
MIA	PHX	01.08.2016	103–95
GSW	POR	01.08.2016	128–108

datasets contain the (x, y) coordinates of the 10 players on the court during the whole game at 25 Hertz or frames per second. The (x, y, z) coordinates are also given for the ball. The event id, date-time, and player details are also included. We obtained the descriptions of the play-by-play events and other meaningful statistics from the official website https://stats.nba.com/. In these files every play is represented by a tuple (row). It includes the play id, date-time, players on the court, score, the outcome of the play, etc.

We joined both datasets based on the play's id. Therefore, the resulting dataset contains for every play the id, the movement of the players, the ball and the outcome, game score, etc. The correct matching was manually verified using the games' video available on https://www.youtube.com/. We observed the players' movement, the outcome of a play, and the score from the videos. This manual matching reduced the likelihood of errors in the data. This data curation and manual matching are time-consuming. Therefore, for this paper we considered 15 random games whose outcomes were summarized in Table 1.

3.2. *Some basketball definitions*. In this subsection we describe some of the basic rules of the game and provide some definitions to help the reader better understand the following sections. However, we cannot cover all the rules and their exceptions in this paper. One can find more details at https://official.nba.com/rulebook/. As we mentioned, basketball games are played between two teams with five players on the court each. The team that hosts the game in its arena is called *home* team. We refer to the visiting as *away* team.

A team is in *possession* when it controls the ball (dribbling, passing, or holding it), and we refer to it as being on *offense* or *attack*. The team on *defense* is the team preventing the other one from getting points scored. The team that is on attack must attempt a field goal within 24 seconds. We refer to a *play* as one action that starts when the team gets possession of the ball and ends when they lose it. For example, team A gets a ball possession after team B scores. The play finishes when team A shoots and misses and team B gets the rebound. If instead, team A gets an offensive rebound, the clock is reset corresponding to a new play. The play encompasses the movements of the players during the possession. Every play has an outcome (e.g., scored, missed, foul, steal, etc.), and a game is composed of many plays.

The term *shot chart* refers here to the positions in the x and y axes of the 10 players when the shot was taken. Our definition is different from the traditional shot charts that consist only

(a) Video frame at the moment of shooting.

(b) Locations from the high-resolution player tracking technology.

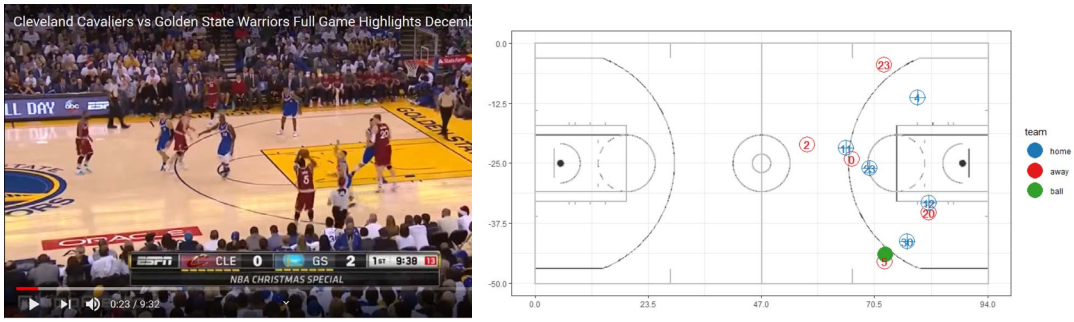

FIG. 1. *Locations of the players and the ball for the first scored three-point field goal of the game Cleveland Cavaliers* (*CLE*) *and the Golden State Warriors* (*GSW*) *on the* 25*th of December* 2015. *This play can be watched at https://youtu.be/jb57MFQLoRo?t=*17.

of the location of the player taking the shot. The term *trajectory* describes the path in the 2D space followed by a player or the ball during a play. Every play is composed of the trajectory of the players and the ball.

In the next section we show the results for three analyses we performed on the described data:

1. *within plays* (Section 4.1). First, we focus on the ID evolution on the movement, distance, and trajectories of the players during a ball possession. This approach accounts for variability in space and time. We illustrate the ID analyses using the game between Cleveland Cavaliers (CLE) and the Golden State Warriors (GSW) on the 25th of December 2015. These two teams made it to the final in that season.
2. *between plays* (Section 4.3). To this extent, we use shot-chart data composed of the locations of the 10 players at the moment of the shot. The resulting data can be seen as generated from a spatial process. Similarly, we again used CLE vs. GSW data from the 25th of December 2015. Figure 1 shows the locations of the players during the first scored three-point field goal of the game by the video screen-shot (a) and the representation of play obtained from the high-resolution player tracking technology (b).
3. *between games* (Section 4.3). This analysis is also performed using shot-chart data, as in point 2 but compares 15 random games.

4. Results.

4.1. *ID in the analysis of movement data.* In this section we assess the change in ID within plays produced by players' movement data in the offensive court, that is, after the ball passes the 47-foot central line, as commonly done (see, e.g., Franks et al. (2015)).

Movement data are highly redundant and correlated when we look at intervals of time recorded on a small scale, like milliseconds. Therefore, the resolution of each play was reduced from 25 to 2.5 frames/second for faster computation and easier visualization, without losing a substantial amount of information. We define the first frame as the first timestamp of the play. Consequently, the number of frames in a game depends on the play's duration. The dataset and the R code employed for the IDs estimation are available at https://github.com/EdgarSantos-Fernandez/id_basketball, along with an animation of the play from Figure 1 that well exemplifies the movement data.

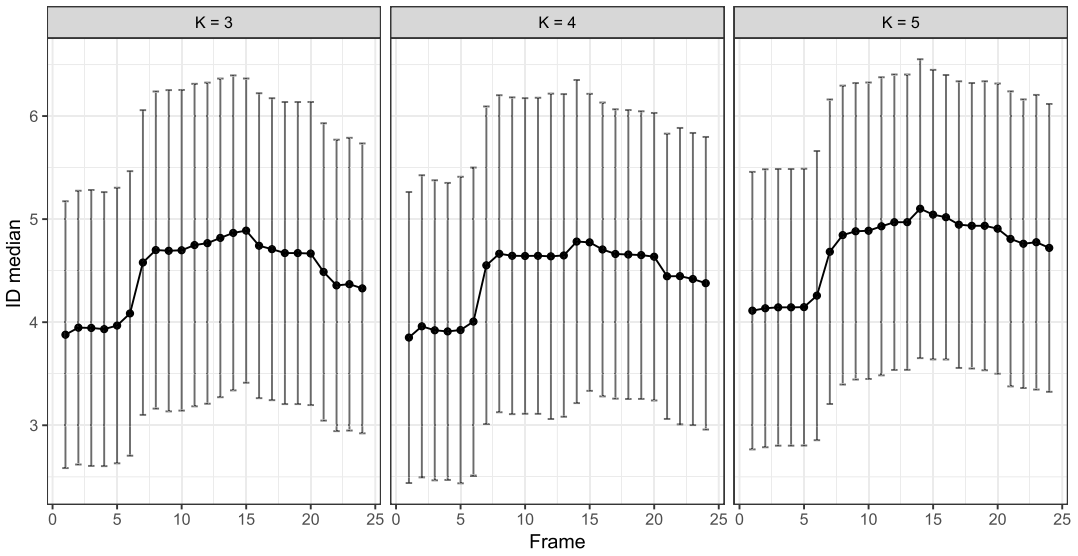

FIG. 2. *Comparison of the posterior ID median values with error bars on the movement data from the play in Figure 1 using several mixture components based on the repulsive prior.*

As a first step we performed a sensitivity analysis to assess the impact of the three different types of prior distributions (conjugate, truncated, and repulsive) on the estimated ID values. In this analysis we found consistent results in the posterior ID median, regardless of the prior we used. We report these results in Appendix A in Figure 11. Analogously, we found stable results regardless of the number of prespecified mixture components K. For example, Figure 2 illustrates the evolution of the ID within a play, using the repulsive prior for $K = 3, 4$, and 5. Similar patterns in ID are obtained across the play, independently of the number of components. From this point on, we use the repulsive variant in our analysis with $K = 3$. Figure 2 also provides some interesting insights. We note in this figure low ID values in frames 1–5. A lower ID, in this context, can be interpreted as the players moving in the same direction, more coherently, and predictably. This usually occurs at the beginning of the play, when all the players get near the defending team's basket. After this first configuration we observe higher ID values. These values are the result of movements with more complex patterns, for example, when the players on attack work on screened plays or create space for the pass, avoiding the players on defense.

We analyzed all the plays during the GSW vs. CLE game mentioned above, computing the posterior ID distributions. Recall that a statistical unit in this context is the vector of the 10 players' coordinates (x, y) in the court at each time point. In Figure 3(a) we show the trajectory of the three players (Irving, Love, and Smith) involved in the play illustrated in Figure 1. In this play, Irving crosses the centerline dribbling, and in frame 7 the players start seeking space, receiving the pass. Then, in frame 13, the ball goes to Love, who passes to Smith, who executes a three-pointer in frame 19.

We computed the posterior coclustering matrix, given the ID estimates. We report the matrix in Figure 3(b). Brighter colors represent a higher probability of being clustered together. For instance, we detect a cluster containing frames 1 to 6. The line plot on top of Figure 3(b) represents the progression of the median ID across the 24 frames of play. Therefore, the evolution in the ID value captures the changes in movements' dynamics and complexities within a play. We note a spike in the ID value produced in stamp 7. As expected, consecutive frames tend to cluster together because players tend to preserve the momentum in short intervals of time. We identify four clusters (in yellow): frames 1–6, 8–13, 17–20, and 21–24. Also, some

(a) Movement of three players and the ball during a play.

(b) Heatmap of the 10 players' movements during the 24 time stamps (bottom) and posterior medians of the ID (top).

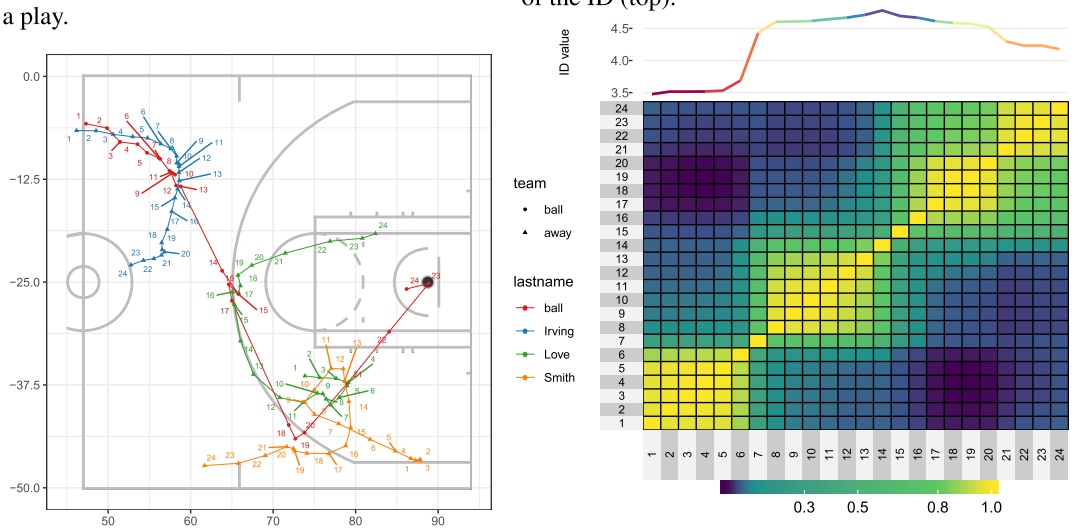

FIG. 3. *The trajectory of the three attacking players involved in the play plus the ball (left panel) and heatmap of the coclustering matrix (lower right) and ID evolution (top right panel) for the first scored three-point field goal of the game by CLE. In (b) the x and y axes represent the frame number. The frequency was set at* 2.5 *frames/second and the play is composed of* 24 *frames. The y axis in the top-right plot is the ID value. https://youtu.be/jb57MFQLoRo?t=17.*

interesting changes can be observed in frames 7 and 15. Given these considerations, we can segment the play into different stages. First, from frames 1 to 6, the ball handler crosses the centerline. Second, from frames 7 to 12, the players create space for the passing of the ball. Passing occurs between frames 13 and 17. Then, from frames 18 to 20, we have the preparation to shoot and the shooting. Finally, from frames 21 to 24, the following through takes place.

Another example of movement analysis during a *driving bank* two-point shot by Kyrie Irving is presented in Appendix B on Figure 12. Figure 12(a) shows the movement of the two players involved and the ball. One can watch the play following the link provided in the caption of the figure. Looking at the coclustering matrix in Figure 12(b) we observe three main clusters: frames 1 to 19, 20 to 26, and 27 to 30. The first cluster contains the ball handler (Irving) crossing the centerline, passing to Dellavedova who passes back to Irving. An interesting change occurs in frame 20, where Irving spins away from the defender and attacks the basket (frames 20 to 26). From frame 20 on, most of the players remain stationary, hence the decline in the ID. From frames 27 to 30, we are again in the following through phase.

We found that during ball possession, where players move in the same direction, the estimated median ID value is low. At the same time, we obtain higher ID values in multidirectional trajectories and complex plays. We illustrate this principle with the following analysis. We consider three simple plays ($idn = 23, 25, 96$) and three complex plays ($idn = 355, 477, 308$), where *idn* is the play identification number. We report the median ID value across the time points, characterizing these plays in Figure 4. Intuitively, simple plays are generally shorter: when they happen, the players reach the *paint zone* without much resistance. Here, in simple plays, five or fewer dimensions are enough to describe the 20 coordinates of the players, showing high redundancy in the data. In contrast, more complex plays require an ID value increasing to $d = 9$. Figures 5(a) and 5(b) show an example of the trajectories of the five players on attack, plus the ball in plays 96 and 308. We also provide a link to the videos of the plays in the Figure captions.

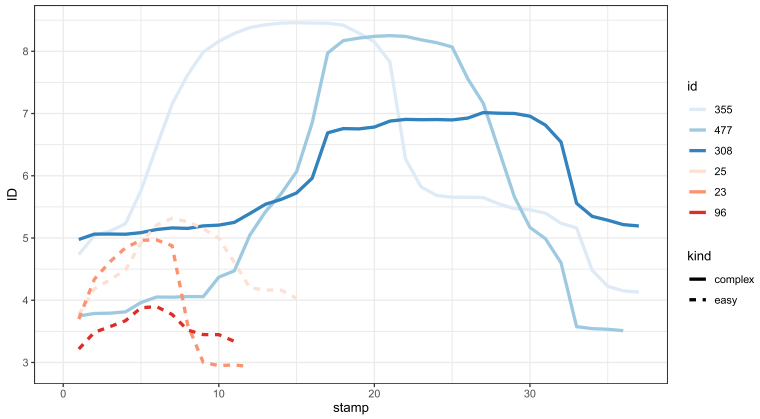

FIG. 4. *Example of the ID in six different plays, three complex (solid line) and three simple plays (dotted line).*

We performed a further analysis based on the Euclidean distance (δ) from the player taking the shot to the basket. Three shot groups were defined as follows:

- *short-distance shots* (dunks, tips, etc.), where $\delta < 6$ feet.
- *midrange shots* (short and long two-point shots), where $6 \leq \delta < 22$ feet.
- *three-point shots* (shots from behind the three-point arc line). This arc measures 22 feet in the sidelines or corners and 23.75 feet from the center of the basket.

Additionally, possessions were divided into two groups: short and long duration. We used an arbitrary cut-off of 12.5 seconds, measured from the moment the ball crosses the centerline. Figure 6 shows the posterior median of the ID value for the game CLE vs. GSW. The x axis represents the frame number (at 2.5 frames/second). Overall, the ID shows patterns of spikes and declines during the execution of the play. Short plays tend to have a peak in ID around frames 10–15 ($\approx$ four to six seconds after the ball reaches the offensive court). However, the ID reaches the pinnacle for long possession times at approximately seconds 6–8 or between frames 15–20.

4.2. *Speed and angle.* A potential limitation of the analysis of movement data is that the players' locations at time t are not independent of the locations at time $t-1$, $t-2$, etc. We extend the analysis using the Euclidean distance traveled at every timestamp (speed) by each player and the angle of the trajectory between timestamps. In both cases the original dimension $D = 10$. In detail, let y and x be the positions in the vertical and horizontal axes on the offensive side of the basketball court, respectively. The speed s (as distance per unit of

(a) Simple play (https://www.youtube.com/watch?v=jb57MFQLoRo&t=91s). (b) Complex play (https://www.youtube.com/watch?v=jb57MFQLoRo&t=227s).

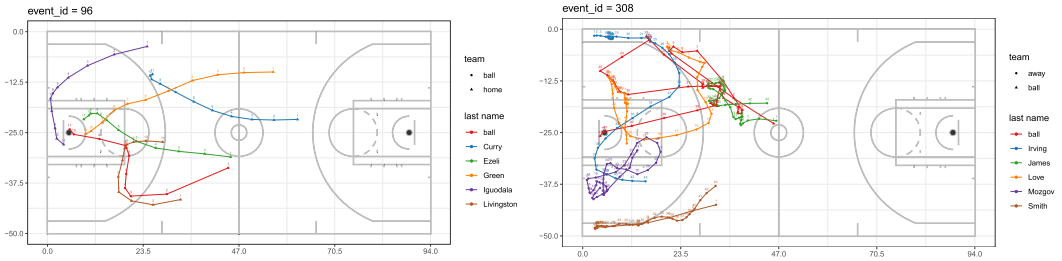

FIG. 5. *Example of players trajectories (indexed by* `last name`*) in simple and complex plays.*

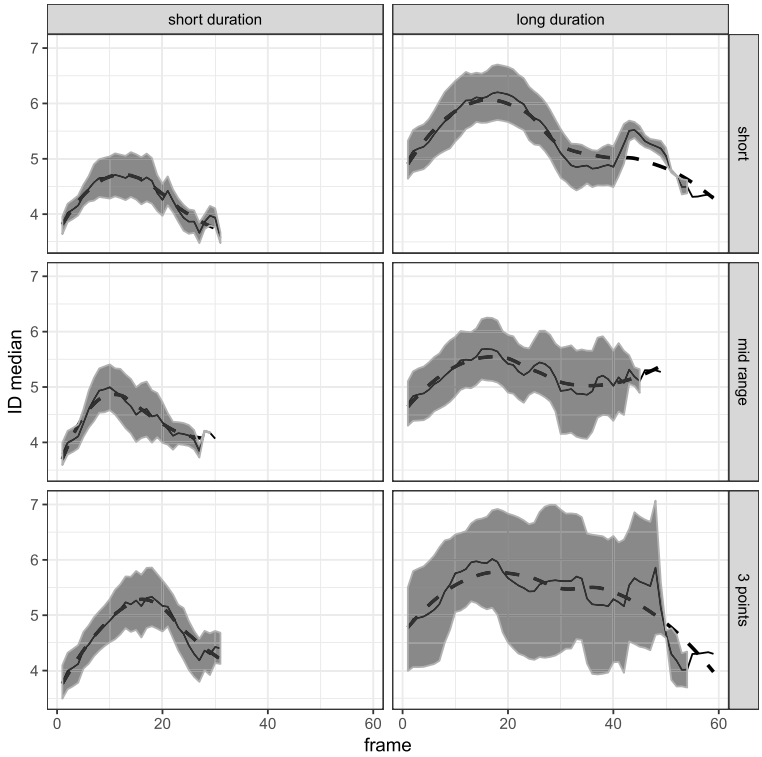

FIG. 6. *Posterior medians and* 95% *credible intervals of the ID during the shots taken by both teams. We consider three distance categories* (`dist_cat`) *short, midrange and three-points attempts as well as short* (*left column*) *and long duration* (*right column*) *possessions* ($t \leq 12.5$ *and* $t > 12.5$ *seconds, respectively*). *The smooth dashed line was obtained by fitting a generalized additive model.*

time) and the angle θ are defined as

$$s_{i,t} = \sqrt{(y_{i,t+1} - y_{i,t})^2 + (x_{i,t+1} - x_{i,t})^2} \tag{7}$$

and

$$\theta_{i,t} = \tan^{-1}\{(y_{i,t+1} - y_{i,t})/(x_{i,t+1} - x_{i,t})\}, \tag{8}$$

where the subscript t represents the time stamp of a play and i denotes the player.

The ID posterior median of the speed and angle datasets are reported in Figure 7. We note a gradual increase in ID on the speed from frames 1 to 20 after which it stabilizes. The players' directions exhibit different behavior, with a sustained increase during the execution of the play. More changes of direction (and, therefore, increments in the play's complexity) happen near the end of the play. We have also investigated how these techniques translate into other team sports, and have found similar ID behaviors. For example, we have conducted preliminary analyses using football tracking data, published on https://github.com/metrica-sports/sample-data; see the Supplementary Material for an example of the evolution of the ID in a play in which the home team scores a goal.

4.3. *ID analysis of shot charts.* In sports like basketball, it is well known that the spatial locations of the attacking and defending players influence the probability of success of a play. Even more important are the locations of the player taking the shot and those guarding them. It has been suggested that a higher variability of the offensive players' locations is associated with attacks that are harder to defend. However, this association has not been

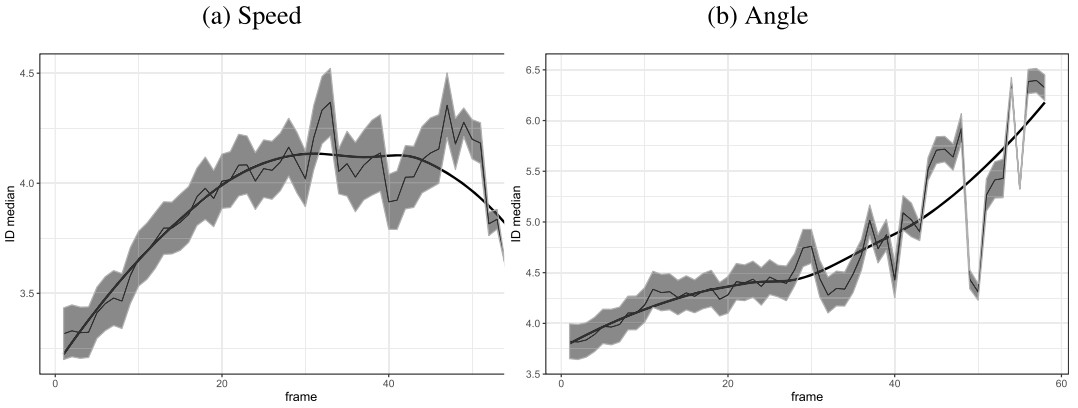

FIG. 7. *Posterior medians and* 95% *credible intervals of the players speed and angle ID per time frame. The smooth line was obtained using a generalized additive model.*

formally assessed before. In this subsection we estimate the ID, simultaneously cluster the shot charts, and examine the association between the ID values and the shots' probability of success.

From each play of the game, we obtained the locations of the players at the moment of the shooting. This point in time was obtained using the z coordinate of the ball (radius). However, we do not consider the location of the ball in the ID estimation. Therefore, for this analysis, we take into account the players' locations and the outcome of a play—missed versus successful shot. We discuss two alternatives. We start with a two-team approach followed by individual team analysis.

4.3.1. *Two-team approach.* We compute the ID using the shot-chart data from the home and away teams. We split the data into two sets as follows. The first set contains the field goal shots, corresponding to when the home team (e.g., GSW) is attacking while the away team (e.g., CLE) is on defense. The second set contains the field goal shots from the away team (CLE) on the attack and the home team on the defense (GSW).

In this case, the number of rows in each dataset is the number of attempted field shots. The number of columns $D = 20$ represents the original dimension of the data (the two coordinates of the player $\{x$ and $y\} \times$ five players $\times$ two teams). The ID, in this context, corresponds to the number of independent directions that embed the 20-dimensional coordinate points.

Clustering by the local ID helps to identify plays that produce substantially lower or higher returns. For example, in Figure 8 we show a heatmap of the posterior coclustering matrix for the plays where CLE was on attack and GSW on defense. Columns and rows were reordered, based on hierarchical clustering, so that plays with a high probability of belonging to the same cluster are grouped. The labels in the y axis represent the game event. Three main clusters are identified (yellow representing a high probability of coclustering between two plays) with an approximately equal number of plays.

For example, $idn = 15$ is the play shown in Figure 1. The right-hand side dot plot shows the outcome of each of the field goals. We depict the missed shots (zero) in orange and the successful shots (one) in green. Plays in cluster 2 (ids between 69 and 112) had a probability of success of 0.21 which contrasts with CLE's 46% field goal success during the season. Clusters 1 and 3 had more than twice this value (0.42 and 0.47, respectively) which is in line with the seasonal average. Similar analyses were carried out for all the teams, identifying clusters of plays that produce above and below-average returns (not shown here due to the limitation of space). Overall, these analyses provide an accurate picture of the game performance, result in a useful tool for coaching, and can be used as the game goes on, if the data is available.

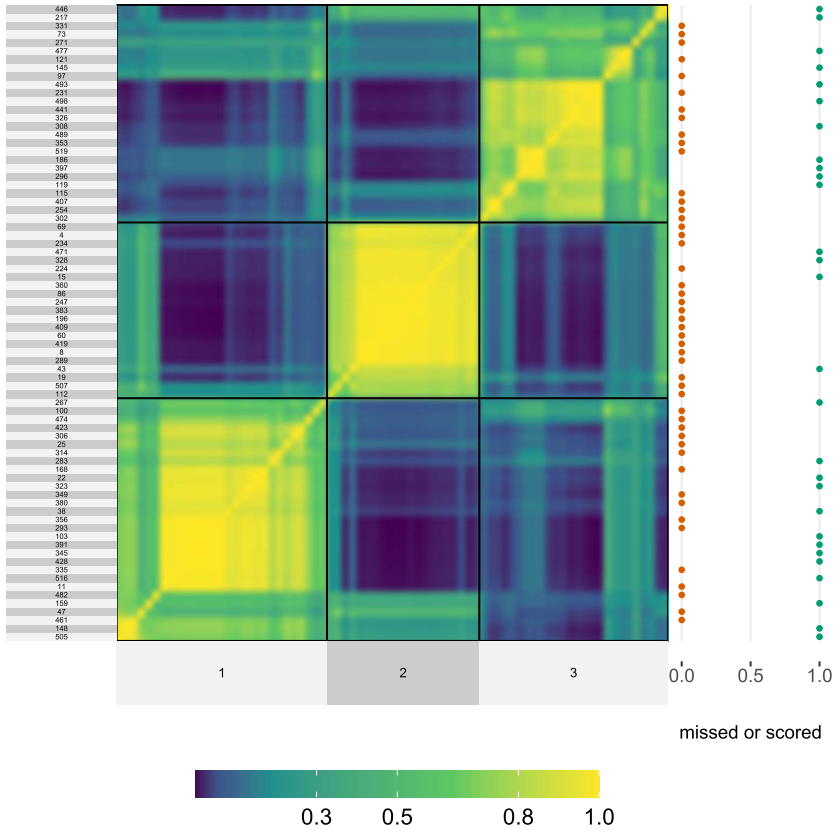

FIG. 8. *Heatmap of the posterior coclustering matrix for the plays where CLE was on attack and GSW on defense. The right dots plot shows the field goals made* (*green dots*) *and missed* (*orange dots*).

We also perform a regression analysis at this individual play level in the shot-charts analysis. We consider the play posterior median ID values as a covariate, along with the distance of the player taking the shot and the shot clock. As expected, the distance from the basket and the probability of scoring are negatively associated. In this regression analysis at the shot level, we do not find evidence that links a higher ID value to a higher probability of scoring. This could be the result of simple plays having a positive outcome, when the defense is poor. However, the relationship is compelling at the game level. We compute the ID values for the shots taken during 15 games of the season. Although the number of games is relatively small, there is a positive association between the overall posterior median of the ID and the game outcome. The boxplots in Figure 9 show the ID for the winning and losing teams. We represent each game by a solid line connecting both teams from the left to the right boxplot. The home team is denoted by the symbol *. For completeness we have also added the final score to the connecting line and the p-value within brackets. The color of the line signals the differences between the posterior medians, using a Mann–Whitney test. In six of these games, the winner had a substantially greater ID. In another seven, there was no difference in the ID between winners and losers, and in the remaining two games, the losers had higher ID values. These results are summarized in Appendix B (Table 2).

Plays generally follow the path of least resistance. We argue that plays tend to have an increased movement complexity when the difference in the score is small, usually due to a tighter defense. We fitted a Bayesian linear model, using as response variable the posterior ID, and the score margin category as a covariate. In Figure 10 we show the marginal effects of different score margins categories {small (zero to five points), medium (six to 10 points),

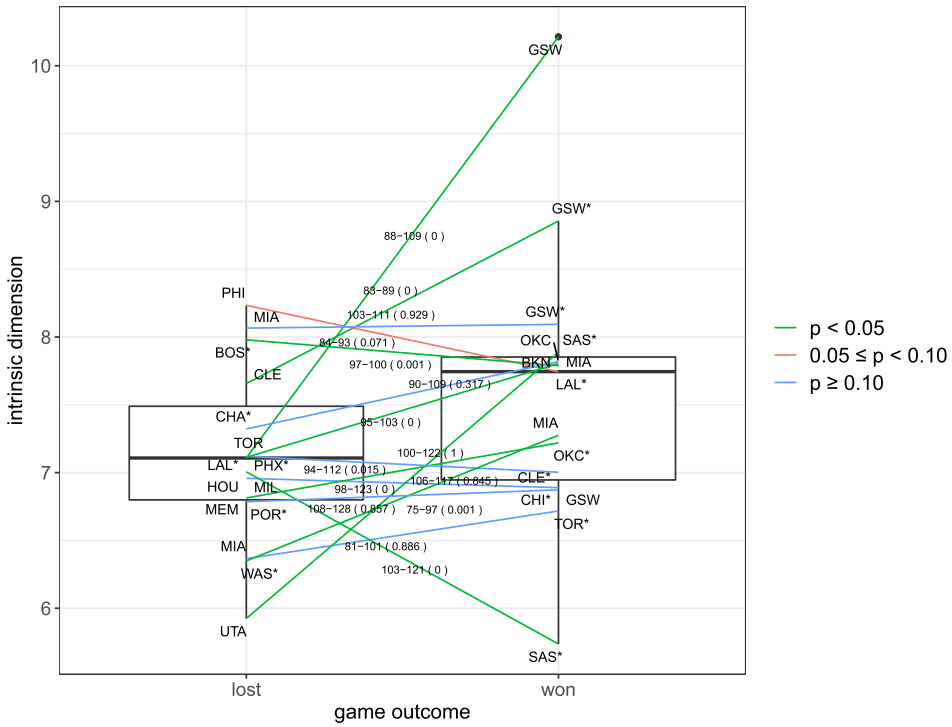

FIG. 9. *Boxplots of the posterior ID of the winners and losers in the* 15 *matches. Each game is represented by a line connecting the two involved teams. The color of the line represents the category in the differences between the posterior distributions of the median ID using a Mann–Whitney test. The game score (points scored by the loser vs. points scored by the winner) is shown in the center of each connecting line in addition to the p-value within brackets.*

large (11–15 points) and huge (≥16 points)} on the posterior ID. This evidence supports the argument that the smaller the score margin, the greater the ID and the complexity. Note the substantial differences between the ID distributions in small and huge score margins.

Similarly, pairwise comparisons, using the Wilcoxon rank-sum test, support this conjecture; see the results in the Appendix (Table 5).

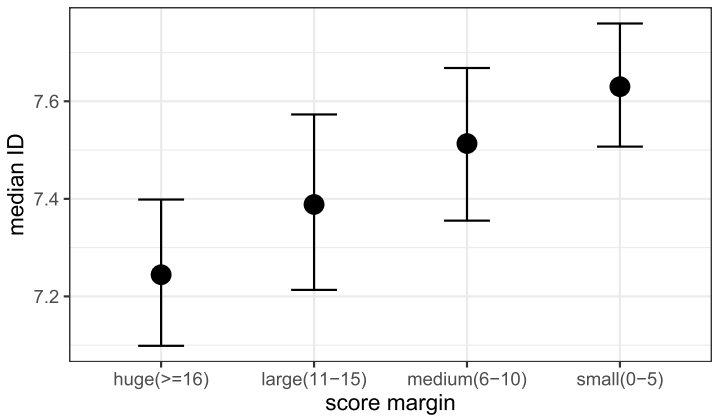

FIG. 10. *Marginal effects of the score margin on the posterior ID.*

We perform a similar analysis for individual datasets, comprising the locations of the five players from each team in attack and then when they are in a defensive role. We have placed these results in Appendix C.

5. Discussion and conclusions. The advent of sports tracking technology is flooding sports analytics and sports data science with large and more complex datasets (Lazar (2014)). Especially in basketball, each game could result in the generation of massive datasets. The high dimension of these datasets makes obtaining competitive advantages and the extraction of meaningful insights, as the indicators of players' (or teams') performance, increasingly challenging. The inference process on key performance measures is also made more laborious. As a result, researchers and practitioners are resorting to dimensionality reduction techniques, so such big datasets can be synthesized, handled, and interpreted more conveniently.

The purpose of the current study is to present a different perspective in the analysis of high-resolution player tracking data from invasion sports, particularly from the NBA. We used the IDs extracted from each player's positions (x, y) in Cartesian coordinates to deepen our comprehension of the game dynamics. In particular, we have determined different stages in the execution of offensive actions and have identified clusters in shot-chart data. Moreover, we have compared and assessed the relationship between the ID and game outcomes. We have employed a model-based methodology developed by Allegra et al. (2020). The mixture model, on which it is founded, allows the segmentation of the observations in groups of homogeneous ID. Building on their proposal, we have introduced different enhancements, ranging from more meaningful prior distributions to better postprocessing of the MCMC output. From the methodological point of view, further work should be done to overcome the limitations of this approach. First, the choice of K, the number of mixture components, does not consider any form of uncertainty. A Bayesian nonparametric extension, employing, for example, a Dirichlet process mixture model (Antoniak (1974), Escobar and West (1995)), would resolve this issue. Second, the analysis of how the ID changes across time frames provides valuable insights but does not satisfy the hypothesis of independence across the observation. This issue paves the way to an interesting research path, where one can combine the model-based ID estimation framework with hidden Markov models (Baum and Petrie (1966)) to take into account the temporal dependence. Finally, the statistical model does not consider any source of noise in the measurement process which can slightly inflate the estimates of the different IDs.

The results show that we can satisfactorily distinguish groups of plays with lower and higher than average returns, according to their ID, which we can interpret as the complexity of a play type. These methods reveal the hidden dynamics for the players' interactions in team sports and translate complex movement patterns into more coherent statistics. Our proposal can guide coaches in planning and designing more effective attacking and defensive plays and provide new competitive advantages.

We found a link between higher games' median ID values and higher play uncertainty and unpredictability in the attack. These results are in line with previous findings: see, for example, Hobbs et al. (2018), where the authors discuss entropy on ball possessions. We also found evidence that larger game ID values from shot charts data are positively associated with winning games. However, this claim needs to be validated with larger sample sizes.

This approach also enhances our understanding of how players' moving tactics and interactions impact the outcome of a game. Stages like ball handling, creating space for passing, shooting, and following through have different characteristics and can be identified using the coclustering matrix along with the median ID curve. We observed an increase in ID values when the players are creating an opportunity for passing and shooting. This behavior is expected, as players on attack tend to move with higher uncertainty and entropy, using, for

example, screen actions. Similarly, we found a decline in ID toward the end of the plays. In those moments the players follow through shots or return to the opposite end of the court.

Although we focus on basketball, the approach we introduce here can be employed to analyze data available from other team sports where player tracking data are available, due to the increased use of global positioning systems or multiple camera systems in sports venues. Examples of such datasets can be found in the National Football League (NFL) (https://github.com/nfl-football-ops/Big-Data-Bowl), football (soccer) (e.g., Pappalardo et al. (2019), De Silva et al. (2018)), and in volleyball (Van Haaren et al. (2016)), to mention a few. More generally, the methodology and analysis can be extended to other fields, such as ecology for modeling animal movement and trajectory data (e.g., Dodge et al. (2013), Mastrantonio et al. (2019)), predator–prey interactions, etc.

We hope that this Bayesian approach can complement manual video game analysis, providing effective clustering in match analyses. This combination, in turn, could produce new insights, usable by coaching staff, to guide the design of more successful strategies. We remark that these analyses can be performed on other sports such as football, NFL, volleyball, and rugby which have implemented a player tracking technology or produce players' spatial data.

As a starting point, we have reported preliminary analyses of football tracking data in the Supplementary Material (Santos-Fernandez et al. ((2022), Section 1.2)). We have found similar patterns in the ID dynamics during the execution of plays, that is, an ID increment as the ball reaches the goal, followed by a decline at the end of the play. We believe that our methodology has the potential to produce new insights, derived from the analysis of movement data in team sports, and opening multiple avenues for further research.

Similarly, the ID analysis can be useful in a large number of applications that go well beyond those discussed here. For example, we are currently investigating the ID of gene expression datasets, grouping patients' gene expression profiles in clusters capable of discriminating different clinical characteristics, such as survival time, disease type, etc.

APPENDIX A: POSTERIOR ID VALUES BASED ON THREE PRIOR TYPES

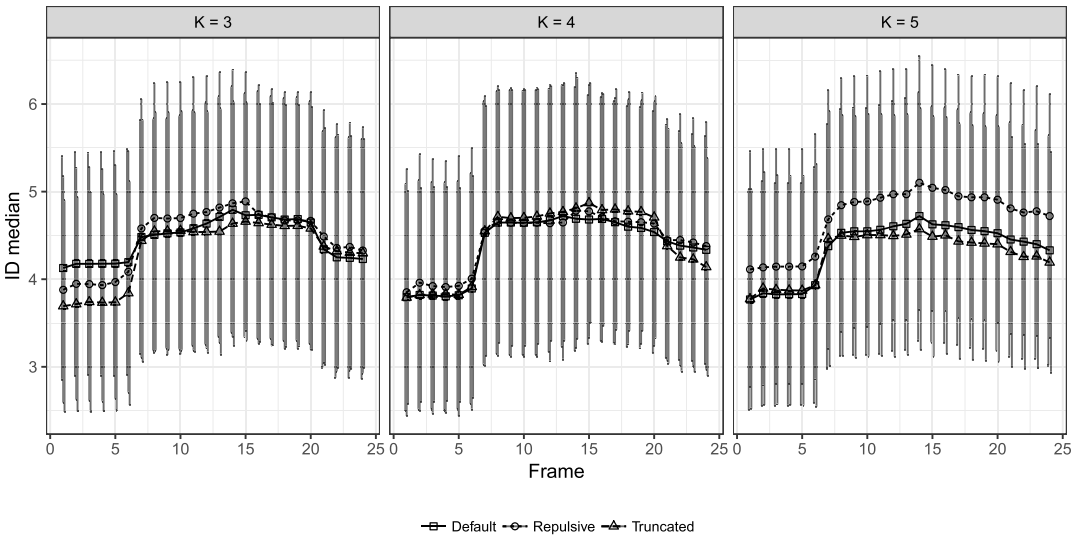

FIG. 11. *Comparison of the posterior median ID values obtained using three methods* (*represented with different point shapes and line types*) *and several mixture components. The results show consistency independently of the method and the number of components used.*

APPENDIX B: OTHER RESULTS

(a) Movement of two players and the ball.

(b) Heatmap of the 30 time stamps during the play (bottom) and posterior medians of the ID (top).

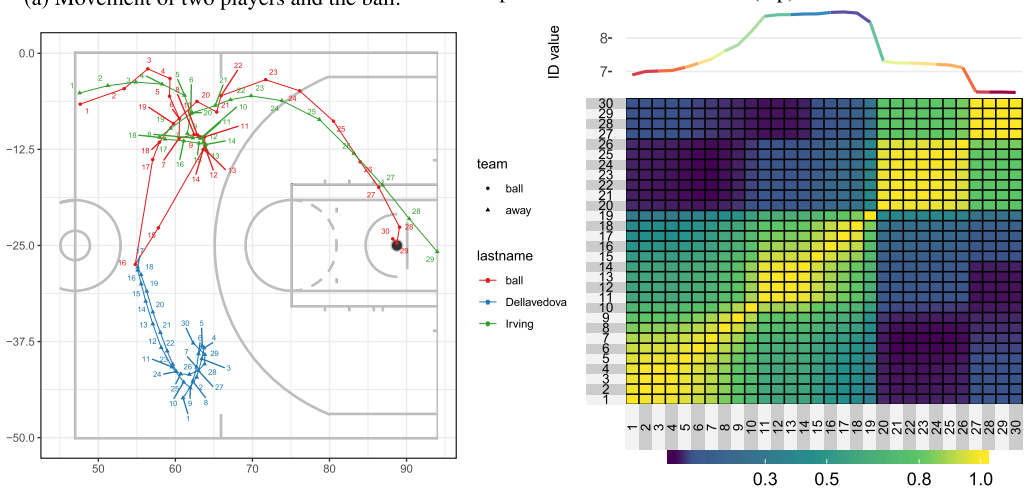

FIG. 12. *(a) Trajectory of the players/ball and (b) coclustering matrix (lower right) and ID in a two-point driving bank shot. https://youtu.be/jb57MFQLoRo?t=180.*

TABLE 2

Posterior median ID values and the p-values from the comparison in the 15 *games analysed during the season* 2015–16

Away	Home	Date	Result	ID Away	ID Home	p-val
GSW	LAL	5-Jan-2016	109–88	10.214	7.108	0.000
MIL	CHI	5-Jan-2016	106–117	6.958	6.889	0.845
MIA	TOR	22-Jan-2016	81–101	6.366	6.717	0.886
CLE	GSW	25-Dec-2015	83–89	7.657	8.855	0.000
HOU	SAS	2-Jan-2016	103–121	7.006	5.737	0.000
PHI	LAL	1-Jan-2016	84–93	8.234	7.745	0.071
MEM	OKC	6-Jan-2016	94–112	6.813	7.219	0.015
UTA	SAS	6-Jan-2016	98–123	5.925	7.886	0.000
BKN	BOS	2-Jan-2016	100–97	7.788	7.979	0.001
TOR	CLE	4-Jan-2016	100–122	7.120	7.003	1.000
MIA	GSW	11-Jan-2016	103–111	8.066	8.093	0.929
OKC	CHA	2-Jan-2016	109–90	7.818	7.322	0.317
MIA	WAS	3-Jan-2016	97–75	7.276	6.349	0.001
MIA	PHX	8-Jan-2016	103–95	7.804	7.113	0.000
GSW	POR	8-Jan-2016	128–108	6.873	6.786	0.857

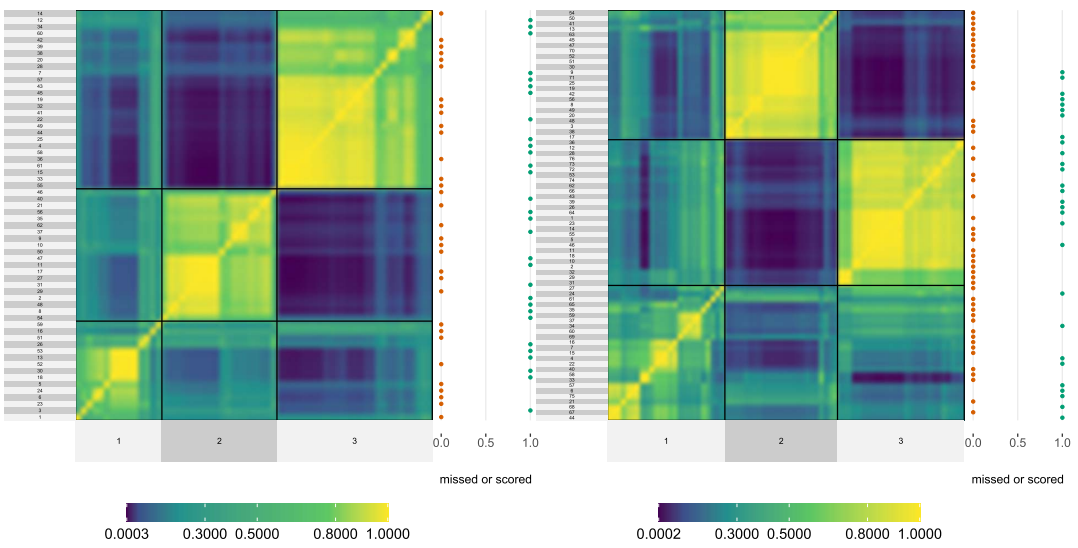

FIG. 13. *Heatmaps and clusters of the shot-chart plays by GSW in attack* (*a*) *and defense* (*b*). *The x axis gives the cluster, and the y axis represent the play number. The top dots plot shows the field goals made* (*green dots*) *and missed* (*orange dots*).

APPENDIX C: INDIVIDUAL TEAM APPROACH

We carried out further analysis using individual teams' data. In this case the dimension is $D = 10$ (location in x and y $\times$ five players). This analysis yields for each team clusters of shot charts with a low and high return in offense and defense. Figure 13 shows the posterior coclustering heatmaps of the plays by GSW. On each of the plots, we defined three clusters. For instance, in subfigure (a) finding cluster 1 in the x-axis, we find that plays $59, 16, \ldots, 1$ have a large probability of belonging to this cluster (in yellow color). The outcome of each play is represented in the dot plot on the right-hand side. Table 3 gives the proportion of successful plays in the clusters. Forty percent of these offensive shots in cluster 1 (a) were successful. Similarly, 55% of the attacking plays in the second cluster were scored.

In (b) we show the clusters from the defensive placements of GSW. For example, cluster 3 shows the weakest defensive outcome for GSW, allowing 40.7% scoring by CLE.

Furthermore, in Figure 14 we present the posterior coclustering of CLE in attack (a) and defense (b). From (b) cluster 1 contains plays where the defense by CLE was ineffective, allowing 50% success for GSW. These results by CLE are summarised in Table 4.

TABLE 3
Probability of success in the offensive and defensive roles for GSW

Role	Cluster	p success
attack	1	0.400
attack	2	0.550
attack	3	0.481
defense	1	0.360
defense	2	0.333
defense	3	0.407

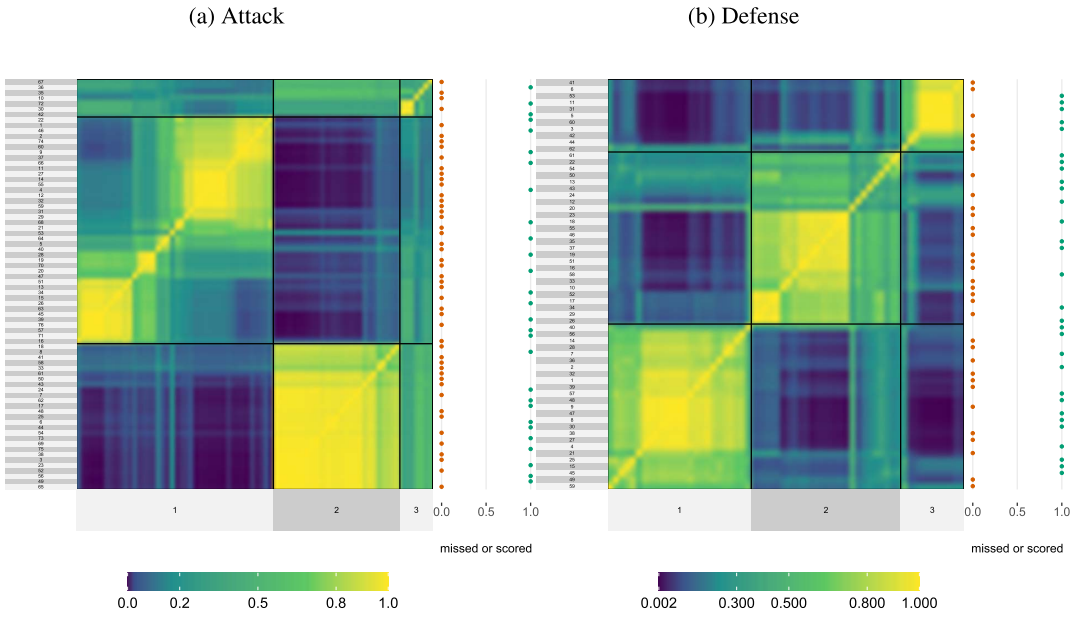

FIG. 14. *Heatmaps and clusters of the shot chart plays by CLE in attack* (*a*) *and defense* (*b*). *The x axis gives the cluster and the y axis represent the play number. The top dots plot shows the field goals made* (*green dots*) *and missed* (*orange dots*).

TABLE 4
Probability of success in the offensive and defensive roles for CLE

Role	Cluster	*p* success
attack	1	0.333
attack	2	0.407
attack	3	0.429
defense	1	0.500
defense	2	0.462
defense	3	0.455

TABLE 5
Pairwise comparisons (*p-values*) *of the distributions of the ID for different scoring margins based on the Wilcoxon rank-sum test. The alternative hypothesis is*: *the category in the rows has greater median ranks than the one in the columns*

	Huge (≥16)	Large (11–15)	Medium (6–10)
Large (11–15)	0.285	·	·
Medium (6–10)	<0.0001	0.334	·
Small (0–5)	<0.0001	0.010	0.505

Acknowledgments. We thank the two anonymous reviewers, the Editor, and the Associate Editor for carefully reading the manuscript and their insightful and constructive comments. These helped to improve the manuscript substantially. All computations and visualizations were carried using R using the packages mcclust (Fritsch (2012)), superheat (Barter and Yu (2017)), tidyverse (Wickham (2017)), gganimate (Pedersen and Robinson (2019)) and ggrepel (Slowikowski (2019)).

Funding. This research was supported by the Australian Research Council (ARC) Laureate Fellowship Program, the Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), and the project "Bayesian Learning for Decision Making in the Big Data Era" (ID: FL150100150), First Investigator: Prof. Kerrie Mengersen. We thank Wade Hobbs and Timothy Macuga for their comments and suggestions. During the development of this article, F. Denti was funded as a postdoctoral scholar by the NIH grant R01MH115697 grant. Previously, he was also supported as a Ph.D. student by University of Milano—Bicocca, Milan, Italy, and Università della Svizzera italiana, Lugano, Switzerland.

SUPPLEMENTARY MATERIAL

ID in football (DOI: 10.1214/21-AOAS1506SUPPA; .pdf). Example of the evolution of the ID values in a football play that produced a goal for the home team.

Codes and data (DOI: 10.1214/21-AOAS1506SUPPB; .zip). It contains some of the R codes and the dataset used in Section 4.

REFERENCES

ALLEGRA, M., FACCO, E., DENTI, F., LAIO, A. and MIRA, A. (2020). Data segmentation based on the local intrinsic dimension. *Sci. Rep.* **10** 16449.

ANSUINI, A., LAIO, A., MACKE, J. H. and ZOCCOLAN, D. (2019). Intrinsic dimension of data representations in deep neural networks.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969 https://doi.org/10.1214/aos/1176342871

BARTER, R. and YU, B. (2017). superheat: A graphical tool for exploring complex datasets using heatmaps. R package version 0.1.0.

BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37** 1554–1563. MR0202264 https://doi.org/10.1214/aoms/1177699147

BENNETT, R. S. (1969). The intrinsic dimensionality of signal collections. *IEEE Trans. Inf. Theory* **15** 517–525. https://doi.org/10.1109/TIT.1969.1054365

CALIŃSKI, T. and HARABASZ, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* **3** 1–27. MR0375641 https://doi.org/10.1080/03610927408827101

CAMASTRA, F. and STAIANO, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Inform. Sci.* **328** 26–41.

CAMPADELLI, P., CASIRAGHI, E., CERUTI, C. and ROZZA, A. (2015). Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Math. Probl. Eng.* Art. ID 759567, 21 pp. MR3417646 https://doi.org/10.1155/2015/759567

CELEUX, G. (1998). Bayesian inference for mixture: The label switching problem. *Compstat* 227–232. https://doi.org/10.1007/978-3-662-01131-7_26

CERVONE, D., D'AMOUR, A., BORNN, L. and GOLDSBERRY, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *J. Amer. Statist. Assoc.* **111** 585–599. MR3538688 https://doi.org/10.1080/01621459.2016.1141685

D'AMOUR, A., CERVONE, D., BORNN, L. and GOLDSBERRY, K. (2015). Move or die: How ball movement creates open shots in the NBA. In *MIT Sloan Sports Analytics Conference*.

DE SILVA, V., CAINE, M., SKINNER, J., DOGAN, S., KONDOZ, A., PETER, T., AXTELL, E., BIRNIE, M. and SMITH, B. (2018). Player tracking data analytics as a tool for physical performance management in football: A case study from Chelsea football club academy. *Sports* **6** 130.

DODGE, S., BOHRER, G., WEINZIERL, R., DAVIDSON, S. C., KAYS, R., DOUGLAS, D., CRUZ, S., HAN, J., BRANDES, D. et al. (2013). The environmental-data automated track annotation (Env-DATA) system: Linking animal tracks with environmental data. *Mov. Ecol.* **1** 3.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510 https://doi.org/10.1080/01621459.1995.10476550

FACCO, E., D'ERRICO, M., RODRIGUEZ, A. and LAIO, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* **7** 12140. https://doi.org/10.1038/s41598-017-11873-y

FRANKS, A., MILLER, A., BORNN, L. and GOLDSBERRY, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *Ann. Appl. Stat.* **9** 94–121. MR3341109 https://doi.org/10.1214/14-AOAS799

FRITSCH, A. (2012). mcclust: Process an MCMC sample of clusterings. R package version 1.0.

FRÜHWIRTH-SCHNATTER, S. (2011). Dealing with label switching under model uncertainty. In *Mixtures: Estimation and Applications*. *Wiley Ser. Probab. Stat.* 213–239. Wiley, Chichester. MR2883354 https://doi.org/10.1002/9781119995678.ch10

GOLDSBERRY, K. (2012). Courtvision: New visual and spatial analytics for the NBA. In 2012 *MIT Sloan Sports Analytics Conference*.

HOBBS, W., MORGAN, S., GORMAN, A. D., MOONEY, M. and FREESTON, J. (2018). Playing unpredictably: Measuring the entropy of ball trajectories in international women's basketball. *Int. J. Perform. Anal. Sport* **18** 115–126.

LAMAS, L., BARRERA, J., OTRANTO, G. and UGRINOWITSCH, C. (2014). Invasion team sports: Strategy and match modeling. *Int. J. Perform. Anal. Sport* **14** 307–329.

LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16** 526–558. MR2351079 https://doi.org/10.1198/106186007X238855

LAZAR, N. (2014). The big picture: Take me out to the ball game. *Chance* **27** 45–48.

LEVINA, E. and BICKEL, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems* 777–784.

LUCEY, P., BIALKOWSKI, A., CARR, P., FOOTE, E. and MATTHEWS, I. A. (2012). Characterizing multi-agent team behavior from partial team tracings: Evidence from the English Premier League. In *AAAI*.

LUTZ, D. (2012). A cluster analysis of NBA players. In *MITSloan Sports Analytics Conference*.

MASTRANTONIO, G., GRAZIAN, C., MANCINELLI, S. and BIBBONA, E. (2019). New formulation of the logistic-Gaussian process to analyze trajectory tracking data. *Ann. Appl. Stat.* **13** 2483–2508. MR4037438 https://doi.org/10.1214/19-aoas1289

METULINI, R. (2018). Players movements and team shooting performance: A data mining approach for basketball. Preprint. Available at arXiv:1805.02501.

METULINI, R., MANISERA, M. and ZUCCOLOTTO, P. (2017). Space-time analysis of movements in basketball using sensor data. Preprint. Available at arXiv:1707.00883.

NISTALA, A. and GUTTAG, J. (2019). Using deep learning to understand patterns of player movement in the NBA. In *Proceedings of the MIT Sloan Sports Analytics Conference* 1–14.

PAPPALARDO, L., CINTIA, P., ROSSI, A., MASSUCCO, E., FERRAGINA, P., PEDRESCHI, D. and GIANNOTTI, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Sci. Data* **6** 1–15.

PEDERSEN, T. L. and ROBINSON, D. (2019). gganimate: A grammar of animated graphics. R package version 1.0.4.

PETRALIA, F., RAO, V. and DUNSON, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems* 1889–1897.

ROBERT, C. P. (2010). Multimodality and label switching: A discussion. In *Workshop on Mixtures*, *ICMS*.

RODRÍGUEZ, C. E. and WALKER, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *J. Comput. Graph. Statist.* **23** 25–45. MR3173759 https://doi.org/10.1080/10618600.2012.735624

ROSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65.

ROZZA, A., LOMBARDI, G., CERUTI, C., CASIRAGHI, E. and CAMPADELLI, P. (2012). Novel high intrinsic dimensionality estimators. *Mach. Learn.* **89** 37–65. MR2967955 https://doi.org/10.1007/s10994-012-5294-7

SAMPAIO, J., DRINKWATER, E. J. and LEITE, N. M. (2010). Effects of season period, team quality, and playing time on basketball players' game-related statistics. *Eur. J. Sport Sci.* **10** 141–149.

SAMPAIO, J., MCGARRY, T., CALLEJA-GONZÁLEZ, J., SÁIZ, S. J., I DEL ALCÁZAR, X. S. and BALCIUNAS, M. (2015). Exploring game performance in the National Basketball Association using player tracking data. *PLoS ONE* **10** e0132894.

SANTOS-FERNANDEZ, E., DENTI, F., MENGERSEN, K. and MIRA, A. (2022). Supplement to "The role of intrinsic dimension in high-resolution player tracking data—Insights in basketball." https://doi.org/10.1214/21-AOAS1506SUPPA, https://doi.org/10.1214/21-AOAS1506SUPPB

SHORTRIDGE, A., GOLDSBERRY, K. and ADAMS, M. (2014). Creating space to shoot: Quantifying spatial relative field goal efficiency in basketball. *J. Quant. Anal. Sports* **10** 303–313. https://doi.org/10.1515/jqas-2013-0094

SKINNER, B. and GOLDMAN, M. (2017). Optimal strategy in basketball. In *Handbook of Statistical Methods and Analyses in Sports*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 229–244. CRC Press, Boca Raton, FL. MR3837238

SLOWIKOWSKI, K. (2019). ggrepel: Automatically position non-overlapping text labels with 'ggplot2'. R package version 0.8.1.

SPERRIN, M., JAKI, T. and WIT, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Stat. Comput.* **20** 357–366. MR2725393 https://doi.org/10.1007/s11222-009-9129-8

TERAMOTO, M., CROSS, C. L., RIEGER, R. H., MAAK, T. G. and WILLICK, S. E. (2018). Predictive validity of National Basketball Association draft combine on future performance. *J. Strength Cond. Res.* **32** 396–408.

VAN HAAREN, J., BEN SHITRIT, H., DAVIS, J. and FUA, P. (2016). Analyzing volleyball match data from the 2014 World Championships using machine learning techniques. In *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 627–634.

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. MR3807860 https://doi.org/10.1214/17-BA1073

WICKHAM, H. (2017). tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1.