On the intrinsic dimension of basketball high resolution data

Edgar Santos-Fernandez^{a,b}, Francesco Denti^{c}, Kerrie L. Mengersen^{a,b}, Antonietta Mira^{c}

[a]*School of Mathematical Sciences. Y Block, Floor 8, Gardens Point Campus. Queensland University of Technology. GPO Box 2434. Brisbane, QLD 4001. Australia.*
[b] *Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers(ACEMS)*
[c]*Università della Svizzera italiana. Lugano, Switzerland*

Abstract

A new range of statistical analysis has emerged in sports after the introduction of the high-resolution player tracking technology, specifically in basketball. However, this high dimensional data is often challenging for statistical inference. We used a Bayesian non-parametric algorithm based on the intrinsic dimension (ID) of the dataset that gives and an indication of uniqueness and complexity. This technique allows classification and clustering of NBA basketball shot charts and player's movement data. We found groups of shots that produce a substantially higher and lower successes. Overall game winners tend to have a larger intrinsic dimension which is an indication of more unique shot placements and unpredictability. Similarly, we found higher ID values in plays when the score margin is small compared to large margin ones. Analyzing movement data, the algorithm identifies key stages on offensive actions such as creating space for passing, preparation/shooting and following through. We found that the ID value spikes reaching a peak between 4 and 8 seconds in the offensive part of the court after which it declines. These outcomes can perhaps be exploited by coaches to obtain better offensive/defensive results.

Keywords: basketball, Bayesian clustering, data complexity, intrinsic dimension, plays classification, movement data, shot charts

1. Introduction

Basketball is a highly dynamic invasion sport, in which a team aims to score in the opposing team's basket. Teams use a large variety of trained plays seeking an increase in the chances of scoring. The introduction of the SportVU NBA player tracking technology brought player movement measurements at 25 frames per second. These high resolution data have motivated several spatial and spatio-temporal statistical analyses (e.g. Goldsberry, 2012; Shortridge et al., 2014; Cervone et al., 2016). However, such high dimensional data are often challenging for statistical inference, computationally expensive and require more sophisticated statistical techniques. It is well known that the placement of the players in attack and defense, and particularly the guard to the player taking the shot are related to the success of a play. However, effective measures of this placement are yet to be developed, which could be useful to measure and improve teams and players' performance.

Recently suggested measures like ball entropy, uncertainty and unpredictability have been regarded as key performance factors in sports games (Lucey et al., 2012; D'Amour et al., 2015; Skinner & Goldman, 2015; Hobbs et al., 2018). In this regard, Skinner & Goldman (2015) pointed out that the expected value in a play will decrease as it is used more often. Similarly, it is argued that teams that have more versatile players in attack produce successful shots from more unique locations in the court and they create more shooting opportunities by passing the ball more effectively. This yields a greater success for the attacking team, making it harder for the defending one due to the greater shooting uncertainty.

In basketball, a large number of individual statistics are collected nowadays. On one hand, teams monitor the players' traditional summary statistics e.g.: the number of points (PTS), defensive rebounds (DREB), assists (AST), field goals made (FGM), 3 Point Field Goals Made (3PM), minutes (MIN), etc. On top of that, several other metrics are estimated from tracking technology: distance feet (Dist. Feet), average speed (Avg Speed), passes

Email addresses: edgar.santosfdez@gmail.com (Edgar Santos-Fernandez), santosfe@qut.edu.au (Edgar Santos-Fernandez), francesco.denti@usi.ch (Francesco Denti), k.mengersen@qut.edu.au (Kerrie L. Mengersen), antonietta.mira@usi.ch (Antonietta Mira)

made and received, etc. This large number of variables × 15 players on active roster × 82 games/season make univariate analysis and comparisons extremely laborious.

Hence, multivariate statistical techniques like clustering are becoming increasingly popular among sports scientists. These methods, despite their greater complexity, allow a better communication of performance to coaches. Lutz (2012), for example, used statistics such as field goals, steals and assist ratio to cluster players with similar features into 10 categories. Other clustering applications can be found in Metulini et al. (2017) and Metulini (2018). In another example of clustering Franks et al. (2015) used nonnegative matrix factorization (NMF) to group defensive players using field goals locations. This approach provides a measure of the impact of defensive players on shot frequency and probability of scoring.

Specifically, clustering algorithms such as k-means have been used for analyzing basketball data. Sampaio et al. (2015), for instance, grouped players based on performance using attacking, defense and passing statistics. More recently, Nistala & Guttag (2019) used clustering for the classification of players' movement based on Euclidean distance. They clustered attacking movements into 20 groups including screen action, movement along each sideline, run along the baseline, etc. Other examples of applications of dimensionality reduction techniques like principal components analysis (PCA) in basketball can be found in Sampaio et al. (2010) and Teramoto et al. (2018).

Positions and movements of players on attack and defense are generally correlated since defensive players guard those in attack. Hence, using techniques like intrinsic dimension makes possible a reduction of redundant data for statistical analysis.

However, little attention has been paid to the complexity of the players' placements in shot charts. Similarly, we found no discussion on the uniqueness of players' movements during possession times.

We add some definitions before delving into the analysis:

- A *possession* means to be in control of the ball.
- The team on *offense* refers to the team handling the ball and trying to score in the basket. The team on *defense* is the team preventing the other one from getting points scored.
- We refer to a *play* as one action that starts when the team gets possession of the ball and it ends when they lose it. For example, team A gets a ball possession after team B scores. The play finishes when team A shots and team B gets the rebound. The play encompasses the movements of the players during the possession. Every play has an outcome (e.g. scored, missed, etc). A game is composed of a large number of plays.
- *Shot chart* refers here to the positions in the x and y axes of the players when the shot was taken. This is different from the traditional shot charts that consist only of the location of the player taking the shot.

Specifically, the purpose of this paper is to:

1. study patterns in shot charts using Bayesian statistical clustering identifying of plays that produce better outcomes;
2. examine if unpredictability in attack and intrinsic dimension are linked to better performance;
3. assess players' movement complexities in 3 points, mid-range and close shots;
4. identify the phases in the execution of a shot (ball handling, creating space for passing, preparation/shooting and following through).

The rest of the article has been divided into three parts. In the next section, we provide a short introduction to the intrinsic dimension of the data and describe of the data used for analysis. This is followed by the Results section in which we examine shot chart data first, followed by an analysis of players' trajectories. Finally, section 4 concludes with a discussion of the finding and limitations.

2. Materials and methods

2.1. Intrinsic dimension of the data

Generally, higher dimensional datasets can be reduced to a subspace of smaller dimensions known as the intrinsic dimension (ID) without losing much information (Levina & Bickel, 2005; Camastra & Staiano, 2016). The ID gives an indication of the complexity, redundancy and unique features in the dataset. Other definitions describe the ID as the number of degrees of freedom.

Several methods for estimating the ID of the data have been developed. Recently, Facco et al. (2017) suggested ID estimation approaches based on the two nearest neighbors (TWO-NN). Following this Allegra et al. (2019) introduced a Bayesian non-parametric modification based on a Dirichlet Process mixture model.

Their approach is broadly described as follows. Consider a dataset X of dimension $N \times D$, where N is the number of observations (or plays). D is the original dimension or the number of variables, which in this case is the coordinates $\times$ the number of players.

Let $\sqrt{\sum_{i=1}^{D}\left(x_i - y_i\right)^2}$ be the Euclidean distance between the vectors of observations or plays x and y. Denoting r_{i1} and r_{i2} as the first and second nearest neighbors respectively to the i^{th} vector of observations. The statistic μ_i is a measure of closeness between this play and the two nearest neighbors.

$$\mu_i = r_{i2}/r_{i1} \tag{1}$$

Allegra et al. (2019) assumes the statistic μ to be a mixture of Pareto distributions with parameter d that represents the intrinsic dimension, being $d < D$. Using a Dirichlet Process mixture model with k components:

$$P\left(\mu_i | d, p\right) \doteq \sum_{k=1}^{K} p_k \text{Pareto}\left(d\right) \doteq \sum_{k=1}^{K} p_k d_k \mu_i^{-(d_k+1)} \tag{2}$$

where p_k are the mixing proportions representing the probability that a point belongs to the manifold k. A Dirichlet distribution with concentration parameter c is used for p_k.

$$\text{Dir}\left(p_1, \cdots, p_k; c_1, \cdots, c_k\right) = \frac{\Gamma\left(\sum_{k=1}^{K} c_k\right)}{\prod_{k=1}^{K}\Gamma\left(c_k\right)} \prod_{k=1}^{K} p_i^{c_i - 1} \tag{3}$$

with $\sum p_k = 1$, $c_k > 0$. The posterior distribution of the parameter d cannot be obtained analytically but using Markov chain Monte Carlo simulations. A gamma distribution with a large variance is used as a prior on the parameter d, e.g. $d \sim \text{Gamma}\left(a = 1.25, b = 0.25\right)$. A non-informative prior is used ($c_k = 1$). See Allegra et al. (2019) for more details.

The Bayesian nonparametric prior assumed in the model allows the estimation of the ID for every observation or play. Given the discrete nature of the Dirichlet process, a potentially different clustering structure is induced in the data at each iteration of the Gibbs sampler, enabling at the same time the estimation of the best partition and the quantification of the uncertainty around it. In particular, we can compute the pairwise co-clustering matrix, where each entry describes the probability that two observations are clustered together.

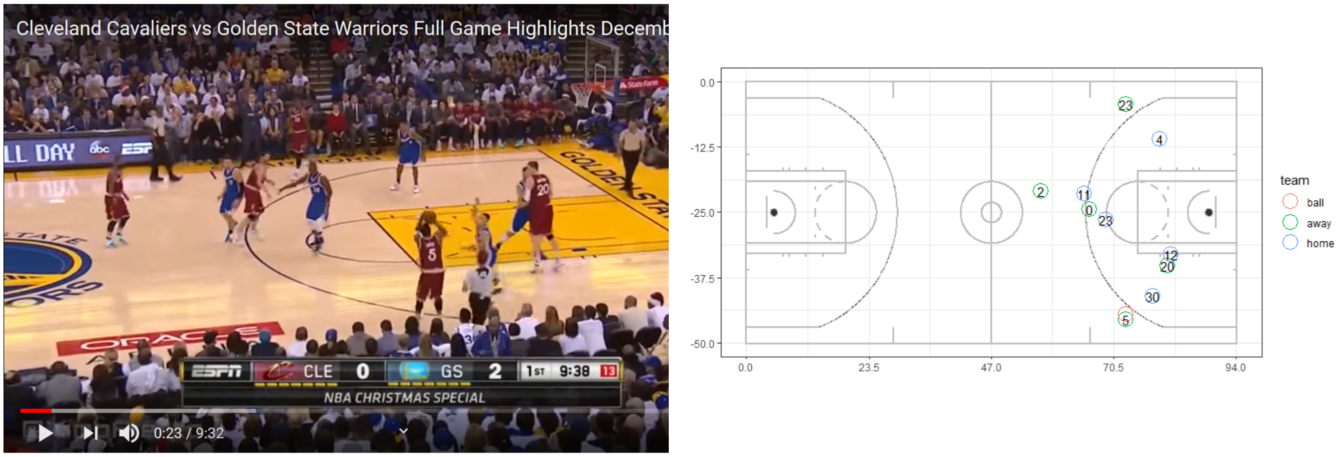
2.2. Description of the dataset

We used STATS SportVU high-resolution player tracking raw data from the NBA season 2015-16. We obtained the play-by-play events description and other statistics from the official website `https://stats.nba.com/`. The match between event in these two files was verified via manual video annotation of the game available on `https://www.youtube.com/`. From each play, we inferred the locations of the players at the moment of the shoot. This point in time was obtained using the z coordinate of the ball (radius). We selected the events = {ShotMissed, ShotMade} and the location of the ball was not considered in the ID estimation. The raw movement data needed curation and manual matching which is time-consuming. Therefore for the purpose of this research, we considered 15 random games (Table 1).

We illustrate the application of the ID method using the game Cleveland Cavaliers (CLE) and the Golden State Warriors (GSW) from the 25th of December 2015. These teams made it to the final in that season. Fig.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).

Table 1: 15 randomly selected games from season 2015-16

Away	Home	Date(MM.DD.YYYY)	Result
GSW	LAL	01.05.2016	109-88
MIL	CHI	01.05.2017	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



(a) Video frame at the moment of the shooting.

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

Figure 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 25th of December, 2015. `https://youtu.be/jb57MFQLoRo?t=17`

3. Results

3.1. Two teams approach

We compute the intrinsic dimension using the shot chart data from the home and away teams. We split the data into two sets as follows: (1) field goals shots taken when the home team (e.g. GSW) is attacking and the away team (CLE) is on defense; and (2) field goals shots from the away team (CLE) on attack and the home team on defense (GSW). The number of rows of each dataset is the number of attempted field shots. The number of columns represent the original dimension of the data i.e. D is 20 (2 players' coordinates (x and y) $\times$ 5 players $\times$ 2 teams). The intrinsic dimension for the set of players (5 vs 5) corresponds to the number of independent directions in which the 20-dimensional points are embedded.

In Fig.2 we show a heatmap of the posterior similarity 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 similar probabilities of belonging to a certain cluster tend to be grouped. The labels in the x axis represent the game event or play identification number (idn). Four main clusters are identified (in yellow color representing high probability), the first three containing most of the plays. For example, $idn = 15$ is the play shown in Fig.1. The right hand side dot plot shows the outcome of each of the field goals. Missed shots (0) are in orange color and shot that were made (1) are in green. We note a large number of unsuccessful plays in cluster 1 (id = 383–19).

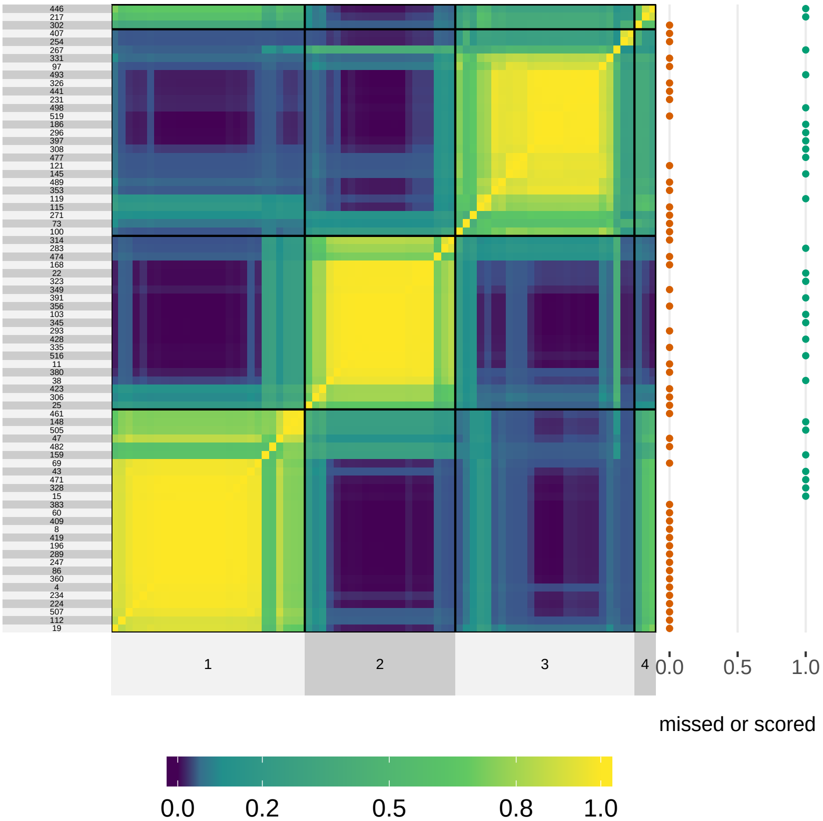


Figure 2: Heatmap of the posterior similarity matrix for the plays where CLE was on attack and GSW on defense.

Equally from the same game we have the field goal plays by GSW, which are represented in Fig. 3. Cluster 1 in the left hand side lower section represents a group of plays that where GSW had probability of success (33%), which is well below the others.

Table 2 contains the probability of success per shot type (short, mid-range and 3 points) and the mean ID value for both teams.

We computed the ID values for the shots taken during 15 games of the season. Although the number of games is relatively small, there seems to be a positive association between the posterior mean of the intrinsic dimension and the game outcome. The boxplots in Fig 4 show the ID for winning and losing teams. Each game is represented by a gray line connecting both teams from the left to the right boxplot. The solid line gives the games where a

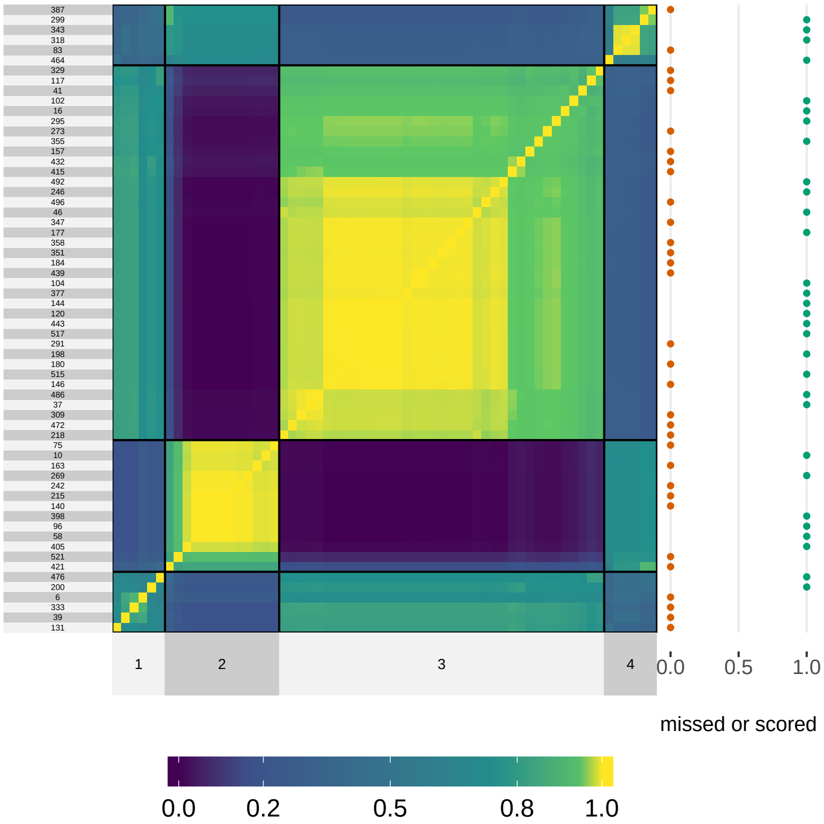


Figure 3: Heatmap of the posterior similarity matrix for the plays where GSW was on attack while CLE was on defense. The right dots plot shows the field goals made (green dots) and missed (orange dots).

significant difference was found between the posterior means using a Mann-Whitney test. The dashed line represents no evidence showing differences between the teams. In seven of these games, the winner had a greater intrinsic dimension. In five, there was no difference in the ID between winners and losers and in three cases the losers had higher ID values.

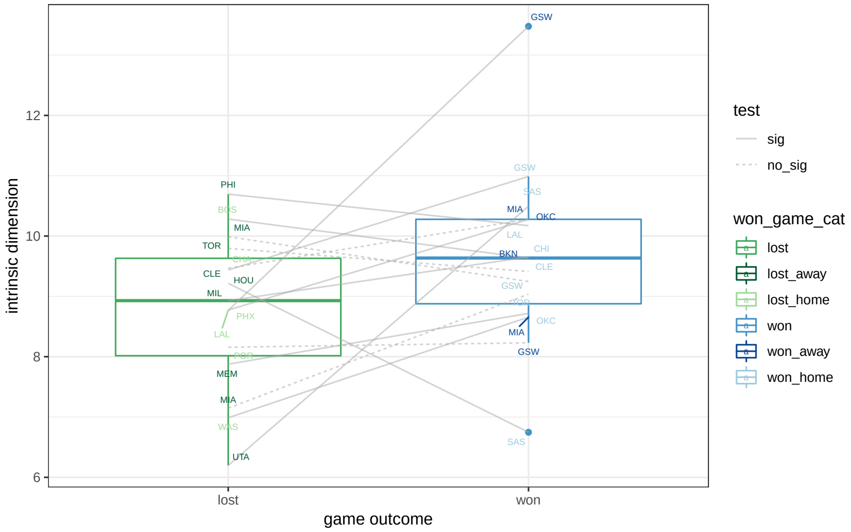


Figure 4: Boxplot of the posterior intrinsic dimension of the winners and losers in the 15 matches. Each game is represented by a gray solid line connecting both teams.

Table 2: Probability of success when each team is on attack per type of shot (short, mid-range and 3 points). In the season, CLE had a probability of scoring of 0.362 in 3 points and 0.514 in 2 points shots. GSW's probability of scoring in 3 and 2 points shots was 0.416 and 0.528 respectively.

team	dist_cat	ns	p success	$\overline{ID}$
CLE	short	40	0.375	9.596
CLE	mid_range	16	0.500	10.182
CLE	3_points	20	0.250	10.052
GSW	short	25	0.640	11.317
GSW	mid_range	22	0.409	11.056
GSW	3_points	15	0.333	11.049

We argue that plays tend to have an increased movement complexity when the difference in the score is small, usually as a result of a tighter defense. In Fig 5 we show the distributions of the ID's posterior means for different score margins categories {small(0-5 points), medium(6-10 points), large(11-15 points) and huge(>=16 points)}.

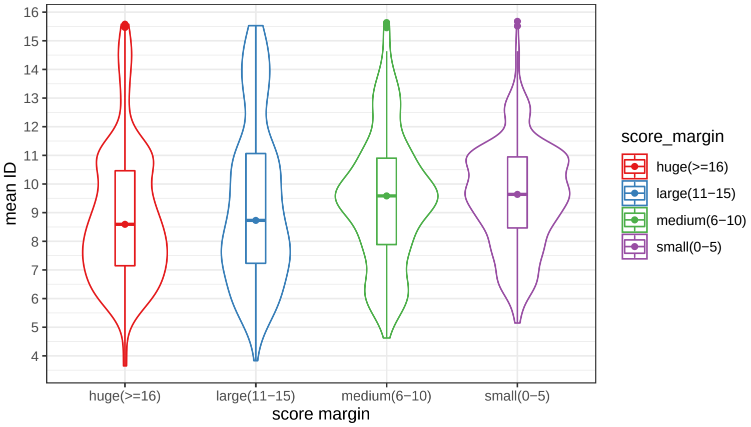


Figure 5: Violin/boxplot of the posterior intrinsic dimension as a function of the score margin in the 15 matches.

Pairwise comparisons using the Wilcoxon rank sum test shows evidence supporting the argument that the smaller score margin the greater the ID and the complexity.

Table 3: 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 mean ranks than the one in the columns.

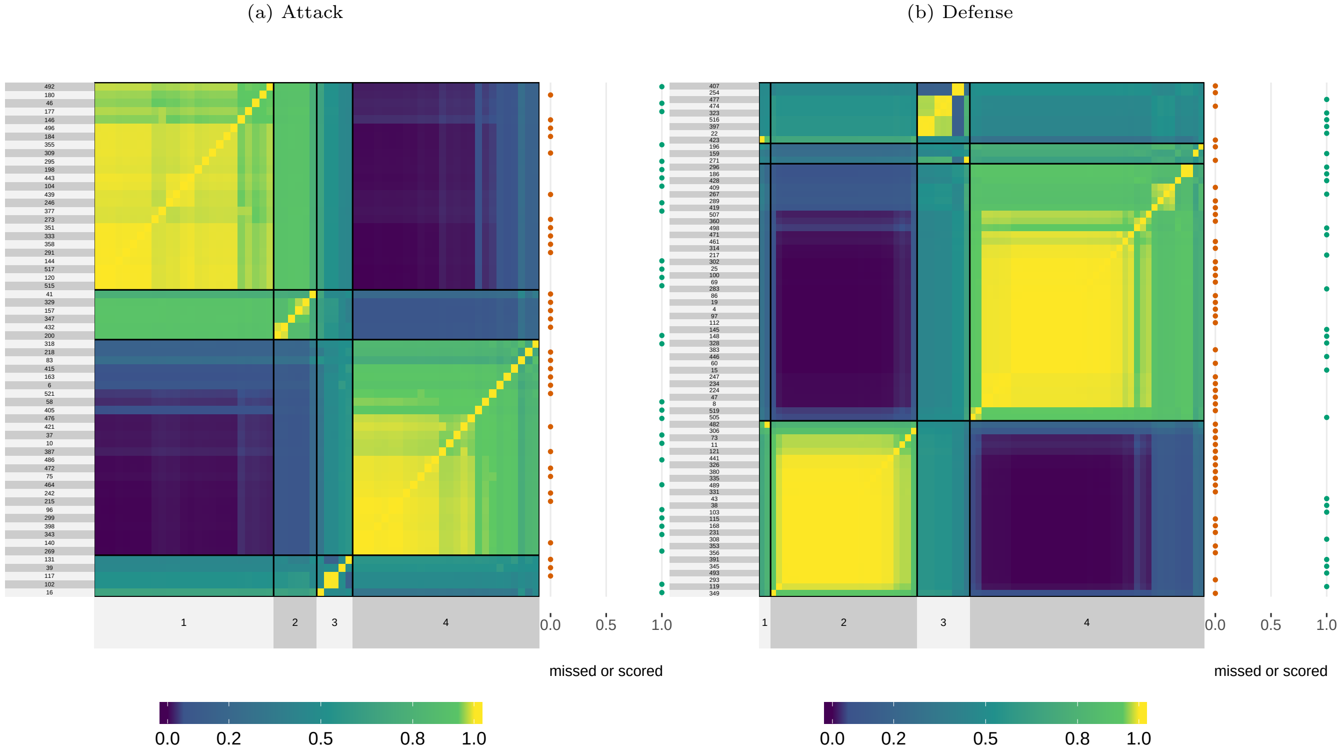
	huge(>=16)	large(11-15)	medium(6-10)
large(11-15)	0.285		
medium(6-10)	<0.0001	0.334	
small(0-5)	<0.0001	0.010	0.505

3.2. Individual Team Approach [Need to update this section]

We ran a similar analysis for individual datasets comprising the locations of the 5 players from each team in attack and then when they are in a defensive role. In this case the dimension is $D = 10$ (location in x and $y \times 5$ players). This analysis yields for each team clusters of shot charts plays with a low and a high return in offense and defense.

Fig 6 shows the posterior similarity heatmaps of the plays by GSW. On each of the plot, we defined four clusters. For instance, in subfigure (a) finding cluster 1 in the x-axis, we find that plays 492, 180, $\cdots$ 120 and 515 have a large probability of belonging to this cluster (yellow color). The outcome of each play is represented in the dot plot on the right-hand side. Table 5 gives the proportion of successful plays in the clusters. 56% of these offensive shots in cluster 1 (a) were successful. Similarly, only 16.7% 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 a poor defensive outcome for GSW, allowing 55% scoring by CLE.


(a) Attack

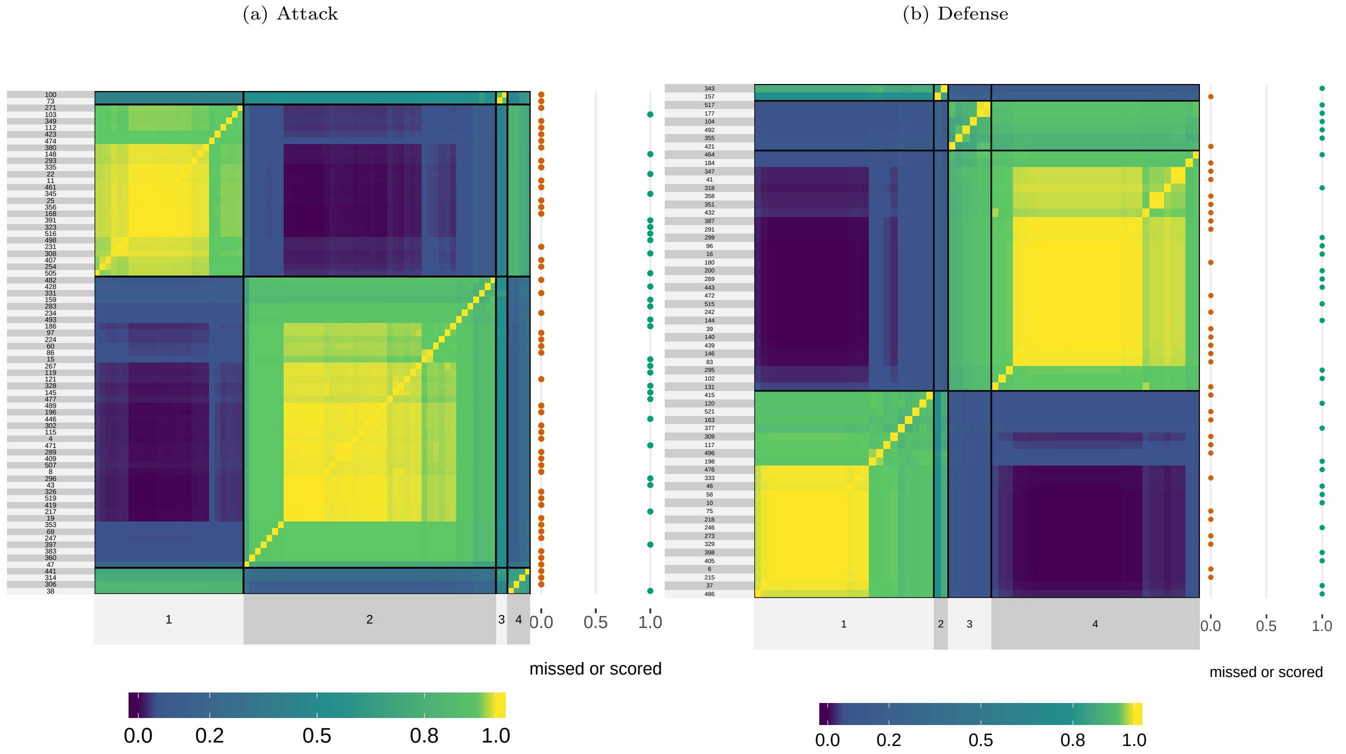
(b) Defense

Figure 6: Heatmap and clusters of the shot chart plays by GSW in attack and defense. 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 GSW.

role	cluster	p success
attack	1	0.560
attack	2	0.167
attack	3	0.400
attack	4	0.500
defense	1	0.000
defense	2	0.320
defense	3	0.556
defense	4	0.375

Furthermore in Fig 7 we present the posterior similarity of CLE in attack (a) and defense (b). From (b) cluster 3 contains six plays where the defense by CLE was ineffective allowing 83% success for GSW.


(a) Attack

(b) Defense

Figure 7: Heatmap and clusters of the shot chart plays by CLE in attack and defense. 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 5: Probability of success in the offensive and defensive roles for CLE.

role	cluster	p success
attack	1	0.385
attack	2	0.386
attack	3	0.000
attack	4	0.250
defense	1	0.480
defense	2	0.500
defense	3	0.833
defense	4	0.414

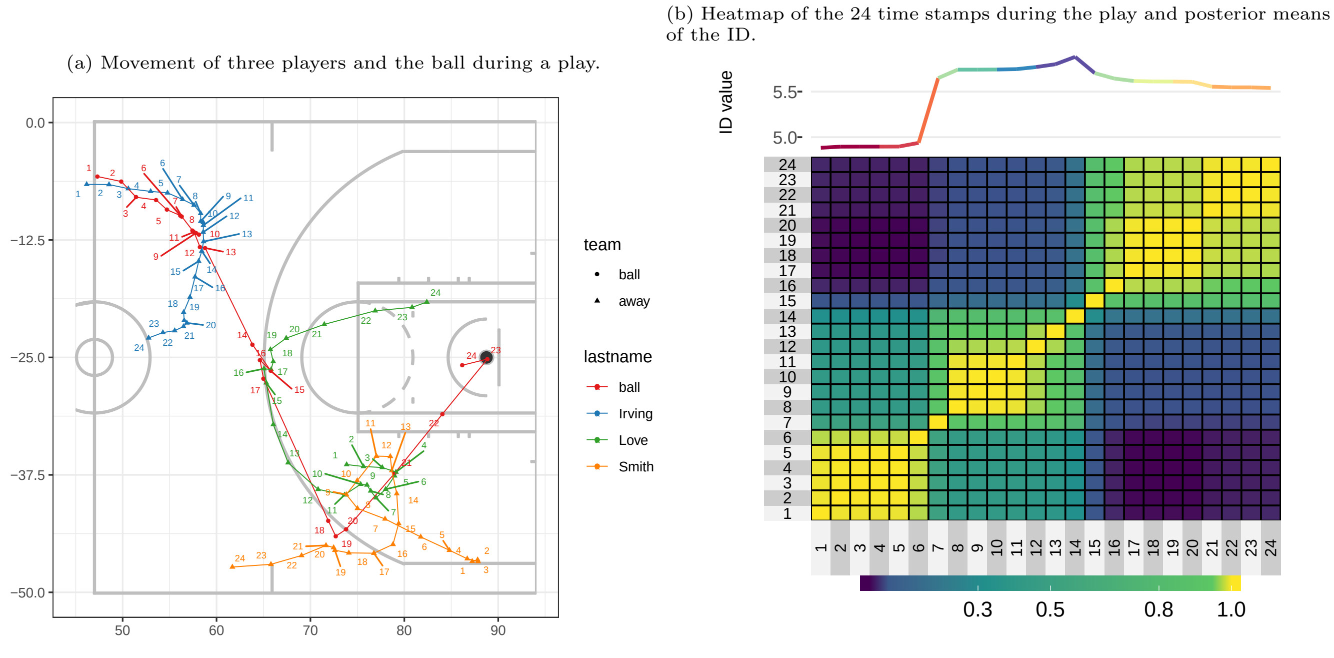
3.3. ID Analysis of Movement Data

In this section, we assess the change in ID along plays using players' movement data in the offensive court (after the ball passes the 47-foot central line). The resolution of each play is reduced from 25 frames/second to 2.5 for faster computation without losing much information. In this way, frame 1 corresponds to the first timestamp during the play. The number of frames in a play is based on the duration.

Fig8a shows the trajectory of the 3 players in handling the ball in the play illustrated in Fig.1. Fig 8b gives the heatmap of the posterior similarity matrix obtained from the Bayesian ID estimation algorithm. The line plot on top represents the evolution of the mean ID across the 24 frames of play.

In this play, Irving crosses the center line dribbling and in frame 7 the players start creating space for passing. Note the spike in the ID value. In frame 13 he passes the ball to Love. The ball goes from Love to Smith that executed a three-pointer in frame 19.

As expected consecutive frames are highly clustered since players tend to preserve the momentum in short intervals of times, but some interesting changes can be observed as in frame 7 and 14. For example, we identify the stages: ball handler after crossing the center line (frames: 1-6), creating space for passing (frames: 7-11), preparation/shooting (frames: 17-20) and following through (frames: 21-24). Another example of a driving bank 2 points shot is presented in Appendix B.


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

(b) Heatmap of the 24 time stamps during the play and posterior means of the ID.

Figure 8: The trajectory of the players/ball, heatmap of the coclustering matrix and ID for the first scored three-point field goal of the game by CLE. The rate is set at 2.5 frames/second and the play is composed of 24 frames. `https://youtu.be/jb57MFQLoRo?t=17`

Possessions, where players move in the same direction, will have a smaller mean ID value, while multi-directional trajectories tend to produce higher values. We illustrate this principle in the following example. We consider three simple (idn = 23, 25, 482) and three complex plays (idn = 446, 477, 145). The mean ID value across these plays is given in Fig. 9. Simple plays are in general shorter where the players reach easily the painted zone without much resistance. Fig. 10a and 10b show an example of the offensive trajectories in plays 25 and 145.

We performed further analysis based on the distance (δ) of from the player taking the shot to the hub. Three shot groups were defined:

- short distance (dunks, tips, etc). Where $\delta < 6$ feet.
- mid-range (short and long two points shots). Where $6 \leqslant \delta < 22$ feet.
- 3 points (shots from behind the line). Where $\delta \geqslant 22$ feet.

Additionally, possessions were divided into two groups: short and long duration. We used the cut-off of 12.5 seconds measured from the moment the ball crosses the center line.

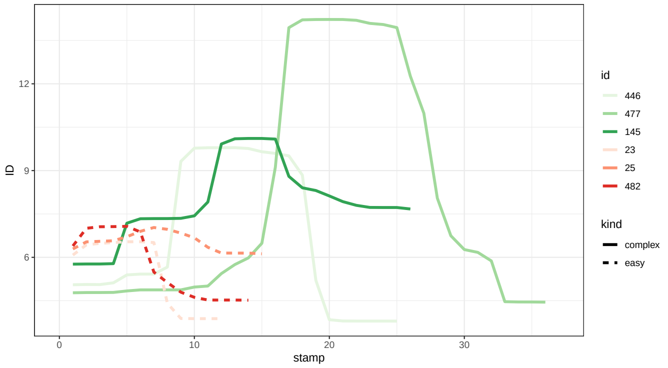
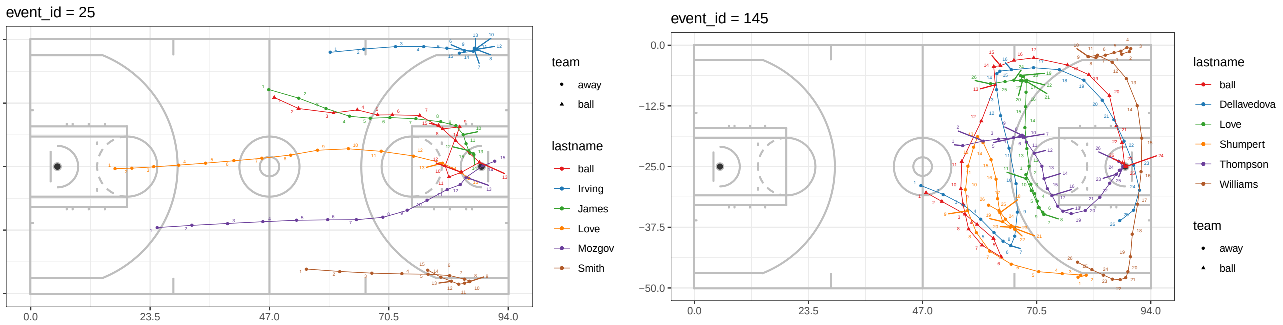


Figure 9: Example of the ID in complex and simple plays.


(a) Simple play.

(b) Complex play.

Figure 10: Example of players trajectories in simple and complex possessions.

Fig. 11 gives the posterior mean of the ID value for the game CLE vs GSW. The x-axis represents the frame number (2.5frames/second). Overall, the ID show patterns of spikes and declines during the execution of the play. Short possessions tend to have a peak in ID around frame 10 (≈ 4 seconds after the ball reaches the offensive court.) However, for long possession times, the ID reaches the pinnacle at approximately twice the time (8 seconds or 20th frame) mainly for short and mid-range shots.

4. Discussion and conclusions

The advent of sports tracking technology is flooding basketball analytics and sports science with large datasets that are increasingly challenging for individual analysis and for making meaningful inferences of players' performance. As a result, researchers and practitioners are resorting to multivariate statistical analysis so that high dimensional data can be reduced, handled and interpreted more conveniently.

The purpose of the current study was to present a different perspective in the analysis of high-resolution player tracking data from the NBA.

We used the intrinsic dimension of the player's positions (x, y) in Cartesian coordinates to:

- identify clusters in shot chart data.
- compare and assess the relationship between intrinsic dimension and game performance.
- determine different stages in the execution of offensive actions.

We opted for the ID approach developed by Allegra et al. (2019) because it has been found to be fast and accurate. The results show that using this Bayesian nonparametric clustering approach we can satisfactorily identify plays with a lower and higher than average return. Our method could help coaches to plan more effective attacking and defensive plays.

Higher games' mean ID values were found to be linked to a higher play uncertainty in the attack. This results are in line with previous findings, see e.g. Hobbs et al. (2018) that discusses ball entropy. We also found that a

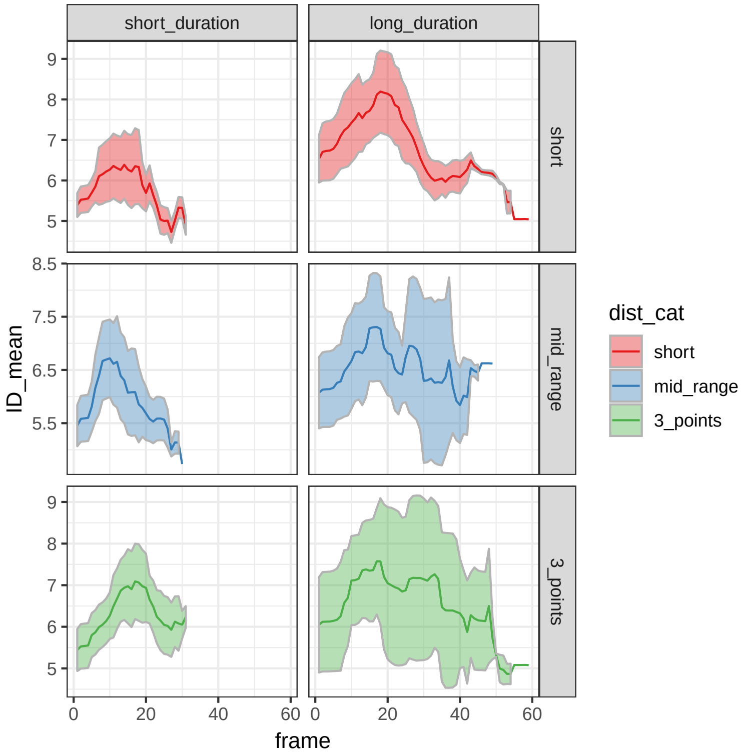


Figure 11: Posterior means and 95% confidence interval of the ID during the shots taken by both teams. We consider short, mid-range and 3 points attempts as well as short and long duration possessions ($t \leqslant 12.5$ and $t > 12.5$ seconds respectively).

larger ID in shot charts is positively associated with winning games. Although this claim needs to be validated using a larger sample size.

This approach also enhances our understanding of how players' moving tactics impact the outcome of a play. 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 mean ID curve. An increase in ID values was found when the players are creating an opportunity for passing and shooting. This is expected as players on attack tend to move with larger uncertainty and entropy using for example screen actions. Similarly, plays show a decline in ID near the end when the players are following through shots or returning to the opposite part of the court.

This Bayesian non-parametric approach could complement manual video game analysis, providing effective and fast clustering. In addition, these analyzes can be easily extended to other sports like football and rugby that have implemented players tracking technology. Further research should be undertaken to assess the link between intrinsic dimension and issues like players energy consumption & fatigue, movement dynamics.

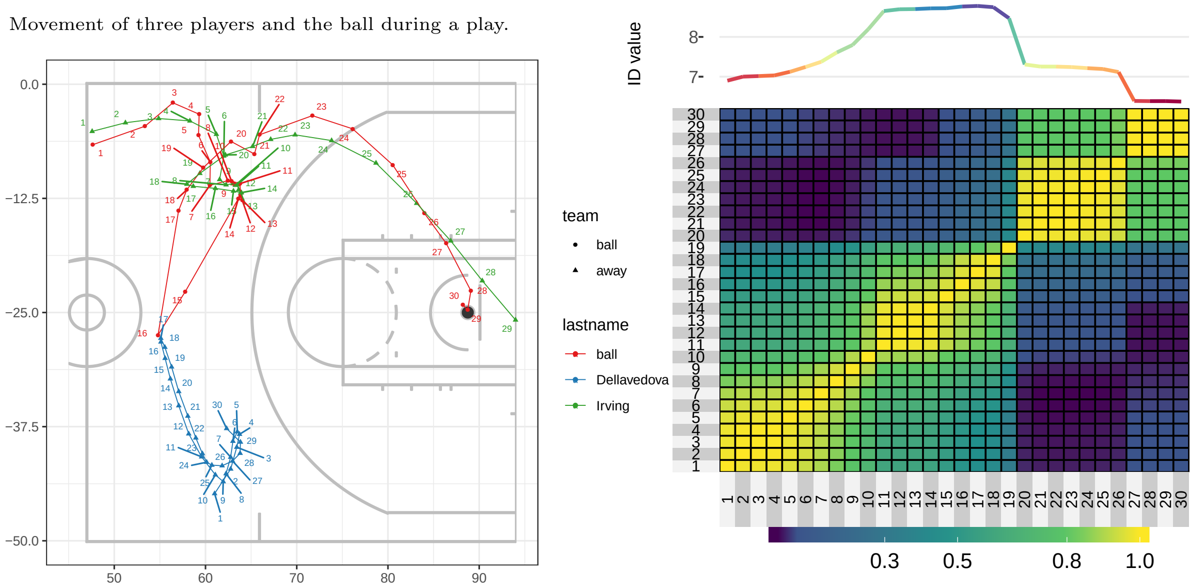
Acknowledgement

This research was supported by the Australian Research Council (ARC) Laureate Fellowship Program, the Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and by the project "Bayesian Learning for Decision Making in the Big Data Era" (ID: FL150100150). First Investigator: Prof Kerrie Mengersen. We thank Wade Hobbs for his comments and suggestions. All computations and visualizations were carried using R using the packages mcclust (Fritsch, 2012), superheat (Barter & Yu, 2017) and tidyverse (Wickham, 2017).

Note

Codes of the intrinsic dimension computation were written by F Denti.

Appendix A. Example of the ID in event 217 of the game (Irving's driving bank shot. Score = 39 vs 37).



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

(b) Heatmap of the 30 time stamps during the play and posterior means of the ID.

Figure A.12: Trajectory of the players/ball and ID in a two-point driving bank shot. `https://youtu.be/jb57MFQLoRo?t=180`

References

Allegra, M., Facco, E., Laio, A., & Mira, A. (2019). Clustering by the local intrinsic dimension: the hidden structure of real-world data. *arXiv preprint arXiv:1902.10459*, .

Barter, R., & Yu, B. (2017). *superheat: A Graphical Tool for Exploring Complex Datasets Using Heatmaps*. URL: `https://CRAN.R-project.org/package=superheat` r package version 0.1.0.

Camastra, F., & Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, *328*, 26–41.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, *111*, 585–599.

D'Amour, A., Cervone, D., Bornn, L., & Goldsberry, K. (2015). Move or die: How ball movement creates open shots in the NBA. MIT Sloan Sports Analytics Conference.

Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, *7*, 12140.

Franks, A., Miller, A., Bornn, L., Goldsberry, K. et al. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, *9*, 94–121.

Fritsch, A. (2012). *mcclust: Process an MCMC Sample of Clusterings*. URL: `https://CRAN.R-project.org/package=mcclust` r package version 1.0.

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., & Freeston, J. (2018). Playing unpredictably: measuring the entropy of ball trajectories in international women's basketball. *International Journal of Performance Analysis in Sport*, *18*, 115–126.

Levina, E., & Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems* (pp. 777–784).

Lucey, P., Bialkowski, A., Carr, P., Foote, E., & 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 *In MITSloan Sports Analytics Conference*.

Metulini, R. (2018). Players movements and team shooting performance: a data mining approach for basketball. *arXiv preprint arXiv:1805.02501*, .

Metulini, R., Manisera, M., & Zuccolotto, P. (2017). Space-time analysis of movements in basketball using sensor data. *arXiv preprint arXiv:1707.00883*, .

Nistala, A., & Guttag, J. (2019). Using Deep Learning to Understand Patterns of Player Movement in the NBA. In *In Proceedings of the MIT Sloan Sports Analytics Conference* (pp. 1–14).

Sampaio, J., Drinkwater, E. J., & Leite, N. M. (2010). Effects of season period, team quality, and playing time on basketball players' game-related statistics. *European Journal of Sport Science*, *10*, 141–149.

Sampaio, J., McGarry, T., Calleja-González, J., Sáiz, S. J., i del Alcázar, X. S., & Balciunas, M. (2015). Exploring game performance in the national basketball association using player tracking data. *PLoS One*, *10*, e0132894.

Shortridge, A., Goldsberry, K., & Adams, M. (2014). Creating space to shoot: quantifying spatial relative field goal efficiency in basketball. *Journal of Quantitative Analysis in Sports*, *10*, 303–313. URL: `https://www.degruyter.com/view/j/jqas.2014.10.issue-3/jqas-2013-0094/jqas-2013-0094.xml`. doi:10.1515/jqas-2013-0094.

Skinner, B., & Goldman, M. (2015). Optimal strategy in basketball. *arXiv preprint arXiv:1512.05652*, .

Teramoto, M., Cross, C. L., Rieger, R. H., Maak, T. G., & Willick, S. E. (2018). Predictive validity of national basketball association draft combine on future performance. *The Journal of Strength & Conditioning Research*, *32*, 396–408.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. URL: `https://CRAN.R-project.org/package=tidyverse` r package version 1.2.1.