tHoops: A Multi-Aspect Analytical Framework Spatio-Temporal Basketball Data Using Tensor Decomposition

Evangelos Papalexakis
University of California, Riverside
Department of Computer Science

Konstantinos Pelechrinis
University of Pittsburgh
School of Computing and Information

Abstract

The shot selection process in basketball can be thought of as an indicator of the identity of a player or a team. Characterization of these processes is important for player and team comparisons, as well as, for pre-game scouting. Typically, shot charts are compared in a heuristic manner. Recently though automated ways have appeared in the sports analytics literature that aim into identifying a set of prototype shooting patterns that can be used as a basis for describing the tendencies of a player or a team. Nevertheless, these approaches are almost exclusively focused on the spatial distribution of the shots. However, there is a multitude of other parameters that can affect the shot selection by a player. For example, the time remaining on the game (or shot) clock, the current score differential, or the defensive player assignment are some contextual factors beyond location that can impact the shot selection of a player. In this work, we propose a framework based on tensor decomposition for obtaining a set of prototype shooting patterns based on (i) spatio-temporal information, as well as (ii) contextual meta-data, that can be used to describe the overall shot selection process for a player or a team. The core of our framework is a three-dimensional tensor $\underline{\mathbf{X}}$, whose element $\underline{\mathbf{X}}(i,j,k)$ is equal to the number of shots that *entity* (player or team) i took from *location* j during *time* k. The granularity of time and location can be defined differently depending on the application. For example, the spatial granularity can be either a fine-grained grid over the court or the shot zones of the court. Using the PARAFAC decomposition we can decompose the tensor into several interpretable patterns, that capture a group of players/teams, that take shots from similar locations during similar times throughout the game. To obtain the results presented in this work we have used as our temporal granularity the game period when the shot occurred (we have grouped all overtimes to a single, 5th period), while as our spatial granularity we have used the shot zones. Using the tensor components, we can express every player/team as a combination of these components with the weights being the corresponding elements in the player/team factor of the decomposition. The framework (which we name tHoops) introduced in this paper can have further applications in the analysis of the spatiotemporal data available from optical player tracking as we showcase by analyzing a small dataset of four games.

1. Introduction

What are the offensive tendencies of your upcoming opponent with regards to their shot selection? Do these tendencies change through the course of the game? Are they particularly ineffective with regards to specific tendencies so as to force them towards them? These are just some of the questions that our proposed framework tHoops, based on tensor decomposition, can answer. The offensive tendencies of a team can be captured through shot selection, player schemes on the court etc. Given the availability of data we use information from shot selection to develop and present tHoops. However, we elaborate on the ways that tHoops can be applied to other kind of data that capture offensive tendencies.

Given a set of shots, described through their location on the court, the player/team that took each shot and the time that each shot was taken[1], tHoops identifies prototype patterns for shooting tendencies. Prior literature exists that aims into identifying similar prototype patterns. For example, in the seminal work by Miller et al. [1] matrix decomposition of a spatial Poisson point process describing the locations of the shots was used to obtain low dimensional representations of the players' shot selection. Intuitively, this task resembles a task of topic discovery in documents, where the set of shots from each player is a document and the *topics* represent shooting patterns. These shot *topics* have been further used to identify the spatial structure of defensive skills [2], a rather ignored topic in the basketball literature. Similar approaches have been used in other sports. For instance, Wang et al. [3] developed a similar unsupervised approach for detecting tactical patterns in soccer, as captured by the spatial distribution of passes. One of the shortcomings of the approaches that exist in the literature today is that they largely consider only the spatial dimension of the underlying data. For example, in the basketball shot selection only the location on the court is utilized. However, as one can imagine, other contextual information can be relevant to this shot selection and offensive strategies. This might include game clock information, score differential, opponent etc. For presentation purposes of our tHoops framework we are going to consider the game clock information as our contextual information. However, we will also discuss how different contexts can reveal even more information.

The rest of the paper is organized as following: in the next section, we will present the basic tensor representation of the shot data, while we will also introduce the notion of tensor factorization/decomposition. Section 3 presents the data we collected and used in our analysis as well as our results, while we also discuss and showcase the applications of tHoops on other types of multi-aspect spatio-temporal basketball data (i.e., player tracking data). Finally, Section 4 concludes our work.

2. Tensor Representation and Decomposition

A n-mode tensor, is a generalization of a matrix (2-mode tensor) in n dimensions. For tHoops we will initially utilize a 3-mode tensor $\underline{\mathbf{X}}$, that will capture the spatiotemporal information of the shot selection for players/teams. The element $\underline{\mathbf{X}}$(i, j, k) will be equal to the number of shots that player/team i took from location j during time k. Figure 1 depicts this structure.

A typical technique for identifying latent patterns in data represented by a 2-mode tensor (i.e., a matrix), is matrix factorization (e.g., Singular Value Decomposition, Non-negative Matrix Factorization etc.). A generalization of the Singular Value Decomposition [4] in n-modes is the Canonical Polyadic (CP) or PARAFAC decomposition [5]. Without getting into the details of the decomposition, PARAFAC expresses $\underline{\mathbf{X}}$ as a sum of F components:

$$\underline{\boldsymbol{X}} \approx \sum_{f=1}^{F} \boldsymbol{a}_f \circ \boldsymbol{b}_f \circ \boldsymbol{c}_f \quad (1)$$

[1] As it will be evident from the description of our method, there is no limit on what information one can incorporate for the shot (e.g., score differential when the shot was taken, opponent etc.).

where $\boldsymbol{a}_f \circ \boldsymbol{b}_f \circ \boldsymbol{c}_f(i, j, k) = \boldsymbol{a}_f(i)\boldsymbol{b}_f(j)\boldsymbol{c}_f(k)$, and they are obtained as the solution of the following optimization problem:

$$\min_{A,B,C} D_{KL}(\underline{\boldsymbol{X}}| \sum_f \boldsymbol{a}_f \circ \boldsymbol{b}_f \circ \boldsymbol{c}_f) \quad (2)$$

where D_{KL} represents the Kullback-Leibler divergence [9]. Simply put, each component, i.e., triplet of vectors, of the decomposition is a rank-one tensor. Each vector in the triplet corresponds to one of the dimensions of the original tensor $\underline{\mathbf{X}}$. In the example given at Figure 1, **a** corresponds to the players, **b** corresponds to the location on the court and **c** corresponds to the time (period). Each of these components can be thought of as a cluster, and the corresponding vector elements as soft clustering coefficients (i.e., if the coefficient is small, the corresponding element does not belong to this cluster). In our application, these clusters correspond to a set of players that tend to take shots from *similar* areas on the court during *similar* times within the game. The vectors (**b**, **c**) essentially correspond to the latent patterns for the spatio-temporal shot selection of players obtained from tensor $\underline{\mathbf{X}}$.

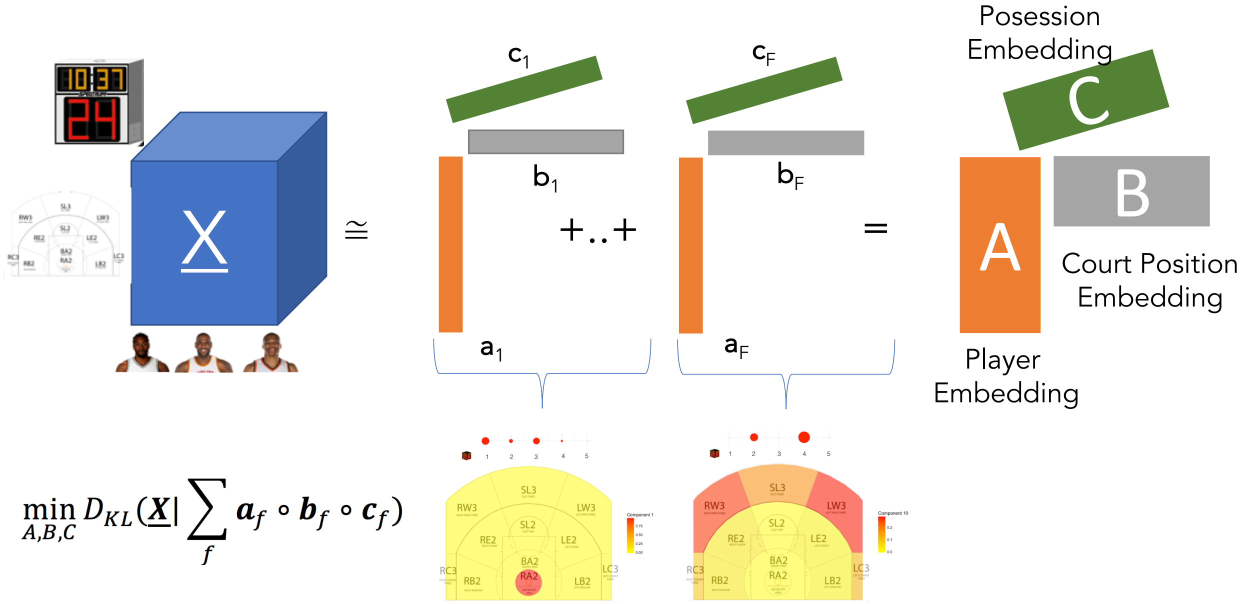

Figure 1 Representing offensive tendencies through tensors allows tHoops to incorporate multiple aspects of the data.

Intuition behind the use of tensors: Tensor decomposition attempts to summarize the given data into a reduced rank representation. PARAFAC tends to favor dense groups that associate all the aspects involved in the data (player, locations and time in the example in Figure 1). These groups need not be immediately visible via inspection of the n-mode tensor, since PARAFAC is not affected by permutations of the mode indices. As an immediate consequence, we expect near-bipartite cores of players who take shots from specific locations on the court during certain periods of the game. The benefit of tensor decomposition over matrix decomposition that has been used until now to analyzing shooting patterns, is the ability to consider several aspects of the data simultaneously. This allows tHoops to obtain a richer set of latent patterns that describe the original data.

One of the challenges associated with tensor decomposition is assessing the quality of the model, which also involves choosing the number of components F. In particular, depending on the structure of the given data, the PARAFAC decomposition can range from (almost) perfectly capturing the data, to performing rather poorly. In order to make sure that our model and choice of number of components we have utilized an elegant diagnostic tool, namely CORCONDIA [6], that serves as an indicator of whether that the model describes the data well or whether there is a problem with the model. Describing the technical details of CORCONDIA is beyond the scope of this work.

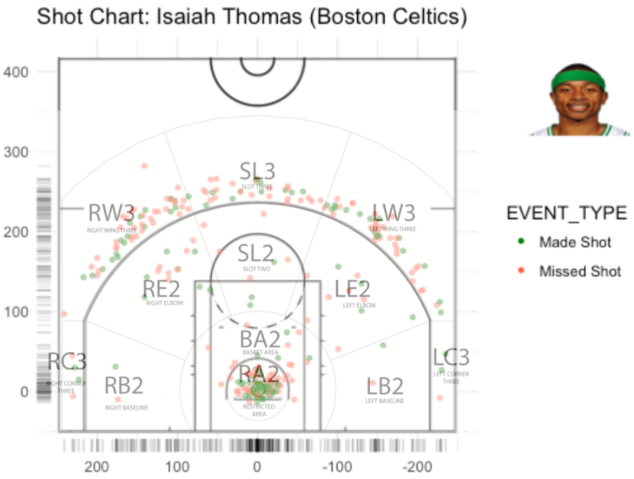

Figure 2 An example shot chart. The zones depicted correspond to the spatial granularity used in tensor $\underline{\mathbf{X}}$.

3. Data and Analysis

In order to examine the applicability of tHoops, we collected a shot dataset from the 2014-15 NBA season through NBA's shotchart API endpoint[2]. This endpoint provides several information for all the shots taken during the season, including: the game that the shot was taken, the player that took the shot, the location on the floor from where the shot was taken, the game clock information, the shot type, and whether the shot was made or missed. In total we collected information for 184,209 shots from 348 different players.

Using these data, we build the shot tensor $\underline{\mathbf{X}}$, where the location dimension corresponds to the 13 zones presented in the sample shotchart at Figure 2. For the temporal dimension, we use the game periods, where we merge all overtimes to a single 5th period. Therefore, the players tensor has a dimensionality of 348x13x5, while the teams tensor has dimensionality 30x13x5.

Player results: We start by presenting our results for the players' latent shooting patterns. One can build two separate tensors, one for made shots $\underline{\mathbf{X}}_{\text{P,Made}}$ and one for missed shots $\underline{\mathbf{X}}_{\text{P,Missed}}$, since one can argue that they encode different information of sorts. Figure 3 presents the spatial and temporal patterns for the 12 components we identified using the $\underline{\mathbf{X}}_{\text{P,Made}}$. Components 1 and 2 are particularly important for showing the difference between tHoops and a similar approach based on matrix factorization. The spatial element of these two components is very similar (almost identical) and represents shots made from the (deep) paint. However, they are different in the temporal dimension. As we can see component 1 includes shots taken mainly during quarters 1 and 3, while

[2] The dataset can be made available upon request.

component 2 mainly covers quarters 2 and 4. The fact that PARAFAC detected two components (instead of of one that covers all the periods), means that there are subgroups of players that take and make these shots in different times during the game. Of course, this difference can be purely based on personnel decisions from the coaching stuff through the game, but tHoops is able to pick

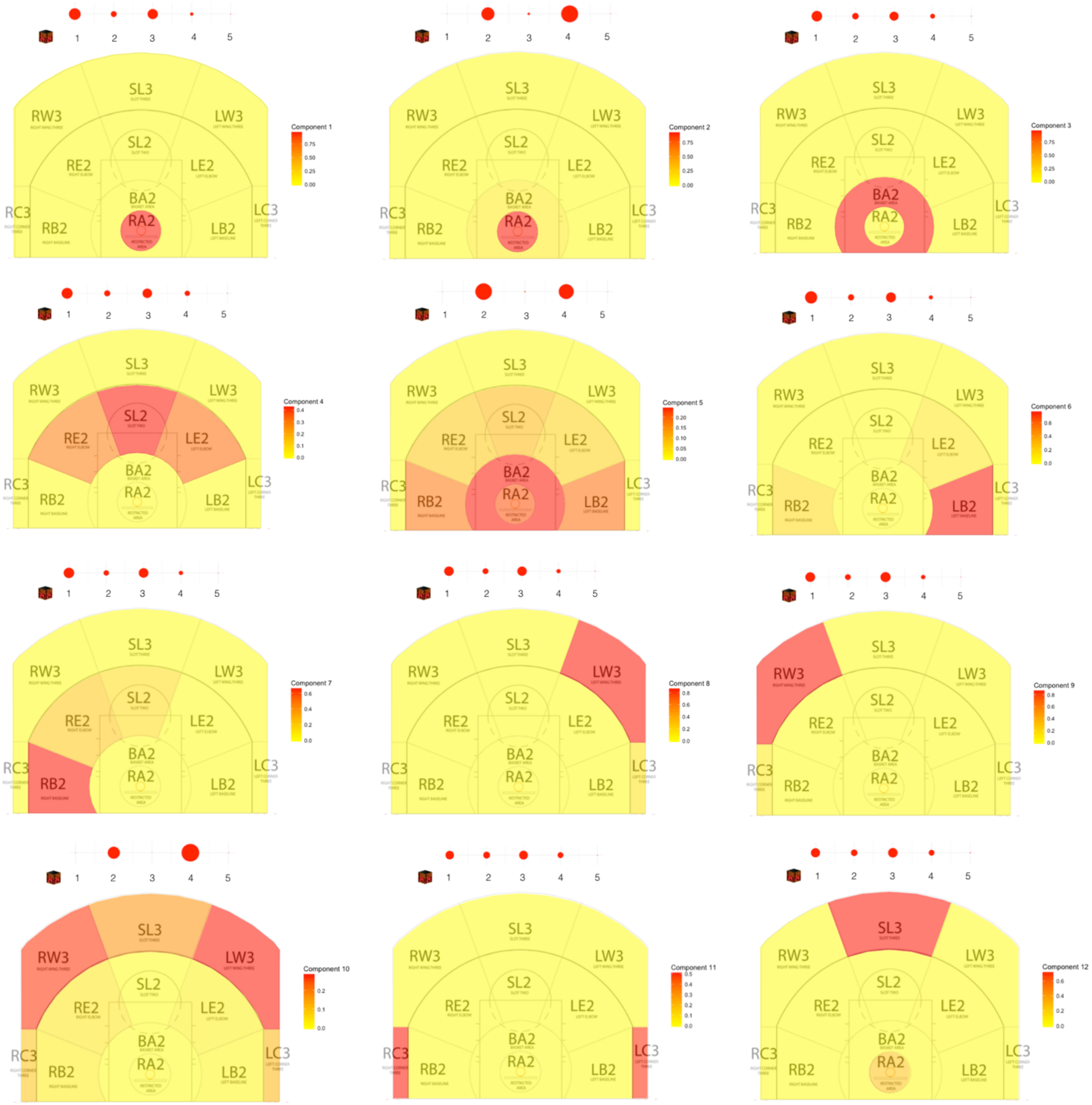

Figure 3 tHoops components for the $\underline{\mathbf{X}}_{P,Made}$ tensor. The spatial and temporal elements are presented. The size of the points for the temporal element correspond to the coefficient for each quarter.

this up and provide us with latent patterns considering all the aspects included in the tensor simultaneously. In contrast, component 11 corresponds to corner 3 (made) shots. There is no other component that includes them (component 10 includes a small fraction of corner 3 shots but it heavily captures above the break 3-point shots), which means that players that are heavily represented in this component take (and make) these shots almost uniformly across the game as it can be seen by the temporal element.

Figure 3 does not provide any information for the player vector of the component. The player vector of the tensor factor informs us which players have a *strong* representation in the component under examination. For example, Table 1 presents the top-10 players (i.e., the players with the largest coefficients in the corresponding player vectors) included in the corner 3 components and the midrange shot component.

Corner 3 (component 11)	**Midrange (component 4)**
Trevor Ariza	Blake Griffin
Matt Barnes	Avery Bradley
Danny Green	Monta Ellis
Harrison Barnes	David West
Klay Thompson	LaMarcus Aldridge
Luol Deng	Anthony David
Kyle Korver	Marc Gasol
JJ Redick	Pau Gasol
O.J. Mayo	Nikola Vucevic
Bojan Bojdanovic	Chris Paul

Table 1 The components obtained from tHoops can provide us with valuable information for the shooting tendencies of players.

As one might have expected Danny Green, Klay Thompson, Kyle Korver and JJ Redick are predominantly featured in the corner 3s component, while players like LaMarcus Aldridge, Chris Paul and the Gasol brothers are featured in the midrange component. Table 1 also serves as an indicator that the components obtained from tHoops are *sensible* and can be *trusted.* Using the coefficients of the player vectors of the tensor components, we can obtain a 12-dimensional latent representation of each player that can be further used to cluster players. These clusters will represent players with similar offensive patterns (with regards to shots made). We use k-means clustering and the gap statistic [7] to determine the appropriate number of clusters, which provides us with a value of k=5. Figure 4 further presents the clusters on a two-dimensional projection using t-SNE [8]. As we can see the clusters are well distinguished – especially considering that t-SNE uses a reduced dimensionality of the data. The largest cluster corresponds to players whose top patterns (i.e., the ones with the highest coefficients) correspond to shots taken from the paint (specifically tensor components 1, 2, and 3). The smallest cluster corresponds to players whose patterns heavily include the 3-point shoot components (tensor components 8, 9, 11 and 12). This cluster includes players such as Steph Curry, James Harden, Kyle Korver, JJ Redick, Gordon Hayward, Kyrie Irving, Klay Thompson and JR Smith. Another distinct cluster includes players whose most dominant components are 4, 6 and 7, i.e., midrange shots. This cluster includes players like DeMar DeRozan, LaMarcus Aldridge, Al Horford, Blake Griffin, Marreese Speights and Anthony Davis. The fourth cluster does not exhibit any specific pattern with regards to the spatial distribution of the shots. However, it includes players who are *offensively active* mainly during quarters 2 and 4 (tensor components 2, 5 and 10). Players that fall into this cluster are mainly bench and role players such as Jamal Crawford, Leandro Barbosa, Patty Mills, Andre Iguodala, J.J. Barea and Vince Carter. This shows that using the information from the tensor components allows us to essentially group players based on different aspects of their game simultaneously. Finally, the last cluster includes players that are a mix of the other 4 clusters, which makes it harder to *profile* them. Nevertheless, considering also the location of this cluster on the t-SNE projection, i.e., surrounded by the other four clusters, it further strengthens our belief that tHoops is able to capture multi-aspect patterns in the shooting data.

Team results: tHoops can also be used for analyzing the teams' offensive/shooting tendencies. In this case the tensor $\underline{\mathbf{X}}$ is obtained by using the shots from all the players of the teams. For the 2014-15 season tHoops identified 7 components, whose spatio-temporal parts are presented in Figure 5.

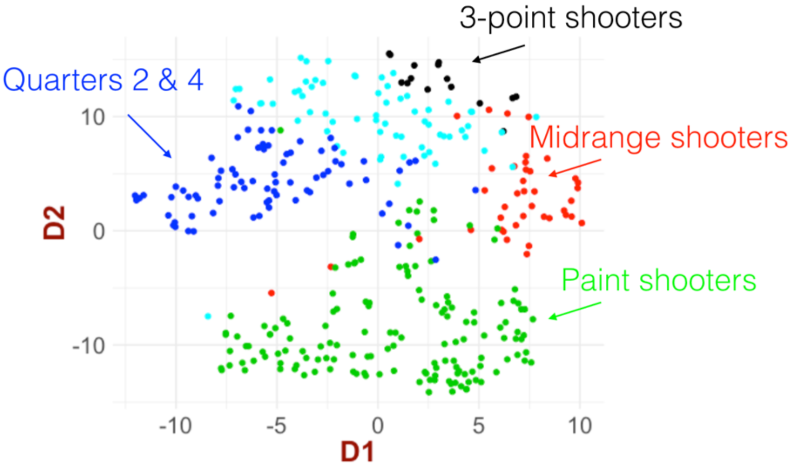

Figure 4 t-SNE visualization of the players' clusters using the components from tHoops.

These patterns can characterize the behavior of teams as a whole – rather than individual players. For example, for the Houston Rockets, their shot selection does not include, components 1, 3, 6 and 7(!!), i.e., the corresponding coefficients are almost 0. Rockets' game is thus only follows the latent patterns of components 2, 4 and 5. Note that these components correspond to three-point shots and shots taken from the paint. Something we should have expected from an analytically savvy team.

However, one might be wondering, what is the need for tHoops to understand that Rockets do not take many midrange shots at any point in the game, or that James Harden, Klay Thompson and Steph Curry are clustered as the 3-point shooters of the league? The analysis in this section should serve as an indicator that tHoops can capture pretty well what really happens. Therefore, when one incorporates information that has previously either ignored or could not be integrated in say a matrix factorization, new insights will be obtained. For example, the shot clock information can be an important factor for shot selection. When the shock clock winds down, a player will simply take a shot (in most of the cases) to avoid a shot clock violation. This shot can be of very bad quality, but the corresponding component obtained from the tensor factorization will inform us for this. The applications of tHoops are only limited by the amount and type of information available to us. In the following section, we elaborate on this and in particular on how it can be used to analyze player tracking data.

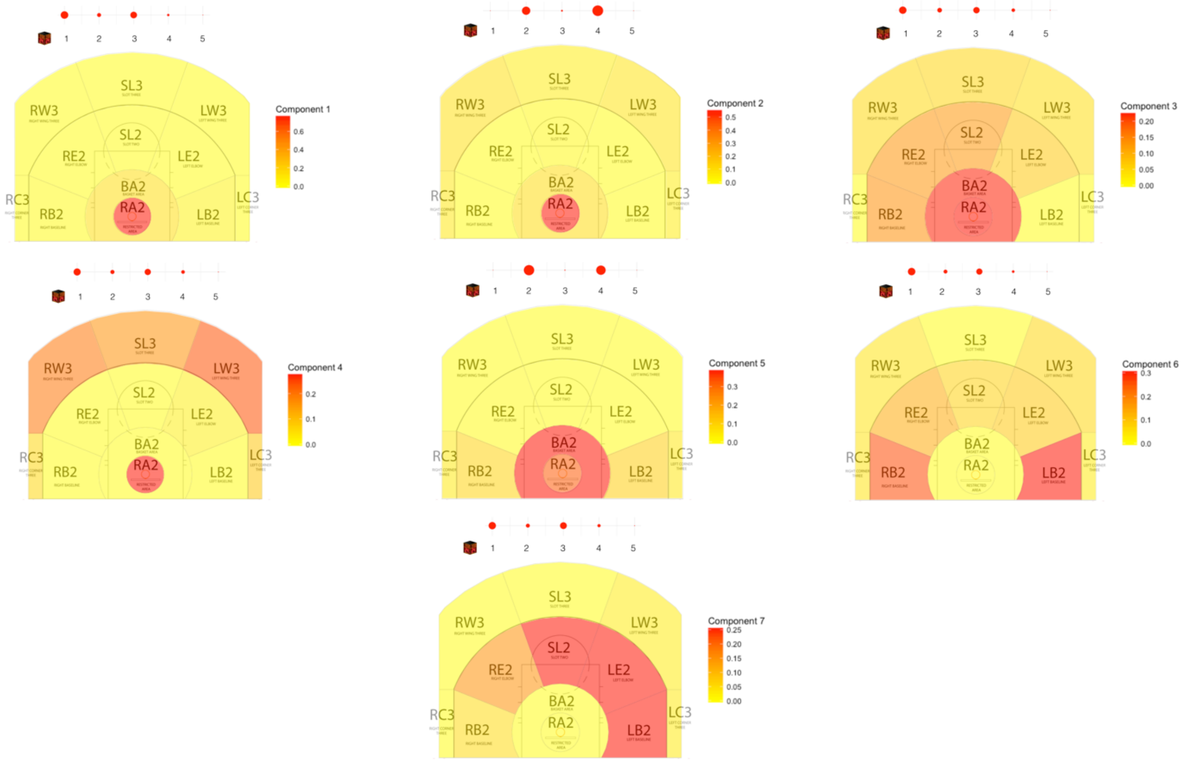

Figure 5 tHoops identified 7 components using $\underline{\mathbf{X}}_{\mathrm{T,Made}}$. The teams' offensive tendencies can then be thought of as a combination of these latent patterns.

3.1 Application of tHoops on tracking data

Offensive (and defensive) tendencies of a team in the NBA can certainly be better described today with player tracking data that capture the location of all the players and the ball 25 times every second. Player tracking data allows us to incorporate a wealth of information in the corresponding tensor including spatial information from all the players on the court. Consequently, one can identify prototype offensive tactics, i.e., sequential formation snapshots, either at the league level or at the team level depending on the setting.

In particular, we can design a three-mode tensor, where the modes represent (i) court zones (or any other spatial granularity of the court – e.g., a grid), (ii) shot clock, and (iii) a possession identifier. The element $\underline{\mathbf{X}}(i,j,k)$ of the tensor represents the number of offensive players on the court zone i, when the shot clock was j during possession k. Simply put, $\underline{\mathbf{X}}(i,j,k)$ can take values from 0 to 5. Further information can be also incorporated in the design of the tensor for additional insight. For instance, a fourth mode can be included that captures the score differential during the possession. This will allow us to identify prototype offensive patterns controlling for the score differential as well. Depending on the type of information it might be more appropriate to include them in a matrix coupled with the tensor. Consider personnel information per possession being available, such as team on offense/defense, player names, boxscore statistics of the players, (adjusted) plus-minus ratings, personal fouls etc. This information is better represented through a matrix $\mathbf{M}_{\mathrm{B}}$ with rows representing the possessions in the dataset and columns representing different attributes. In this case matrix $\mathbf{M}_{\mathrm{B}}$ is coupled with tensor $\underline{\mathbf{X}}$ at the possession dimension. Therefore, we can obtain the tensor components through a coupled matrix-tensor factorization, which essentially provides a low dimensional embedding of the data in a common contextual subspace, by solving the following optimization problem:

$$\min_{A,B,C,D} \left\| \underline{\boldsymbol{X}} - \sum_{f} \boldsymbol{a}_f \circ \boldsymbol{b}_f \circ \boldsymbol{c}_f \right\|_F^2 + \left\| \boldsymbol{M}_B - \boldsymbol{A}\boldsymbol{D}^{\boldsymbol{T}} \right\|_F^2 \tag{3}$$

where **A**, **B**, and **C** are matrices whose columns are vectors $\boldsymbol{a}_f$, $\boldsymbol{b}_f$, and $\boldsymbol{c}_f$, and **D** is a factor matrix for $\mathbf{M}_{B.}$ Given that tensor $\underline{\mathbf{X}}$ and matrix $\mathbf{M}_B$ are coupled in the possession dimension (factors $\boldsymbol{a}_f$), matrix **A** is common to the latent patterns for both the tensor and the matrix. The solution of optimization problem (3) will provide us with components that include prototype patterns for offensive strategies controlling also for the meta-information of the possessions. This for example can reveal the potential impact that the *abilities* of the players (as captured by individual statistics) have on the choice of offensive formations.

Example: We have obtained access to optical tracking data for four NBA games from the 2015-16 season. Clearly the sample size is small (714 non-transition possessions in total) to obtain deep insights but still can show the power of our method. Using the data from these games we build the three-mode tensor $\underline{\mathbf{X}}$ as described above. Figure 6 presents two representative patterns that we obtained.

Component 1 (top part of the figure) is observed when there are between 9-16 seconds left on the shot clock. This component showcases a good spread of the offense, with players outside the three-points line above the break, the left midrange elbow (potentially setting up a screen for the player outside the three-point line) and the right baseline. Component 7 (bottom part of the figure) includes mainly players outside the three-point line and in the basket area. These areas generate the most efficient shots and as we can see it is indeed observed when there are between 5-9 seconds left on the shot clock (i.e., towards the end of a possession). The interesting thing is that all the components identified from tHoops have a very well defined temporal component, i.e., most of the elements are zero. This is a very good property of tHoops, since it allows for synthesizing the various components to identify full offensive schemes.

Apart from the applications for on-court strategy and scouting, the components identified by tHoops can drive the development of a system that allows for flexible search in a database of possessions. This can automate and facilitate tasks related with film study. For example, one can imagine querying the system using as input a (probabilistic) spatial distribution for the offense, the shot clock and any other information available for the possessions used to build tensor $\underline{\mathbf{X}}$.

4. Conclusions

In this paper, we have introduced tHoops, a tensor decomposition-based framework for analyzing multi-aspect basketball data. tHoops is able to identify prototype patterns in the underlying data by simultaneously considering information from different *sources*. We have showcased the applicability and power of tHoops using a large scale shotchart dataset including 184,209 from the 2014-15 NBA season, as well as a dataset with optical tracking data from a small sample of 4 NBA games from the 2015-16 season. We believe that tHoops can automate a lot of the functionalities related with pre-game scouting and film studies.

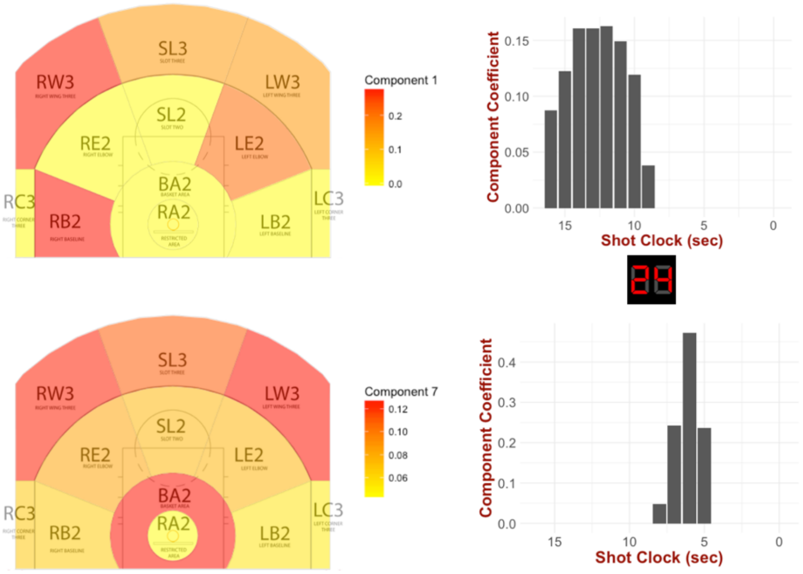

Figure 6 tHoops can also be used to process and analyze player optical tracking data.

5. References

[1] A. Miller, L. Bornn, R. Adams and K. Goldsberry, "Factorized Point Process Intensities: a Spatial Analysis of Professional Basketball", in International Conference on Machine Learning (ICML), 2014.

[2] A. Franks, A. Miller, L. Bornn and K. Goldsberry, "Characterizing the Spatial Structure of Defensive Skill in Professional Basketball", in The Annals of Applied Statistics, Vol. 9, No. 1, pp. 94-121, 2015.

[3] Q. Wang, H. Zhu, W. Hu, Z. Shen and Y. Yao, "Discerning Tactical Patterns for Professional Soccer Teams: An Enhanced Topic Model with Applications", in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[4] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman, "Indexing by latent semantic analysis", in JASIST, Vol. 41, No. 6, pp. 391-407, 1990.

[5] R. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an *explanatory* multimodal factor analysis", 1970.

[6] R. Bro and H.A. Kiers, "A new efficient method for determining the number of components in PARAFAC models", in Journal of Chemometrics, Vol. 17, No. 5, pp. 274-286, 2003.

[7] R. Tibshirani, G. Walther and T. Hastie, "Estimating the Number of Data Clusters via the Gap Statistic", in Journal of the Royal Statistical Society B, Vol. 63, pp. 411-423, 2001.

[8] L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE", in Journal of Machine Learning Research, Vol. 9, No. 11, pp. 2579-2605, 2008.

[9] Chi, Eric C., and Tamara G. Kolda. "On tensors, sparsity, and nonnegative factorizations." SIAM Journal on Matrix Analysis and Applications 33.4 (2012): 1272-1299.