13TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 1 - 2, 2019 | BOSTON, MA
presented by ESPN

From Markov models to Poisson point processes: Modeling movement in the NBA

Basketball Track

Jacob Mortensen
Simon Fraser University
jmortens@sfu.ca

Luke Bornn
Simon Fraser University, Sacramento Kings
lbornn@sfu.ca

1 Introduction

With the rise of optical tracking data, the ability to accurately model player movement has become a key competitive advantage in many sports. In the NBA, this tracking data is available at a rate of 25 measurements per second and provides (x, y) coordinates for all ten players on the court and (x, y, z) coordinates for the ball. Analysis of this data presents a substantial challenge, due to both the scale of the data, which can consist of more than 100 million rows for a given season, and to the sophisticated methods required to make sense of it. Current approaches generally fall into two categories: black-box methods and Markov models.

Black-box methods have proven useful in modeling defensive situations [1] and in creation of sequence models to forecast player movement and detect plays [2]. In these projects, researchers successfully implemented neural networks that recreate player movement to a high degree of complexity, but unfortunately these models lack clear interpretation. Markov models, on the other hand, simplify movement to treat the players as billiard balls so that their position at time t depends only on their position at time $t-1$, retaining a clear interpretation. This approach was used to great effect by Cervone et. al. [3], who relied on Markov transition probabilities to propogate action through a possession, thereby determining its expected value based on the current state. However, Cervone et. al. assume a simplistic model which estimates a players current position as a function of location, velocity, and acceleration, limiting the complexity of movement that can be accurately represented.

In this work, we combine elements of traditional Markov approaches with tools from spatial statistics to develop a flexible nonparametric method which allows for complicated patterns of movement and incorporates the presence of meaningful spatial features (such as the three-point line), while remaining completely interpretable.

1.1 Motivating Example

Let $S_t = (x_t, y_t)$ represent a player's location on the court at time t. The sequence of player locations during the game $S_1, \ldots, S_T$ can be treated as a Markov chain with conditional transition density $p(S_t \,|S_{t-1})$. Estimation of $p(S_t \,|S_{t-1})$ is frequently achieved through use of a dynamic linear model [4], which posits a relationship between sequential elements in the Markov chain of the form $S_t = \boldsymbol{\rho} f(S_{t-1}) + \epsilon$, where $\boldsymbol{\rho}$ is a vector of coefficients and $f(\cdot)$ is a function that can output quantities such as location, velocity, and acceleration. Often ϵ is assumed to follow a $N(0, \sigma^2)$ distribution, though other distributions can be used. An advantage of this representation is that it

offers a simple way to estimate $p(S_t \,|S_{t-1})$ even in regions where data is sparse or completely unavailable. However, as data becomes increasingly nonlinear or non-Gaussian, fitting such models can become prohibitively difficult. Nonparametric approaches have a similar functional form, but rather than making distributional assumptions about the error term, ϵ, it is assumed to be nonparametric and vary as a function of S_{t-1}. This allows ϵ to naturally incorporate both spatial variation and nonlinear relationships.

The simplest means of nonparametric estimation for Markov transitions is to use empirical maximum likelihood estimates for transition probabilities. These can be calculated by partitioning the basketball court C into K regions $A_1, \dots, A_K$ and letting the transition probabilities $p_{ij} = \mathbf{P}\big(S_t \in A_j \,\big|\, S_{t-1} \in A_i\big)$, i.e., the probability of transitioning to region A_j at time t given that a player was in region A_i at time $t-1$, for all $j, i = 1, \dots, K$. These probabilities are commonly stored in a transition probability matrix

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{K1} & \cdots & p_{KK} \end{bmatrix},$$

so that row i contains the probabilities of transitioning to all other states given that the chain is currently in state i.

The maximum likelihood estimate for p_{ij} is $\hat{p}_{ij} = N_{ij}/N_{i\cdot}$, where N_{ij} is the number of transitions from state A_i to A_j and $N_{i\cdot} = \sum_{j=1}^{K} N_{ij}$. Unfortunately, this only works with a very coarse partition because as the number of states approaches infinity, $\hat{p}_{ij} = 0$ for the majority of i and j even though the true underlying p_{ij} are non-zero. We desire a method that allows us to estimate not just coarse transition probabilities, but continuous transition densities. One way to deal with this is to make assumptions about the structure of the transition probabilities that generates a continuous density as the number of elements in the partition increases.

We propose a general framework for transition density estimation which borrows strength across the rows and columns of the transition matrix P by reparameterizing a Markov model as a Poisson point process in the combined input/output space. This defines a relationship between a discrete state correlation structure and a stochastic process that allows us to easily step between completely continuous or discrete states of any resolution. In Section 2 we detail the three relationships that make this possible: first, equivalence between the likelihood for a Markov model and the multinomial distribution; second, the multinomial-Poisson transformation; and third, the connection between the Poisson distribution and a Poisson point process. Once the mathematical underpinnings for our work have been established, we demonstrate how it might be applied to optical tracking data in Section 3. Finally, in Section 4 we discuss future work and provide concluding remarks.

2 Moving from a Markov chain to a Poisson point process

2.1 Multinomial representation of a Markov chain

Consider an ergodic first-order Markov chain that has values in some domain $D \subseteq \mathbb{R}^F$, $F > 0$, at times $t = 0, \dots, T$. As in Section 1.1, partition the domain into K states $A_1, \dots, A_K$. Given the initial observation X_0, the conditional likelihood for the Markov chain is

$$L(P) = \prod_{t=1}^{T} \mathbf{P}\left(X_t \in A_j \middle| X_{t-1} \in A_i\right) = \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{N_{ij}} \tag{1}$$

where $\mathbf{P}\left(X_t \in A_j \middle| X_{t-1} \in A_i\right) = p_{ij}$ and $N_{ij} = \sum_{t=1}^{T} \mathbf{I}[x_{t-1} \in A_i, x_t \in A_j]$. The conditional likelihood listed in (1) is proportional to the product of K independent multinomial likelihoods [5]. Therefore, we can estimate transition probabilities by assuming that transitions from state i are realizations of a multinomial$(N_{i\cdot}, p_{i1}, \dots, p_{iK})$ distribution for all i, where $N_{i\cdot} = \sum_{j=1}^{K} N_{ij}$.

2.2 Multinomial-Poisson transformation

Suppose $\mathbf{y} = (y_1, \dots, y_M)$ are independent Poisson random variables with means $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_M)$. The joint distribution of $\mathbf{y}$ factorizes into the product of a multinomial distribution and a Poisson distribution over $n = \sum_{j=1}^{M} y_j$, i.e.,

$$f(\mathbf{y}) = \prod_{i=1}^{M} f(y_i) = \text{Poisson}(n|\Lambda)\text{Multinomial}(\mathbf{y}|\boldsymbol{\alpha}, n).$$

Here $\Lambda = \sum_{j=1}^{M} \lambda_j$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$ where $\alpha_i = \lambda_i / \sum_{j=1}^{M} \lambda_j$. Birch [6] showed that under the constraint $n = \Lambda$ the maximum likelihood estimate for $\boldsymbol{\alpha}$ is equivalent whether we maximize over the multinomial density or the product of independent Poisson densities (see also [7], [8]). The connection to modelling Markov transitions follows directly; if we assume $N_{ij} \sim \text{Poisson}(\lambda_{ij})$ then we can estimate the underlying transition probabilities by letting $p_{ij} = \lambda_{ij} / \sum_{j=1}^{K} \lambda_{ij}$.

2.3 Poisson point process

Our review of Poisson point processes is necessarily brief, but the interested reader is directed to [9] or [10] for a more thorough exposition. A point process is a stochastic process over a domain D where a realization from the process is a finite set of points $S = \{\mathbf{s}_1, \dots, \mathbf{s}_n : \mathbf{s}_i \in D\}$. The distribution of S is governed by an intensity function $\Lambda(\mathbf{s}) : D \longrightarrow [0, \infty)$, and a point process is said to be Poisson if, for any subset $B \subseteq D$, the number of points falling within B (denoted $N(B)$) follows a Poisson(δ) distribution, where $\delta = \int_B \Lambda(\mathbf{s})\, d\mathbf{s}$. Estimation of the intensity function is critical because it influences both the number of points and where they fall in the domain.

The likelihood for the Poisson point process is

$$L(\Lambda(\mathbf{s}); \mathbf{s}_1, \dots, \mathbf{s}_n) = \prod_{i} \Lambda(\mathbf{s}_i) \exp\left(-\Lambda(D)\right), \tag{2}$$

where $\Lambda(\cdot)$ is the point process intensity and $\Lambda(D) = \int_D \Lambda(\mathbf{s}) d\mathbf{s}$. Suppose we partition D into a series of discrete regions $B_1, \dots, B_M$. Conditional on the intensity function, if two regions B_1 and B_2 are

disjoint, then $N(B_1)$ and $N(B_2)$ are independent Poisson random variables. Due to this property, it follows that the likelihood over this partition is:

$$\prod_m \exp\big(-\Lambda(B_m)\big)\big(\Lambda(B_m)\big)^{N(B_m)}/N(B_m)!.$$

As the partition grows increasingly fine, $N(B_m) = 1$ or 0, depending on whether or not there is an observation in B_m, and in the limit we reach (2) [10].

To illustrate the connection to Markov transition density estimation, let (x_{t-1}, x_t) be successive elements of our Markov chain for arbitrary t. To estimate the conditional transition density we would set $p(x_t|x_{t-1}) = \Lambda(x_{t-1}, x_t)/\int \Lambda((x_{t-1}, \xi))d\xi$ which, by properties of intensity functions, is a valid density. Comparing this with the transition probability estimation outlined in Subsection 2.2, we can see that this is simply the continuous extension of $\lambda_{ij}/\sum_{j=1}^{K}\lambda_{ij}$. This relationship is convenient because it allows us to estimate transition probabilities or densities for any size partition. Estimating continuous transition densities permits us to balance fidelity and interpretability, generating transition surfaces that are both highly detailed and easy to explain.

2.4 Model choice and inference

In summary, we have created a general inference framework wherein one can estimate continuous Markov transition densities in two steps: first, assume that sequential elements of the Markov chain (X_{t-1}, X_t) are realizations from a Poisson point process (note that X_{t-1} and X_t can be real or vector valued) and estimate the corresponding point process intensity function $\Lambda((X_{t-1}, X_t))$. Second, calculate the conditional transition density for fixed X_{t-1} as

$$p(X_t|X_{t-1}) = \frac{\Lambda\big((X_{t-1}, X_t)\big)}{\int \Lambda\big((X_{t-1}, \xi)\big)d\xi}.$$

Because our estimate of the transition density is a valid density, variates from it can be easily simulated using any method that allows for sampling from a nonstandard distribution, i.e., rejection sampling [11].

A key attribute of the method presented in this paper is that the intensity function can be estimated by whatever method the user deems appropriate. One of the simplest examples is the kernel density estimator introduced by [12], where it is proposed that the intensity function is estimated by

$$\widehat{\Lambda}(\mathbf{s}) \equiv \frac{1}{p_b(\mathbf{s})}\sum_{i=1}^{m} K_b(\mathbf{s} - \mathbf{s}_i)$$

where $K_b(\cdot)$ is a kernel function with bandwidth b, and $p_b(\mathbf{s})$ is an edge correction that scales the intensity function to integrate to the appropriate count. While this kernel estimator is useful, in order to incorporate additional structural information a more sophisticated approach is necessary; a log Gaussian Cox process (LGCP) provides an elegant solution. Defined simply, a LGCP is a Poisson point process with $\Lambda(s) = \exp(Z(s))$, where $Z(s)$ is a Gaussian process (for a fuller treatment see

[9]). Modelling transition densities as a LGCP allows us to account for structural information by regressing the mean of $Z(s)$ on relevant covariates and through choice of covariance function. Because LGCP's are doubly stochastic they are computationally challenging to fit. Many methods exist to simplify this computation, including integrated nested Laplace approximation (INLA) [13], nearest-neighbors Gaussian processes [14], and predictive processes [15]. In the following results, we use process convolutions to overcome the computational expense, a method that replaces the latent Gaussian process $Z(s)$ with a basis function expansion reliant on a smoothing kernel [16].

3 Application to tracking data

Having established the mathematical relationship that makes representing a Markov model as a Poisson process possible, we now show how it can be applied. We are interested in modeling the movement of the ball around the court. Specifically, we will examine how that movement differs across teams, and how individual players and coaches impact the way the ball moves. We analyze data from all 1230 regular season games for the 2015-16 NBA season. In addition to the coordinate information for the players and the ball, the tracking data also includes variables indicating when an action such as a dribble, pass, or shot occurs. In order to prevent all of the transition density mass being centered around the current location of the ball, we thin the data to only use ball locations where an event is recorded, leaving us with 2,280,281 total transitions.

The movement of the ball is assumed to be a first-order Markov chain with shots, fouls, and turnovers serving as absorbing states. By considering each data point in terms of its origin and destination locations, we can model this as a four dimensional LGCP (one dimension for each x and y location in the origin and destination) using a process convolution model [16], as mentioned before.

Due to the scale of the data, some concessions must be made to improve computational performance. To this end, we assume that the covariance function is separable in the origin and destination dimensions and use a truncated normal distribution as our kernel for the process convolution. A compact kernel results in a sparse design matrix, and separability allows us to construct the full design matrix via Kronecker product. Additionally, we divide the court into 1.25×1.25 foot squares, resulting in $n = 1480$ grid cells and $n^2 = 2{,}190{,}400$ origin-destination grid cell combinations. This grid cell size is somewhat arbitrary, but provides a good balance between computational feasibility and a high level of resolution for the transition surface. In order to calculate the process convolution weights, we place $k = 50$ kernel locations in a hexagonal grid over the court region, resulting in $k^2 = 2500$ kernel locations in the origin-destination space.

A key factor in our decision to model transitions with a process convolution model is that it readily admits the inclusion of additional covariates. Because the location on the court where a shot originates impacts the value of a made basket, we do not expect the transition surface to vary smoothly across the three point line. We account for this discontinuity by dividing the court into $r = 15$ regions and include information about transitions between the different regions in the design matrix, resulting in a model of the form: $\log(\Lambda(\mathbf{s})) = Z\boldsymbol{\alpha} + X\boldsymbol{\beta}$. In this equation Z is a $n^2 \times r^2$ matrix of indicator variables, with a 1 in column $r(i-1)+j$ indicating that the transition began in court region i and ended in court region j, and X is a $n^2 \times k^2$ matrix of kernel weights. Thus, our process convolution model consists of both macro transitions between court regions and fine scaled spatial variation.

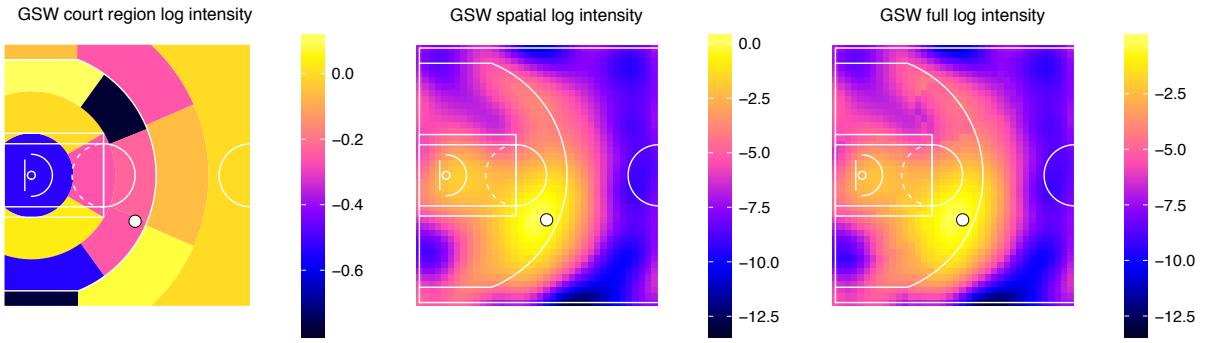

Figure 1: Log intensity surface for the Golden State Warriors (right), decomposed into court region transitions (left) and a spatial component (middle).

Before considering comparisons, we do an initial examination of the transition surface for a single team. Figure 1 shows the log transition surface for the Golden State Warriors. Note that because the full surface is four dimensional, we can only display what the surface looks like conditioned on a single point, indicated in these figures by the white circle. A web application that allows the user to explore these surfaces for all potential conditioning locations is available at https://jwmortensen.shinyapps.io/transition_surface_visualization. In Figure 1 we have decomposed the full log intensity surface into its component parts: the macro transitions between court regions and the continuous spatial surface. These plots depict the intensity surface on the log scale, so they do not have the clean interpretation of the Markov transition densities, but considering plots on the log scale makes it easier to see how the transition surface varies over space. Here we see that the full log intensity manifests the discontinuous effects of the macro transitions, most pronounced at the top right arc of the three point line and near the left corner three. In the area directly surrounding the conditioning location, the probability of transitioning in front of or behind the three point line is close to equal, but the discontinuity grows larger further away from the conditioning location. This could potentially be explained by the fact that transition probabilities due to dribbling are likely less impacted by the three point line than transitions due to passing.

By calculating transition surfaces under a variety of different conditions we can begin to understand how these conditions impact ball movement. First we compare two different teams, the 2015-16 Cleveland Cavaliers and 2015-16 Golden State Warriors. In order to compare and interpret surfaces, we convert log intensities, like the one shown in Figure 1, into conditional transition densities. These surfaces can be interpreted as showing the likelihood of transitioning to any location on the court from the location indicated by the white dot. In order to facilitate easier comparison, we take the difference of the density surfaces, shown in the final panel of Figure 2. This surface is calculated by subtracting the Golden State surface from the Cleveland surface so that positive values indicate higher transition density for the Cavaliers and vice versa. Much of the surface shows that there is essentially no difference in transition densities, but we can see that the Warriors have a higher probability of moving into the key directly in front of the white dot than the Cavaliers. A glance at the scale may lead one to believe that this difference is so small as to be meaningless, but bear in mind that the reason these densities are small is because the

normalization for the conditional density is occurring over such a large region. If we integrate over the blue area in front of the white circle (the area indicating larger transition density for Golden State) then we get -0.1109, which can be interpreted as, "for every 100 possessions that start at the white circle the Warriors will have approximately 11 more possessions move into that region than the Cavaliers." Because the nature of basketball is such that a few points per one hundred possessions often means the difference between winning and losing, this is a potentially consequential difference.

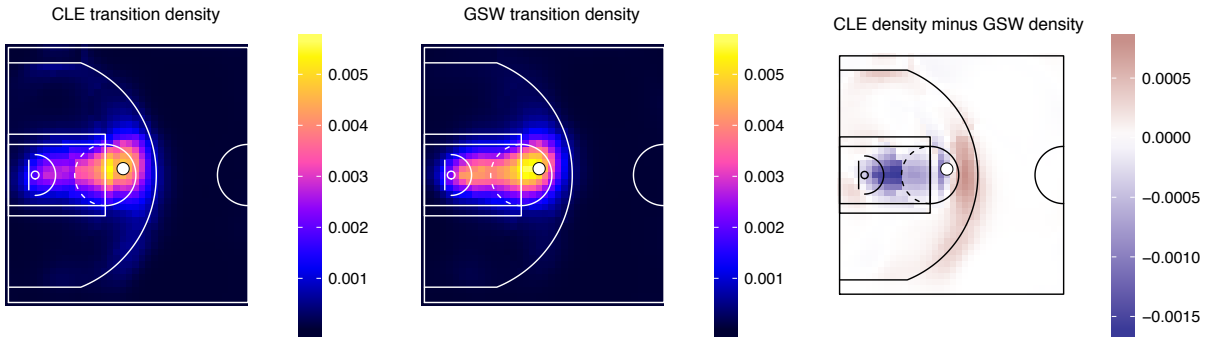

Figure 2: Transition density surfaces for the Cleveland Cavaliers (left) and Golden State Warriors (middle) for the 2015-16 season. The origin point for these surfaces is indicated by the white circle. The final plot is the difference between Cleveland and Golden State's transition density surfaces. This is calculated by subtracting the surface for the Warriors from the surface for the Cavaliers, so negative (blue) values indicate areas for which the Warriors have higher transition densities, while positive (red) values indicate areas where the Cavaliers have higher transition densities.

We can perform a similar comparison to examine how individual actors impact ball movement on the court, which we do here for Golden State Warrior's point guard Stephen Curry and for former Cleveland Cavalier's head coach David Blatt. To assess how ball movement changes when Stephen Curry is present, we partitioned all of Golden State's transitions based on whether or not Curry was currently on the court. The left panel of Figure 3 shows the difference surface for Curry, conditioned on the same location we used to compare Golden State and Cleveland. Of course, any observed changes cannot be directly attributed to Curry due to collinearity with his fellow players, but interestingly, we see a near reversal of the pattern revealed in Figure 2. With Curry on the court the probability of moving into the red region in front of the white dot is 0.151 higher than when he is off, indicating that the increased probability we saw for Golden State to move into the key in Figure 2 is primarily due to Curry and whichever players share most of his minutes.

In addition to examining player effects, we wanted to see if we could capture differences in coaching. David Blatt, who was the head coach of the Cleveland Cavaliers at the start of the 2015-16 NBA season until he was fired on 22 January 2016, presented an ideal opportunity. We compared ball movement for the 41 games he served as Cleveland's coach to the 41 regular season games the Cavaliers played after his firing in an attempt to assess what affect coaching may have had on the Cavaliers' play style. From the plot on the right side of Figure 3 we can see that when David Blatt was coach the probability of moving from that specific location on the right side of the key to a wide range of locations was slightly elevated, whereas the probability mass has a much higher concentration around the conditioning location for the games played after his coaching

tenure was over. This pattern would seem to suggest that there was greater offensive entropy and increased ball movement with Blatt as coach.

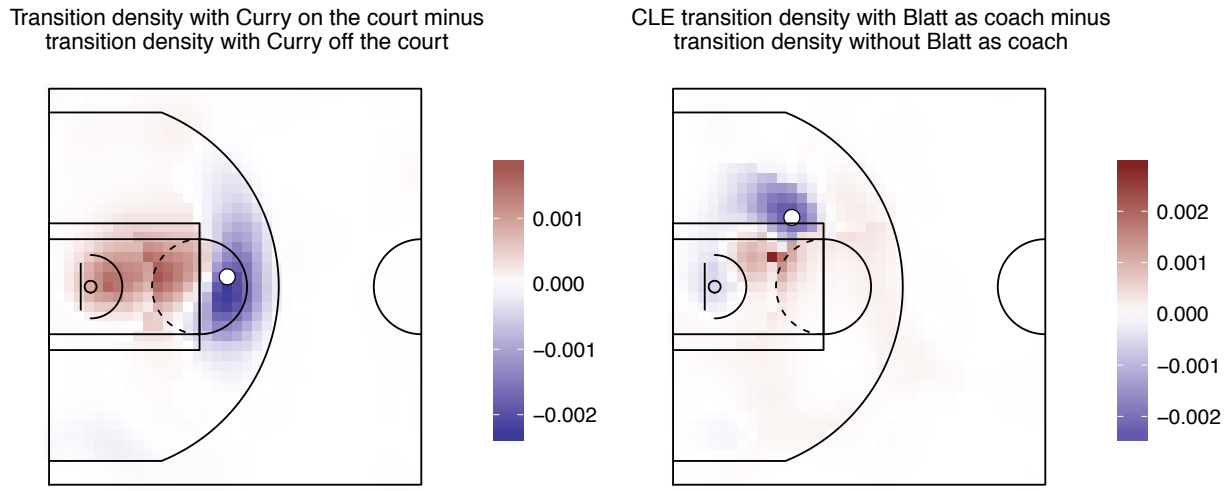

Figure 3: Difference surfaces for Stephen Curry (left), the starting point guard for the Golden State Warriors, and David Blatt (right), former head coach for the Cleveland Cavaliers. For Curry, areas of red indicate regions where the transition density is higher when Curry is on the court. For Blatt, red regions indicate areas where the transition density was higher while he was coach.

4 Future work and conclusions

In Section 2 we established the mathematical foundation for modeling Markov transitions as Poisson point processes, and in Section 3 we showed how they could be used, but there are still several opportunities for further research. Thus far we have assumed that transition densities are temporally homogeneous. Due to the nature of sports in general, this is a simplifying assumption that will frequently be violated; figuring out how to adapt this method to account for time-varying transitions is an intriguing option for further work. An additional research path is to focus on scalability. We have successfully modeled two-dimensional data, but borrowing strength across origins and destinations required us to fit a four-dimensional Poisson process. More work needs to be done to extend this method to three-dimensional data and beyond. Fortunately, for most sports, tracking data is only available in two-dimensions anyway, so this should not impede adoption of this framework.

Our examples in this paper have been confined to the NBA, but this approach can be used to produce transition estimates in all sports where tracking data is available. Additionally, although we estimated transition surfaces in Section 3 and used them to examine teams, players, and coaches, this is merely one example of how this methodology can be used. This framework is valuable anywhere accurate representations of movement are required, and could provide immediate benefits in cases, such as expected possession value (EPV) [3], that have previously been using simple parametric models to capture player movement. One particularly nice feature of the point process approach as it applies to sports is that it simplifies the inclusion of spatial features

that can impact transitions, shown by the discontinuity we captured at the three point line in Figure 1.

In conclusion, we have presented a straightforward framework to modeling movement data that allows great complexity while maintaining interpretability. By using Poisson point processes to model Markov transitions, we have extended Markov models to continuous space nonparametrically, which balances the high fidelity provided by black-box methods without losing the inherent meaning provided by conditional transition densities.

References

[1] H. M. Le, P. Carr, Y. Yue and P. Lucey, "Data-driven ghosting using deep imitation learning," *Sloan Sports Analytics Conference,* pp. 1-15, 2017.

[2] K.-C. Wang and R. Zemel, "Classifying NBA offensive plays using neural networks," *Sloan Sports Analytics Conference,* pp. 1-9, 2016.

[3] D. Cervone, A. D'Amour, L. Bornn and K. Goldsberry, "A multiresolution stochastic process model for predicting basketball possession outcomes," *Journal of the American Statistical Association,* vol. 111, no. 514, pp. 585-599, 2016.

[4] M. West and J. Harrison, Bayesian Forecasting and Dynamic Models, Berlin, Heidelberg: Springer-Verlag New York, Inc, 1997.

[5] M. B. Rajarshi, Statistical Inference for Discrete Time Stochastic Processes, India: Springer India, 2013.

[6] M. W. Birch, "Maximum Likelihood in Three-Way Contingency Tables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 25, no. 1, pp. 220-233, 1963.

[7] J. Palmgren, "The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables," *Biometrika,* vol. 63, no. 19, pp. 1-25, 1981.

[8] S. G. Baker, "The multinomial-Poisson transformation," *Journal of the Royal Statistical Society. Series D (The Statistician),* vol. 43, no. 4, pp. 495-504, 1994.

[9] J. Møller and R. P. Waagepetersen, Statistical Inference and Simulation for Spatial Point Processes, Boca Raton, FL: CRC Press, 2004.

[10] S. Banerjee, B. P. Carlin and A. E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, Boca Raton, FL: CRC Press, 2015.

[11] J. von Neumann, "Various techniques used in connection with random digits," in *Monte Carlo Method*, Washington D.C., National Bureau of Standards Applied Mathematics Series, 1951, pp. 36-38.

[12] P. Diggle, "A kernel method for smoothing point process data," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 34, no. 2, pp. 138-147, 1985.

[13] F. Lindgren and H. Rue, "Bayesian spatial modeling with R-INLA," *Journal of Statistical*

Software, vol. 63, no. 19, pp. 1-25, 2015.

[14] A. Datta, S. Banerjee, A. O. Finley and A. E. Gelfand, "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets," *Journal of the American Statistical Association,* vol. 111, no. 514, pp. 800-812, 2016.

[15] S. Banerjee, A. E. Gelfand, A. O. Finley and H. Sang, "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 70, no. 4, pp. 825-848, 2008.

[16] D. M. Higdon, "Space and space-time modeling using process convolutions," *Quantitative methods for current environmental issues,* pp. 37-54, 2002.

[17] H. Rue, S. Martino and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 71, no. 2, pp. 319-392, 2009.