Real-Time Prediction of Basketball Outcomes Using High-Resolution Spatio-Temporal Tracking Data

Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry

Department of Statistics, Harvard University Center for Geographic Analysis, Harvard University

Goal

Using optical tracking data, estimate the value of *every action* in a basketbal game, including non-terminal actions like passes and dribble penetrations that do not show up in the box score.

NBA Optical Tracking Data

The SportVU optical tracking data consists of x, y locations on the court of all 10 players—and x, y, z locations of the ball—25 times per second. A full season of data is over a billion points in space-time.

Expected Possession Value (EPV)

Expected Possession Value (ν_t) is the number of points (X) the offense is expected to score on a particular possession, given its spatiotemporal history up to time t ($\mathcal{F}_t$):

$$\nu_t = \mathbb{E}[X|\mathcal{F}_t].$$

ν_t averages over all possible future courses of the possession. Temporally coherent estimation requies a stochastic process model of basketball.

Multiresolution Modeling

Jointly model full-resolution basketball process (Z_t) and coarsened view of the process ($C_t = \mathscr{C}(Z_t)$).

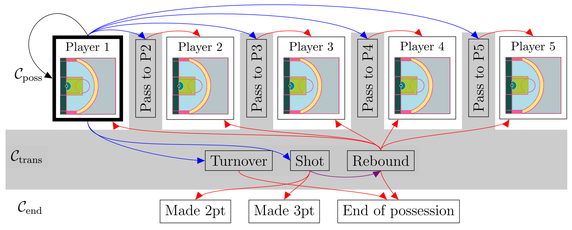

Figure: Schematic of coarsened state process C_t. Macrotransition states shaded in gray.

- dcervone@fas.harvard.edu
- damour@fas.harvard.edu
- bornn@stat.harvard.edu
- kgoldsberry@fas.harvard.edu

Estimated EPV Curve

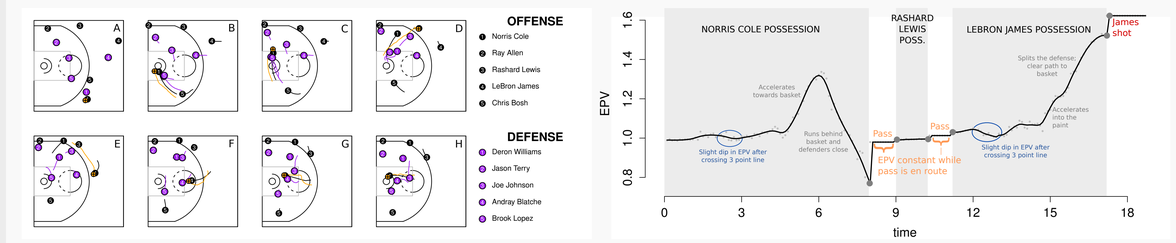

Figure: Miami Heat possession against Brooklyn Nets with estimated EPV curve. Norris Cole wanders the perimeter (A) then drives toward the basket (B). Instead of taking the shot, he runs underneath the basket (C) and passes to Rashard Lewis (D), who passes to LeBron James (E). After entering the perimeter (F), James slips behind the defense (G) and scores an easy layup (H).

Model Components: Two Transition Types

Macrotransition Model

Macrotransitions encapsulate discrete basketball actions (e.g. passing and shooting). Modeled using competing risks [4]. Assuming player ℓ is the ballcarrier, the hazard of macrotransition j at time t is

$$\log(\lambda_j^\ell(t)) = [\mathbf{W}_j^\ell(t)]'\boldsymbol{\beta}_j^\ell + \xi_j^\ell\left(\mathbf{z}_\ell(t)\right),$$

where $\mathbf{W}_j^\ell(t)$ are time-varying covariates and ξ_j^ℓ a Gaussian Process-distributed spatial random effect.

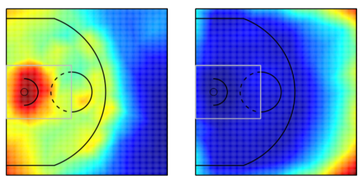

Figure: Posterior mean and SD for spatial effect ξ in LeBron James' shot-taking hazard model.

Microtransition Model

Microtransitions represent player movement between macrotransitions.

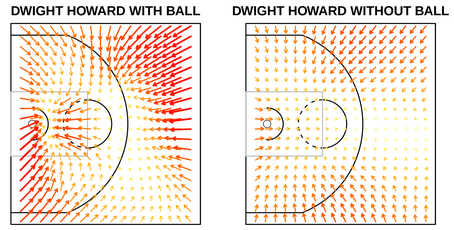

Figure: Acceleration fields (μ_x, μ_y) for Dwight Howard.

Classical dynamics motivates an ARI(1,1) model for forecasting a player's location ($\mathbf{z}(t)$) conditional on a fixed ballcarrier. Acceleration fields modeled with a Gaussian Process prior.

Evaluating EPV with Multiresolution Transitions

Letting τ and δ be the start and end times of the next macrotransition to complete, $M_j(\tau)$ be the identity of this macrotransition, and Z_s and C_s denote the full-resolution and coarsened states at time s, under conditional independence assumptions, rewrite EPV as

$$\nu_t = \sum_{c \in \mathcal{C}} E[X|C_\delta = c]\left(\int_t^\infty \int_{\mathcal{Z}} \mathbb{P}(C_\delta = c, M(\tau)|Z_\tau = z, \tau = s)\mathbb{P}(Z_\tau = z, \tau = s|\mathcal{F}_t)dzds\right).$$

- Computed from long-run distribution of C, derived from Macrotransition model.
- Computed directly from Macrotransition model.
- Computed directly from Microtransition model.
- Integration by Monte Carlo.

Independence Assumptions

Assumptions necessary to make EPV estimation computationally tractable

- Macrotransitions decouple the outcome from fine-grained information in $\mathcal{F}_t$.
- Coarsened process C_t is marginally Semi-Markov (sojourn times not exponential).

Hierarchical Modeling

Hierarchical modeling of unknown components in our transition models provides shrinkage necessary for good predictive performance.

- Functional basis representation of spatial surfaces ξ_j^ℓ more appropriate for hierarchical models [3].
- A player similarity network, learned by the spatial distribution of players' time on the court, provides CAR priors for model parameters across players.
- Inference performed using R-INLA software [5].
- Scale of data necessitates implementation on high-performance, distributed computing systems.

Results

EPV curves (ν_t) calculated for every NBA possession in 2013-2014 season, revealing the real-time "stock price" of a possession. Using these, analysts can identify the most impactful actions during a possession and estimate novel ways in which players contribute value to the offense. For more details, please see the authors' additional work on this project [1, 2].

References

[1] D. Cervone, A. D'Amour, et al. "A Multiresolution Semi-Markov Process Model for Predicting Basketball Possession Outcomes." *In preparation* (2014).

[2] D. Cervone, A. D'Amour, et al. "Pointwise: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data." *MIT SSAC* (2014).

[3] F. Lindgren, H. Rue, S. Martino, et al. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *JRSSB* ***73.4*** (2011): 423–498.

[4] R. Prentice, J. Kalbfleisch, et al. "The analysis of failure times in the presence of competing risks." *Biometrics* (1978): 541–554.

[5] H. Rue, et al. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *JRSSB* ***71.2*** (2009): 319–392.