A Bayesian Joint Model for Spatial Point Process with Application to Basketball Shot Chart

Jieying Jiao
Department of Statistics, University of Connecticut
Joint work with Guanyu Hu and Jun Yan

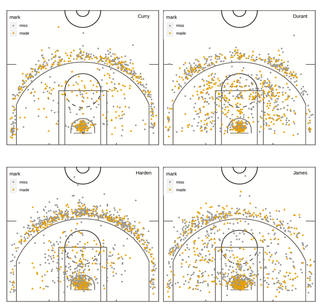
Introduction

- Shot chart are important summaries for basketball coaches. But no work has been done to jointly model shot location and shot success indicator, which is called mark.
- Shot success rate maybe higher at locations where more shots are made.
- We propose a Bayesian joint model of marked spatial point process to model shot location and mark simultaneously. And use spike-slab prior to do variable selection.

Data Overview

NBA player shot chat data of 2017-2018 regular season (`stats.nba.com`).

Player Name	Shot Number	Success rate (%)	2-point shot(%)	Period(%)
Curry	740	50.0	43.4	(35.0, 20.7, 34.1, 9.9, 0.4)
Durant	1032	52.5	66.7	(30.9, 23.6, 30.4, 14.7, 0.3)
Harden	1286	45.6	50.6	(28.7, 22.3, 27.7, 21.0, 0.3)
James	1409	54.3	74.6	(29.1, 21.5, 25.3, 23.6, 0.4)



Model Setup

- With location data $\mathbf{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ and associated mark data $\mathbf{M} = (m(\boldsymbol{s}_1), \ldots, m(\boldsymbol{s}_N))$, we propose to use non-homogeneous Poisson point process (Diggle, 2013) and logistic regression, respectively. Intensity dependent model (Ho and Stoyan, 2008) is incorporated for the joint model.
- Joint Likelihood: $L(\boldsymbol{\Theta} \mid \mathbf{S}, \mathbf{M}) \propto L(\lambda_0, \boldsymbol{\beta} \mid \mathbf{S}) \times L(\xi, \boldsymbol{\alpha} | \mathbf{M}, \boldsymbol{\lambda})$, $\boldsymbol{\Theta} = (\lambda_0, \boldsymbol{\beta}, \xi, \boldsymbol{\alpha_0})$.
- Intensity Model:

$$L(\lambda_0, \boldsymbol{\beta} \mid \mathbf{S}) \propto \left(\prod_{i=1}^{N} \lambda(\boldsymbol{s}_i) \right) \exp\left(- \int_{\mathcal{B}} \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \right),$$
$$\lambda(\boldsymbol{s}_i) = \lambda_0 \exp\left(\mathbf{X}^\top(\boldsymbol{s}_i) \boldsymbol{\beta} \right),$$

where λ_0 is the baseline intensity, $\mathbf{X}(\boldsymbol{s}_i)$ is a $p \times 1$ spatially varying covariate vector, and $\boldsymbol{\beta}$ is the corresponding coefficient vector.

- Mark Model:

$$L(\xi, \boldsymbol{\alpha} | \mathbf{M}, \boldsymbol{\lambda}) = \prod_{i=1}^{N} \theta(\boldsymbol{s}_i)^{m(\boldsymbol{s}_i)} \left(1 - \theta(\boldsymbol{s}_i)\right)^{1 - m(\boldsymbol{s}_i)},$$
$$\mathrm{logit}\big(\theta(\boldsymbol{s}_i)\big) = \xi \lambda(\boldsymbol{s}_i) + \mathbf{Z}^\top(\boldsymbol{s}_i) \boldsymbol{\alpha},$$

where $\lambda(\boldsymbol{s}_i)$ is the intensity defined above with a scalar coefficient ξ, $\mathbf{Z}(\boldsymbol{s}_i)$ is a $q \times 1$ spatially varying covariate vector, and $\boldsymbol{\alpha}$ is a $q \times 1$ vector of coefficient.

Bayesian Inference

- Prior Specification: Conjugate prior for λ_0 and vague prior for other parameters.

$$\lambda_0 \sim \mathrm{G}(a, b), \qquad \boldsymbol{\beta} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \mathbf{I}_p),$$
$$\xi \sim \mathrm{N}(0, \delta^2), \qquad \boldsymbol{\alpha} \sim \mathrm{MVN}(\mathbf{0}, \delta^2 \mathbf{I}_q),$$

- Variable Selection: Spike-Slab prior on covariates coefficients.

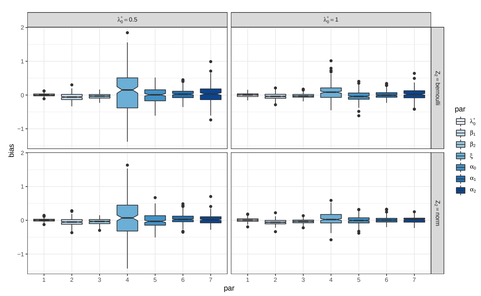
$$\alpha_i \sim \mathrm{N}(0, \delta_i^2), \qquad \delta_i^2 = 0.01(1 - \gamma_i) + 100\gamma_i,$$
$$\gamma_i \sim \mathrm{Bernoulli}(\phi_i), \qquad \phi_i \sim \mathrm{Beta}(0.5, 0.5).$$

- MCMC Sampling Schemes: Metropolis-Hasting (MH) within Gibbs algorithm (Using R package nimble).

  Sampler: `RW()` for ϕ_i's, `binary()` for γ_i's, and `RW_llFunction()` for others.
- Bayesian Model Selection Criteria: mDIC and mLPML on mark model only (Ma, Chen and Hu, 2018) to test intensity independent or dependent model ($\xi = 0$ or not).

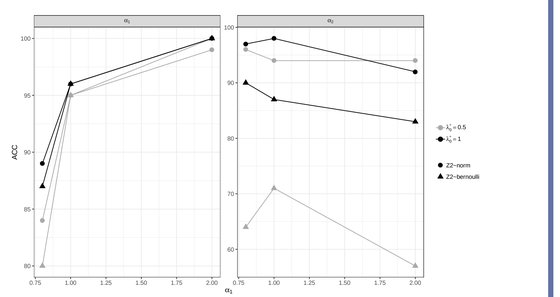
Simulation–Estimation

- On area $\mathcal{B} = [-1, 1] \times [-1, 1]$, $\mathbf{X}^\top(\boldsymbol{s}_i) = (x_i, y_i)$, $\mathbf{Z}^\top(\boldsymbol{s}_i) = (1, Z_{1i}, Z_{2i})$ from $\mathrm{N}(0, 1)$ or $\mathrm{Bern}(0.5)$, $\boldsymbol{\beta}^\top = (2, 1)$, $\xi = \alpha_0 = 0.5$, $\alpha_2 = 1$, $\lambda_0 = 100\lambda_0^*$, $\lambda_0^* \in \{0.5, 1\}$, $\alpha_1 \in \{0.8, 1, 2\}$. No variable selection and vague prior was used. 20,000 MCMC chain with 10,000 burn-in.
- No obvious trend in estimation performance when α_1 increases. Plot below shows the results when $\alpha_1 = 1$.



Simulation–Variable Selection

- New covariates $Z_3, \ldots, Z_4 \sim \mathrm{N}(0, 1)$ in marks model, $\alpha_3, \ldots, \alpha_4$ equals 1 or 0, number of significant $Z_i's$ are $\{2, 4, 6\}$ (Spike-Slab used). Others unchanged.
- Results shows that the selection accuracy rates for those zero valued coefficients are always 100%.
- Plot below shows the results when only α_1, α_2 are non-zero.



Real Data Analysis

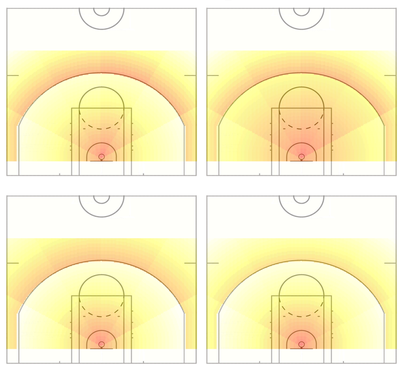
- Intensity Model Covariates:

Covariates	Comment
3-point shot	binary
distance1	standardized, distance to origin for 2-point shot, 0 for 3-point shot
distance2	standardized, 0 for 2-point shot, distance to 3-point line for 3-point shot
angle	reference level is $[-\frac{\pi}{2}, \frac{\pi}{6}]$

- Mark Model Covariates:

Covariates	Comment
intensity value	
seconds	left towards the end of the period, scaled with 100
3-point shot	binary
period	reference level is first period
opponent	1 for playoff teams and 0 for others
distance to origin	standardized
angle	reference level is $[-\frac{\pi}{2}, \frac{\pi}{6}]$

- Fitted intensity heat plot for four players and parameters estimation results for Curry's data.



Model	Covariate	Posterior Mean	Posterior SD	95% Credible Interval
Intensity	baseline (λ_0)	0.74	0.06	(0.62, 0.87)
	3-point shot	−3.15	0.22	(−3.58, −2.74)
	distance1	−2.10	0.05	(−2.20, −2.01)
	distance2	−2.32	0.17	(−2.67, −1.99)
	angle $[\pi/6, \pi/3]$	0.42	0.09	(0.23, 0.60)
	angle $[\pi/3, \pi/2]$	0.58	0.09	(0.40, 0.75)
	angle $[\pi/2, 2\pi/3]$	0.58	0.09	(0.40, 0.76)
	angle $[2\pi/3, 5\pi/6]$	0.41	0.09	(0.23, 0.60)
	angle $[5\pi/6, 3\pi/2]$	−0.22	0.11	(−0.43, 0.00)
Mark	Intercept	−0.75	0.15	(−1.04, −0.46)
	λ	0.09	0.01	(0.06, 0.11)

Future Work

- Interaction among different players.
- Multivariate model for more players jointly.
- Non-parametric method for intensity model.
- DIC and LPML for joint model.