A Bayesian Enhanced Baseline-and-Residual Model for Advanced Basketball Analytics

Zhengze Zhang

January 20, 2025

Abstract

This paper presents a novel "baseline-and-residual" model for evaluating basketball players' on-court impact, leveraging a comprehensive set of non-traditional metrics (including defensive matchups, hustle events, spacing analysis, and Synergy-based play-type data). Our framework replaces the typical box-score-centric approach with a Bayesian-based baseline prior that helps address small-sample player uncertainty. We then apply a multi-round residual correction, reminiscent of a regularized plus-minus method, on a possession-level basis to distribute on-court success or failure across players actually on the floor. The Bayesian prior mitigates overfitting for players with limited data and enables us to assign confidence intervals to final ratings. Empirical considerations suggest this approach can reduce reliance on raw box-score "inflation" and better highlight under-the-radar contributions such as screen assists, forced turnovers, or synergy-driven playmaking.

1 Introduction

Advanced basketball analytics has long debated how to separate genuine impact from superficial box-score accumulation. Models such as BPM [1], RAPM [2, 3], and more recent hybrids like EPM [4] or LEBRON have stressed partial solutions: some rely heavily on box-score data, others rely purely on on/off plus-minus. Often, small-sample players or outlier performances remain problematic, leading to noisy or misrepresented estimates of impact.

We propose here a Bayesian Enhanced Baseline-and-Residual Model (**BEBRM**), designed to (i) incorporate a broad non-box-score dataset—covering defensive matchups, hustle records, spacing metrics, and synergy-based play-type information, and (ii) impose a

Bayesian prior to regularize small-sample player estimates and produce confidence intervals. In the second stage, the model applies a regularized plus-minus approach at the possession level, distributing the team's performance deviation from the baseline across the actual on-court players, via iterative ridge regression. This approach aims to identify genuine contributions such as forced turnovers, space creation, or specialized synergy skill sets, rather than weighting simple points or rebounds alone.

2 Data Sources

The method described requires multiple NBA API endpoints, which offer more detailed insights beyond traditional box scores. Let $\mathcal{N}$ index all players of interest in a given season. We summarize the key data sources below:

1. **PlayByPlay (v1/v2/v3).** Provides possession-level events (scoring, missed shots, substitutions, etc.). Used to:
   - Identify which players are on/off the floor for each possession.
   - Retrieve the actual scoring outcome of each possession.
2. **Matchups (BoxScoreMatchupsV3 / LeagueSeasonMatchups).** Records:
   - Opponent field goal attempts, made shots, help blocks, forced turnovers, etc., in direct defensive matchups.

   This underlies a more precise defensive prior.
3. **HustleStatsBoxScore.** Provides:
   - Contested shots, deflections, charges drawn, screen assists, among other hustle metrics.

   Useful for capturing intangible frontcourt or off-ball contributions.
4. **ShotChartDetail / ShotChartLeagueWide.** Supplies:
   - Spatial shot coordinates $(\text{LOC\_X}, \text{LOC\_Y})$, shot distance, and outcome.
   - Potential for spacing/shot-distribution metrics and partial luck (three-point outlier) adjustments.

5. **TeamPlayerOnOffSummary / BoxScoreAdvanced / FourFactors.** Macro-level on/off efficiency (OFF_RATING, DEF_RATING), four-factor stats, and advanced metrics (TS_PCT, EFG_PCT, etc.). These provide additional prior knowledge.

6. **SynergyPlayTypes.** Key to capturing specific *play-type* performance such as isolation, pick-and-roll (PnR) ball-handling, spot-ups, or post-ups, each with points-per-possession and usage rate. These form specialized prior components for offensive and defensive roles.

In the sequel, we form each player's feature vector (for offense, defense, or both) by merging these endpoints.

3 Bayesian Baseline Prior

3.1 Feature Construction

Define for each player $i \in \mathcal{N}$ an expanded feature vector

$$\mathbf{X}_i = \big[X_{i,1},\, X_{i,2}, \ldots, X_{i,m}\big],$$

where each component $X_{i,k}$ is derived from the above data (matchups, hustle, synergy, etc.). Typical examples include:

- **Offensive synergy metrics** (e.g. isolation PPP, pick-and-roll ball-handler PPP, transition PPP);
- **Hustle-based screen assists**, e.g. ratio of `SCREEN_ASSISTS` to `MIN`;
- **Matchup-based FG% difference** for defenders faced, if relevant to offense or potential mismatch creation;
- **ShotChart spacing improvement**, measuring how much the player's presence elevates corner-threes or rim attempts for teammates;
- **On/Off-based net rating differentials**, to capture macro-level patterns;
- **Four-factor advanced stats**, capturing usage, turnover rates, eFG%, etc.

3.2 Bayesian Regularized Regression

We then define a response variable Y_i for each player i, typically obtained from a known large-sample measure of real plus-minus or an aggregated team impact measure. Denote $\boldsymbol{\theta}$ as the coefficient vector to be fitted. The linear model is:

$$Y_i \;=\; \boldsymbol{\theta}^{\top}\mathbf{X}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

A Gaussian prior is placed on $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \;\sim\; \mathcal{N}(\mathbf{0},\, \alpha^2\,\mathbf{I}). \tag{2}$$

This corresponds to a "Ridge-like" Bayesian perspective, ensuring shrinkage of $\boldsymbol{\theta}$. Posterior inference can be performed via Bayesian ridge regression in closed form, or via MCMC/variational Bayes. The posterior distribution is:

$$p(\boldsymbol{\theta} \;|\; \{(\mathbf{X}_i, Y_i)\}_{i=1}^{N}) \;\propto\; \exp\Bigl\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \boldsymbol{\theta}^{\top}\mathbf{X}_i)^2 - \frac{1}{2\alpha^2}\|\boldsymbol{\theta}\|^2\Bigr\}. \tag{3}$$

Hence, the posterior mean $\hat{\boldsymbol{\theta}}$ yields the best linear combination of features for approximating Y_i, while small-sample or uncertain features are aggressively shrunk.

Let $\hat{\boldsymbol{\theta}}$ denote the final posterior mean. We split it into an offensive portion $\hat{\boldsymbol{\theta}}^{(\mathrm{off})}$ and a defensive portion $\hat{\boldsymbol{\theta}}^{(\mathrm{def})}$ by training (1)–(3) separately for an offensive target and a defensive target. Then, each player i's baseline rating is:

$$\beta_i^{(\mathrm{base,\,OFF})} \;=\; \bigl(\hat{\boldsymbol{\theta}}^{(\mathrm{off})}\bigr)^{\top}\mathbf{X}_i^{(\mathrm{off})}, \quad \beta_i^{(\mathrm{base,\,DEF})} \;=\; \bigl(\hat{\boldsymbol{\theta}}^{(\mathrm{def})}\bigr)^{\top}\mathbf{X}_i^{(\mathrm{def})}. \tag{4}$$

Crucially, the posterior variance indicates how reliably we know each coefficient; small-sample players i see heavier shrinkage or broader intervals.

3.3 Small-Sample Robustness and Uncertainty

This Bayesian approach naturally addresses player i who has limited minutes or synergy possessions: the regression infers large posterior variance for relevant coefficients. Concretely, $\beta_i^{(\mathrm{base})}$ is forced closer to zero (or average) unless sufficient data implies otherwise. One can further provide a $(1-\alpha)$ posterior interval for $\beta_i^{(\mathrm{base,OFF})}$, revealing the level of confidence in this baseline rating.

4 Possession-Level Residual Correction

4.1 Motivation

While $\beta_i^{(\text{base,OFF/DEF})}$ incorporates essential synergy/hustle/matchups knowledge, it does not explicitly factor in *which lineups* or *opponents* a player faces in each possession, nor does it incorporate real-time on-court synergy across teammates. Hence, we apply a *possession-level* regression reminiscent of RAPM to capture the difference between actual outcomes and the predicted baseline sum.

4.2 Model Setup

Let N_{pos} be the total number of offensive possessions in the dataset. For each possession ν $(1 \le \nu \le N_{\text{pos}})$:

- $\mathcal{A}_\nu$: the set of offensive players on the floor for that possession;
- y_ν^{OFF}: actual points scored (possibly after partial luck adjustment for three-point anomalies if desired).

We define an initial baseline prediction:

$$\hat{y}_\nu^{(0,\text{OFF})} \;=\; \frac{1}{|\mathcal{A}_\nu|} \sum_{i \in \mathcal{A}_\nu} \beta_i^{(\text{base,OFF})}, \tag{5}$$

where $\beta_i^{(\text{base,OFF})}$ is from (4). The *residual* for possession ν is

$$r_\nu^{\text{OFF}} \;=\; y_\nu^{\text{OFF}} \;-\; \hat{y}_\nu^{(0,\text{OFF})}. \tag{6}$$

4.3 Iterative Ridge Regression on Residual

We proceed with an iterative approach to regularized regression on r_ν^{OFF}:

1. **Initialization:** set $\beta_i^{(0,\text{OFF})} \;=\; \beta_i^{(\text{base, OFF})}$ for all i.
2. **At iteration t:**

$$r_\nu^{\text{OFF},(t)} \;=\; y_\nu^{\text{OFF}} \;-\; \frac{1}{|\mathcal{A}_\nu|} \sum_{i \in \mathcal{A}_\nu} \beta_i^{(t-1,\text{OFF})}, \tag{7}$$

   and let $x_{\nu,i} = 1$ if $i \in \mathcal{A}_\nu$ and 0 otherwise. Solve the ridge problem

$$\{\Delta_i^{\text{OFF},(t)}\} \;=\; \arg\min_{\{\Delta_i\}} \sum_{\nu=1}^{N_{\text{pos}}} \Big(r_\nu^{\text{OFF},(t)} \;-\; \sum_i x_{\nu,i}\,\Delta_i \Big)^2 \;+\; \lambda_{\text{OFF}} \sum_i (\Delta_i)^2. \tag{8}$$

3. **Update:** for each player i,

$$\beta_i^{(t,\mathrm{OFF})} \;=\; \beta_i^{(t-1,\mathrm{OFF})} \;+\; \rho\,\Delta_i^{\mathrm{OFF},(t)}, \tag{9}$$

   where ρ is a step-size (often set to 1 for a direct update).

4. **Repeat** until convergence or a fixed iteration T.

After final iteration T, define

$$\beta_i^{(\mathrm{final},\,\mathrm{OFF})} \;=\; \beta_i^{(T,\mathrm{OFF})}. \tag{10}$$

Comments

- This multi-round update ensures we do not rely on a single pass of (r_ν^{OFF}), letting the model iteratively refine each player's offset from the baseline.
- If a small-sample player i has large *posterior variance* in $\beta_i^{(\mathrm{base},\,\mathrm{OFF})}$, initially the rating is drawn toward average. Only repeated positive outperformance in actual possessions will significantly push $\beta_i^{(t,\mathrm{OFF})}$ upward across iterations.

4.4 Defensive Possessions

A parallel procedure applies for defensive possessions ($\mathcal{B}_\nu$ indicates defenders, y_ν^{DEF} is points allowed). We similarly solve

$$r_\nu^{\mathrm{DEF},(t)} \;=\; y_\nu^{\mathrm{DEF}} \;-\; \frac{1}{|\mathcal{B}_\nu|}\sum_{j\in\mathcal{B}_\nu}\beta_j^{(t-1,\mathrm{DEF})}, \tag{11}$$

using a ridge solution to distribute r_ν^{DEF} among $j \in \mathcal{B}_\nu$.

4.5 Final Ratings

The final overall rating can be the sum of final offensive and defensive ratings:

$$\beta_i^{(\mathrm{final},\,\mathrm{TOTAL})} \;=\; \beta_i^{(\mathrm{final},\,\mathrm{OFF})} \;+\; \beta_i^{(\mathrm{final},\,\mathrm{DEF})}. \tag{12}$$

Optionally, separate offensive and defensive values can be reported, along with posterior intervals from the baseline Bayesian step.

5 Discussion of Advantages

5.1 Mitigating Small-Sample Instability

By incorporating a Bayesian prior in the baseline step, the model *automatically* shrinks ratings of players with scarce data. This mitigates inflated or deflated assessments triggered by a few outlier events.

5.2 Highlighting Non-Box-Score Contributions

Our baseline vector $\mathbf{X}_i$ includes synergy play-type effectiveness, hustle stats (screen assists, deflections, etc.), spacing metrics from shot charts, and advanced on/off data. Hence, the baseline rating focuses less on raw box-score tallies and more on intangible or context-driven strengths. The iterative residual correction further accounts for actual in-game lineup synergy.

5.3 Posterior Uncertainty

Unlike classical linear regression or BPM-like approaches, the Bayesian prior yields a posterior distribution, delivering confidence intervals on the baseline rating. This clarity is especially beneficial in front-office or analytics discussions, where a small-sample breakout performer might show a wide confidence band, emphasizing the potential risk or reward.

5.4 Comparison to Existing Models

- *BPM* [1] relies on box-score only, ignoring hustle or synergy data. Our approach extends coverage to non-traditional metrics, plus possession-based alignment.
- *RAPM* [2, 3] uses pure on/off lineups, but lacks a baseline layer. Our Bayesian prior aggregates synergy, hustle, matchups, thus clarifying player roles before distributing residual performance.
- *LEBRON/EPM* similarly embed a prior plus an on/off regression. However, many public models rely mainly on standard or partially advanced box-score features. Our approach explicitly incorporates synergy-based PPP, hustle, and spacing, which can further highlight intangible contributions.

6 Conclusion

We have introduced a new Bayesian Enhanced Baseline-and-Residual Model (**BEBRM**) that blends non-traditional metrics (defensive matchups, hustle data, synergy-based play-type analysis, and shot-chart spacing) with an iterative plus-minus approach. The Bayesian prior in the baseline stage addresses small-sample bias by reverting uncertain cases to more conservative estimates and enabling confidence intervals. A subsequent multi-round ridge regression on possession-level residuals refines each player's final rating according to real-time on-court performance.

Empirically, this model promises improved recognition of undervalued skill sets (e.g., off-ball screen assists, forced turnovers, and high-efficiency synergy play types) that do not inflate a typical box score. By mitigating small-sample volatility and awarding partial credit for intangible or team-oriented contributions, the proposed method offers a fresh avenue in the pursuit of a more holistic basketball analytics framework. Future research may extend Bayesian hierarchical structures to further account for rookies' transitions or player aging curves.

References

[1] C. Anella and P. Stone. Box Plus/Minus: A Box-Score Estimate of NBA Player Performance. *Journal of Quantitative Analysis in Sports*, 2016.

[2] J. Sill. Improved NBA adjusted +/- using regularization and out-of-sample testing. *Proceedings of the 5th MIT Sloan Sports Analytics Conference*, 2010.

[3] N. Winston. A Bayesian Hierarchical Model for Adjusted Plus-Minus in Basketball. *Journal of Sports Analytics*, 2014.

[4] D. Guest. Estimated Plus/Minus: A Modern Approach to Basketball Player Evaluation. *Public Analytics Blog*, 2020.