From Prior Beliefs to Lineup Truths: Bayesian Inference for Lineup Performance

Christos Petridis
Temple University
Konstantinos Pelechrinis
kpele@pitt.edu
University of Pittsburgh
Zoran Obradovic
Temple University

Article

Keywords:

Posted Date: February 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8714352/v1

Additional Declarations: No competing interests reported.

From Prior Beliefs to Lineup Truths: Bayesian Inference for Lineup Performance

Christos Petridis[1], Konstantinos Pelechrinis[2], and Zoran Obradovic[1]

[1]Temple University, Department of Computer and Information Science, Philadelphia, 19122, USA, {christos.petridis, zoran.obradovic}@temple.edu
[2,*]University of Pittsburgh, Department of Informatics and Networked Systems, Pittsburgh, 15260, USA, kpele@pitt.edu
[*]Corresponding author. Email: kpele@pitt.edu

ABSTRACT

One of the key responsibilities of a team's coach is to identify the lineups that provide the best chance of winning a game. Traditional metrics such as offensive and defensive ratings summarize past performance, but they are inherently noisy and not predictive. To address this limitation, we adopt a fully Bayesian approach to estimate the posterior predictive distribution of each lineup's offensive and defensive rating. Specifically, we assume a normal prior for these ratings, while the observed points scored and allowed per possession for each lineup serve as our evidence. Given the normal likelihood and the conjugacy of the normal model, the posterior predictive distribution is also normal, with updated mean and variance reflecting both prior beliefs and observed data. Our out-of-sample evaluations show that forecasts based on the posterior predictive distribution outperform the baseline model considering only the past lineup performance observed. The performance gap also increases as the observation lineup samples get smaller. In addition to the prediction improvements, the proposed Bayesian framework naturally quantifies uncertainty in lineup performance, enabling the generation of different rankings, including probabilistic comparisons that better reflect the inherent variability in basketball outcomes.

Introduction

In the 2011 NBA Finals, the Dallas Mavericks made two pivotal decisions that ultimately turned the series in their favor. Throughout the earlier playoff rounds, the Mavericks had rested their superstar, Dirk Nowitzki, by substituting Peja Stojakovic in his place. However, this lineup proved less effective during the Finals, leading the coaching staff to adjust by replacing Stojakovic with Brian Cardinal in Nowitzki's minutes. The second crucial adjustment was inserting J.J. Barea into the starting lineup. Without getting into the tactical reasons these changes succeeded (interested readers can refer to[1]), it is evident that lineup configuration plays a central role in determining a team's success. Indeed, lineup choices can spell the difference between winning an NBA championship and spending the offseason haunted by "what if" questions.

In this study, we introduce a fully Bayesian framework that provides both (a) predictive estimates of a lineup's offensive and defensive ratings and (b) an explicit quantification of the uncertainty associated with its performance. While numerous player rating systems have been developed and used in the public sports analytics community, there has been little publicly available research focused on evaluating lineups. Player ratings are typically estimated via regression models in which the independent variables represent the players on the court during a possession or other time segment, and the dependent variable is the number of points scored in that interval. Each player is represented by indicator variables denoting whether they are on offense, defense, or off the court. Early approaches[2] to estimating player impact suffered from multicollinearity, as specific subgroups of players tended to share the floor for most of their minutes. To mitigate this issue and improve predictive performance, later models employed regularized regression techniques that shrink coefficients toward zero[3]. In the Bayesian interpretation, ridge regression corresponds to imposing a normal prior centered at zero on all player effects[4]. However, this assumption may be overly restrictive: different players warrant different prior expectations. For instance, if both LeBron James and a rookie have coefficients shrunk toward zero, the model requires substantially more data to recover the true difference between them. To address this limitation, subsequent models such as ESPN's Real Plus-Minus (RPM)[5] incorporated informative priors derived from box-score statistics, effectively shrinking player effects toward values reflecting prior performance. Extending this idea further, a fully Bayesian regression framework allows the analyst to specify a prior distribution for each player's rating, naturally capturing both player-specific prior beliefs and the uncertainty surrounding them[6].

More recent work by researchers[7,8] and practitioners[9] has explored various approaches to evaluating lineups. For example, `LinNet`[7] employs a low-dimensional embedding derived from a lineup matchup network as input to a logistic regression model that estimates the probability of lineup A outperforming lineup B. Two notable limitations of this approach are: (a) the limited interpretability of the learned low-dimensional space, and (b) the absence of uncertainty quantification. Kalman and

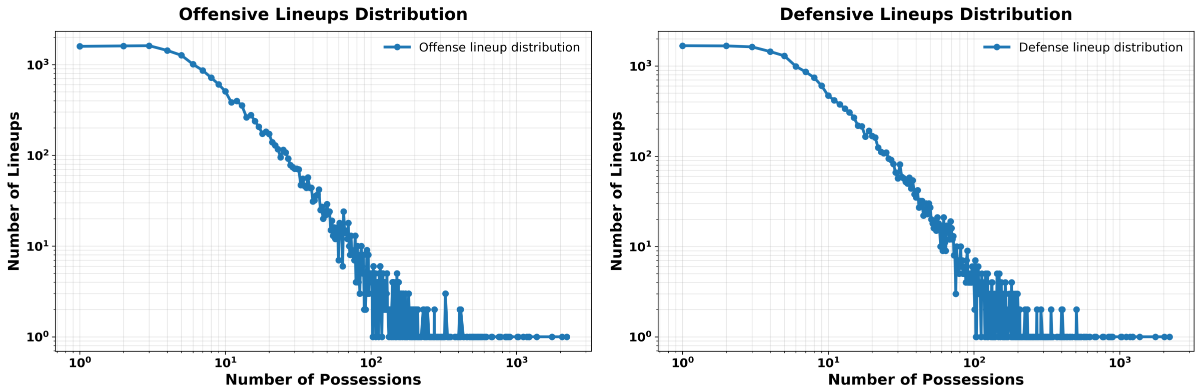

Figure 1. The majority of the lineups during a season appear on the court for a small number of possessions. The distribution (in log-log scale) for the number of offensive (left) and defensive (right) possessions per lineup is right skewed, with a small number of lineups playing for thousands of possessions, while the majority of them plays for less than 50 possessions.

Bosch[8] proposed a clustering-based model that groups players according to basic and advanced basketball statistics to infer their roles on the court. These inferred roles serve as a data-driven alternative to traditional basketball positions and are then used to model a lineup's net rating, defined as the difference between points scored and allowed per 100 possessions. The proposed model improves upon `LinNet` by offering greater interpretability in terms of how player composition and skill sets contribute to success. The work most closely related to ours is the Bayesian-adjusted rating approach proposed by Pelechrinis[9] (subsequently adopted by Kalman and Bosch[8]), in which a lineup's raw offensive, defensive, or net rating is adjusted via a Bayesian average. However, this method produces only a single point estimate for the adjusted rating, without quantifying the uncertainty in lineup performance, an issue that becomes particularly salient when a lineup has limited on-court data.

In our work, we start by considering a league-average prior belief for a lineup's offensive and defensive abilities. Then, using the observed data from the lineup's possessions, we update this belief to obtain the posterior predictive distribution for each component separately. This posterior predictive distribution enables us not only to generate forecasts for future matchups, and rank lineups based on the posterior predictive mean, but also to produce a probabilistic comparison of lineups that accounts for both their expected performance and the associated uncertainty. We evaluate the predictive accuracy of our framework against a frequentist baseline that relies solely on historical lineup data. Our findings indicate that incorporating Bayesian inference yields substantial improvements in predictive performance. More specifically, the posterior predictive distribution improves the out-of-sample predictive performance, as captured by an absolute relative error metric, by approximately a factor of x3.5, with the exact improvement depending on the observation sample size, while the defensive performance predictions enjoy slightly higher improvement. The benefit is higher for lineups with a small number of possessions, which is also the norm for the vast majority of the lineups. Figure 1 shows the distribution of the number of possessions played per lineup on offense (left) and defense (right). As we can see, this distribution is right skewed, with the vast majority of the lineups playing less than 50 total possessions. For these lineups the raw offensive and defensive ratings will be extremely noisy and hence, not reliable predictors of future lineup performance. Figure 2 presents for a single lineup the sequential updates we obtain for our posterior mean and for the raw average points per possession scored for the lineup. As we can see, and as one might have expected, our Bayesian approach stabilizes much quicker. Even more importantly, most of the lineups reside in the grayed out area (zoomed in at the inset figure). As we can see, the raw ratings for the lineups in that area are still very noisy and consequently any predictions or decisions made using these numbers will not be robust as well.

The main contributions of our study are as follows:

- To the best of our knowledge, we introduce the first publicly available framework for lineup evaluation that **explicitly incorporates the uncertainty** around each lineup's expected ability.
- We conduct a detailed empirical evaluation comparing our Bayesian framework to baseline predictions derived exclusively from historically observed lineup data.
- We demonstrate how the posterior predictive distributions for a lineup's offensive and defensive ratings can be used to generate more reliable rankings for assessing relative lineup performance, as well as, probabilistic comparisons between them.

In the following sections we are going to present in detail our setting and the Bayesian framework we are using. We will also provide a detailed evaluation of the predictive power of the framework and finally, showcase its application on probabilistic ranking.

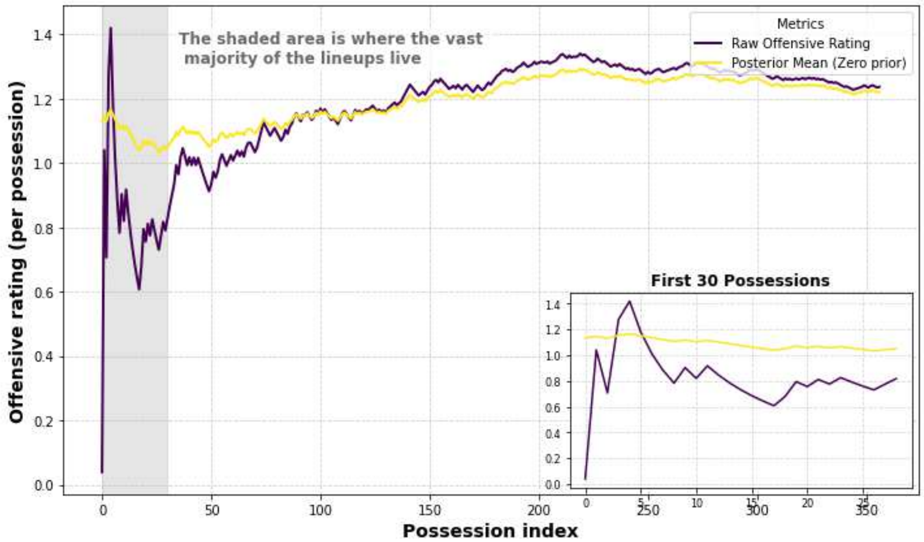

Figure 2. Raw versus Bayesian-adjusted offensive lineup ratings. The raw estimate exhibits substantial variability due to limited sample size, while the Bayesian estimate shrinks the lineup rating toward the league mean, reducing noise and yielding a more stable assessment of lineup performance.

Methods

Our objective is to estimate lineup ratings - both offensive and defensive - that exhibit stronger predictive performance than the official ratings reported by the league. The latter are computed as the number of points scored (or allowed) per 100 possessions by each lineup. We model the lineup ratings within a Bayesian Normal-Normal framework. For each lineup l, we specify a prior distribution for its offensive (and defensive) rating, denoted by $p_{prior,o} \sim \mathcal{N}(\mu_o, \tau_o^2)$ and $p_{prior,d} \sim \mathcal{N}(\mu_d, \tau_d^2)$, respectively.[1] This prior represents our best estimate of lineup l's (offensive or defensive) rating before observing any data. The prior can be subjective or informed by historical data such as player ratings from previous seasons.

After observing possession-level data y_i for lineup l, that is, the number of points scored during possession i (with n total possessions) by lineup l (or the number of points allowed by l if we are modeling the defensive rating), we update our beliefs and obtain the posterior distribution $f_{posterior,o}$ (or $f_{posterior,d}$), corresponding to the offensive (or defensive) rating of the lineup. Taking advantage of conjugacy, we do not need to sample the posterior through MCMC, but the posterior distribution for each lineup remains Normal, with the following posterior mean μ_n and standard deviation τ_n[10]:

$$\boxed{\mu_n = \frac{\dfrac{n\bar{y}}{\sigma^2} + \dfrac{\mu_0}{\tau_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\tau_0^2}}} \quad \boxed{\tau_n = \sqrt{\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1}}}$$

where σ^2 represents the known observation variance, i.e., the variance of the data-generating process (the likelihood). In practice, we use the empirical variance of the observed possession-level outcomes across all lineups, $\widehat{\sigma}^2 = \mathrm{Var}(y)$, as a plug-in estimate for the observation variance σ^2. This common variance assumption simplifies the comparison across lineups and yields an empirical-Bayes version of the Normal–Normal model.

Like any Bayesian framework, our model, visualized in Figure 3, combines the data observations (likelihood) with prior beliefs, to update and obtain the posterior beliefs for the lineup ratings. The prior acts as a regularizer, which is particularly important when we have unreliable or small samples. The ratings that are reported on the league's website only consider the observations, and can be thought of as a Maximum Likelihood Estimation, instead of a Maximum A Posteriori estimation that we perform by considering a prior.

[1]For simplicity, we omit the subscript for the lineup l in the notation. Each model is, however, specific to a particular lineup.

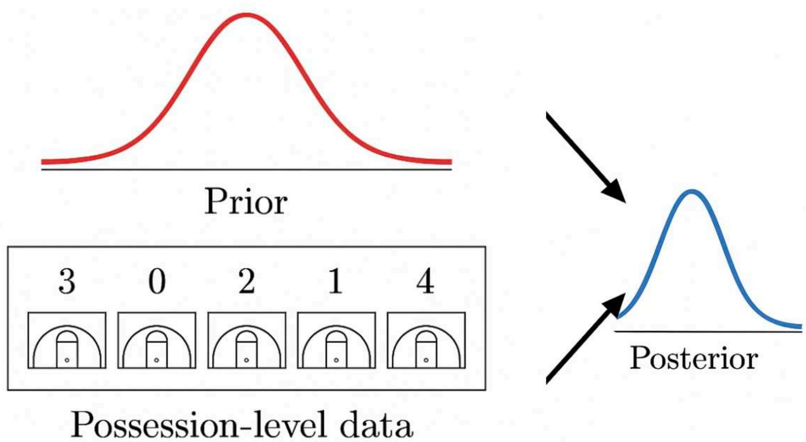

Figure 3. Bayesian updating of a lineup's offensive rating. A league-average prior distribution is combined with possession-level scoring data (the likelihood) to produce a posterior distribution for the lineup's offensive rating. The posterior balances prior information with observed performance, shrinking noisy maximum-likelihood estimates toward the league mean when sample sizes are small.

The choice of the prior influences the resulting posterior, particularly in settings where the number of observations for a given lineup is small, a situation that applies to the majority of lineups. There are several reasonable specifications for the prior means μ_o and μ_d. One of them is the sum of the offensive player ratings (and respectively defensive player ratings) for the players included in the lineup. Alternatively, one could adopt a more agnostic approach by centering the prior means to the league average points per 100 possessions scored, thereby treating each lineup as *a priori* average both on offense and defense. This is the choice we make in this work. While we later discuss some options for player ratings that can further improve lineup performance forecast, our objective in this study is to evaluate the Bayesian framework we introduce and not identify accurate player ratings. This drives us to treating every lineup as a league-average one before seeing any observations.

The second key aspect of the prior is its variance, denoted by τ_o^2 and τ_d^2, which controls how strongly the prior information influences the posterior. Small prior variances correspond to *strong* priors that exert substantial influence when the sample size n is small, whereas large prior variances correspond to *weak* or *diffuse* priors that allow the data to dominate even when the sample is small. These different prior specifications can lead to substantially different posterior inferences, especially for lineups with limited possession-level data.

Given that our main objective is lineup performance forecast we will focus our evaluation on out-of-sample prediction. However, one of the practical issues with this evaluation is that we do not know the *true* ability of a lineup. Therefore, what will we compare our predictions against? To overcome this limitation, we will focus our evaluation on lineups for which we have observed more than 300 possessions (on each offense and defense). This way we can treat the final observed offensive/defensive ratings $r_{o,l}$ and $r_{d,l}$ for a lineup l as the *ground truth*, while at the same time calculating the posterior over different windows of observations (i.e., using the first ν number of possessions). With $\mathbb{E}[f_{posterior,o}(\nu)]$ (and $\mathbb{E}[f_{posterior,d}(\nu)]$) being the posterior mean for offense (and defense) after ν observations, we will calculate the relative absolute error in our predictions for different values of ν for offense and defense respectively:

$$\begin{aligned}\varepsilon_{r,o} &= \frac{\mathbb{E}[f_{posterior,o}(\nu)] - r_{o,l}}{r_{o,l}}, \\ \varepsilon_{r,d} &= \frac{\mathbb{E}[f_{posterior,o}(\nu)] - r_{d,l}}{r_{d,l}}\end{aligned} \tag{1}$$

While, as discussed in the following section, we initially consider a discrete set of candidate priors for evaluation, it quickly becomes evident that predictive performance depends not only on the choice of the prior but also on the amount of available evidence. In particular, two lineups with different numbers of observed possessions ν may warrant different degrees of regularization. This observation motivates moving beyond a single, fixed prior specification for all lineups. Instead, we introduce a model that learns the appropriate prior standard deviation as a function of the number of observed possessions, allowing the strength of the prior to adapt continuously to the available data.

To formalize this idea, we consider a Normal–Normal hierarchical model at the lineup level in which the strength of the prior depends explicitly on the amount of observed data. Let $y_{l,i}$ denote the outcome of possession i for lineup l, with ν_l total possession and sample mean $\bar{y}_l$. We assume:

$$y_{l,i|\mu_l} \sim \mathcal{N}(\mu_l, \sigma_y^2), \tag{2}$$

where μ_l represents the latent offensive (or defensive) rating of lineup l. Prior to observing data, we place a Normal prior on μ_l:

$$\mu_l \sim \mathcal{N}(\mu_{0,l}, \tau_l^2), \tag{3}$$

with prior mean $\mu_{0,l}$ and prior standard deviation τ_l^2. Rather than fixing τ_l across all lineups, we allow it to vary as a smooth function of the number of observations ν_l. Specifically, we parametrize the prior standard deviation using a log-log regression:

$$\log \tau_l = \alpha + \beta \cdot \log(\nu_l) \tag{4}$$

which ensures positivity and enforces a monotonic dependence of prior strength on sample size, with the direction and rate of change governed by β. This formulation captures the intuition, and the evaluation results, that the degree regularization should decrease as more evidence become available. Integrating out the latent parameter μ_l yields a closed-form marginal distribution for the observed lineup mean:

$$\bar{y}_l \sim \mathcal{N}(\mu_{0,l}, \tau_l^2 + \frac{\sigma_y^2}{\nu_l}), \tag{5}$$

which serves as the basis for estimating the regression parameters α and β. These parameters are learned by maximizing the marginal likelihood $\mathcal{L}(\alpha, \beta)$ across lineups. Assuming independence:

$$\boxed{\mathcal{L}(\alpha, \beta) = \prod_{l=1}^{L} \mathcal{N}\left(\bar{y}_l \;\middle|\; \mu_{0,l},\ \tau_l^2(\alpha, \beta) + \frac{\sigma_y^2}{\nu_l}\right)} \tag{6}$$

The marginal likelihood evaluates how plausible the observed lineup sample means are under a model in which prior dispersion increases smoothly with sample size, allowing us to learn the rate at which lineup abilities are permitted to deviate from the league average. We also use a weighted objective, to emphasize the calibration in the small-sample regime. Once estimated, the resulting function $\tau(\nu)$ defines a sample-size-adaptive prior that is subsequently used for posterior predictive inference.

Someone here might argue that as the amount of evidence for a lineup increases, the effect of the prior reduces, and hence, we should only worry about lineups with small sample sizes. While this is exactly true, the prior variance determines the amount of pooling across lineups at finite samples. While the posterior converges to the likelihood as the number of observations increases, the rate at which this occurs is regulated by the prior standard deviation, which we learn as a function of sample size. So the proposed model adaptively modulates the strength of prior regularization based on the amount of available efixence, stabilizing the small-sample estimates while preserving standard Bayesian asymptotic behavior.

Results

In this section we will present our detailed evaluations. We use data from the 2023-24 NBA season. Our data consist of 271,419 total possessions and 16,455 unique lineups. For each possession, we have information for the 5 players on offense, the 5 players on defense, and the points scored during this possession. Table 1 shows some basic statistics from our dataset. These descriptive stats make it very clear, that in the majority of the cases, we have to work with very small sample sizes for lineups and thus, a more robust approach is needed when dealing with them. As aforementioned we focus on lineups that have at least 300 possessions on each side of the ball, which gives us a set of 73 lineups in total. For each one of these lineups we restrict the available observations we use for calculating the posterior to the first n, for various values of n, and we use bins of 10 possessions. We consider several different normal priors in our evaluations, all with a mean of the league-average points scored/allowed per possession. The standard deviation of these priors range from very strong to diffuse taking values of $0.01s$, $0.02s$, $0.03s$, $0.04s$, $0.05s$, $0.1s$, $0.2s$, $0.5s$, $1s$, $1.5s$, $2s$, $5s$ and $10s$, where s is the standard deviation of the points scored over all the possessions in the data. While we present all of our results in the Supplementary Material, Figure 4 (notice the logarithmic scale on x-axis) presents the results for 3 representative prior strengths of $0.01s$ (strong), $0.1s$ (informative) and $2s$ (weak). In particular, we present the relative absolute error ε_r (Equation 1) as a function of the number of observations/possessions for each lineup.

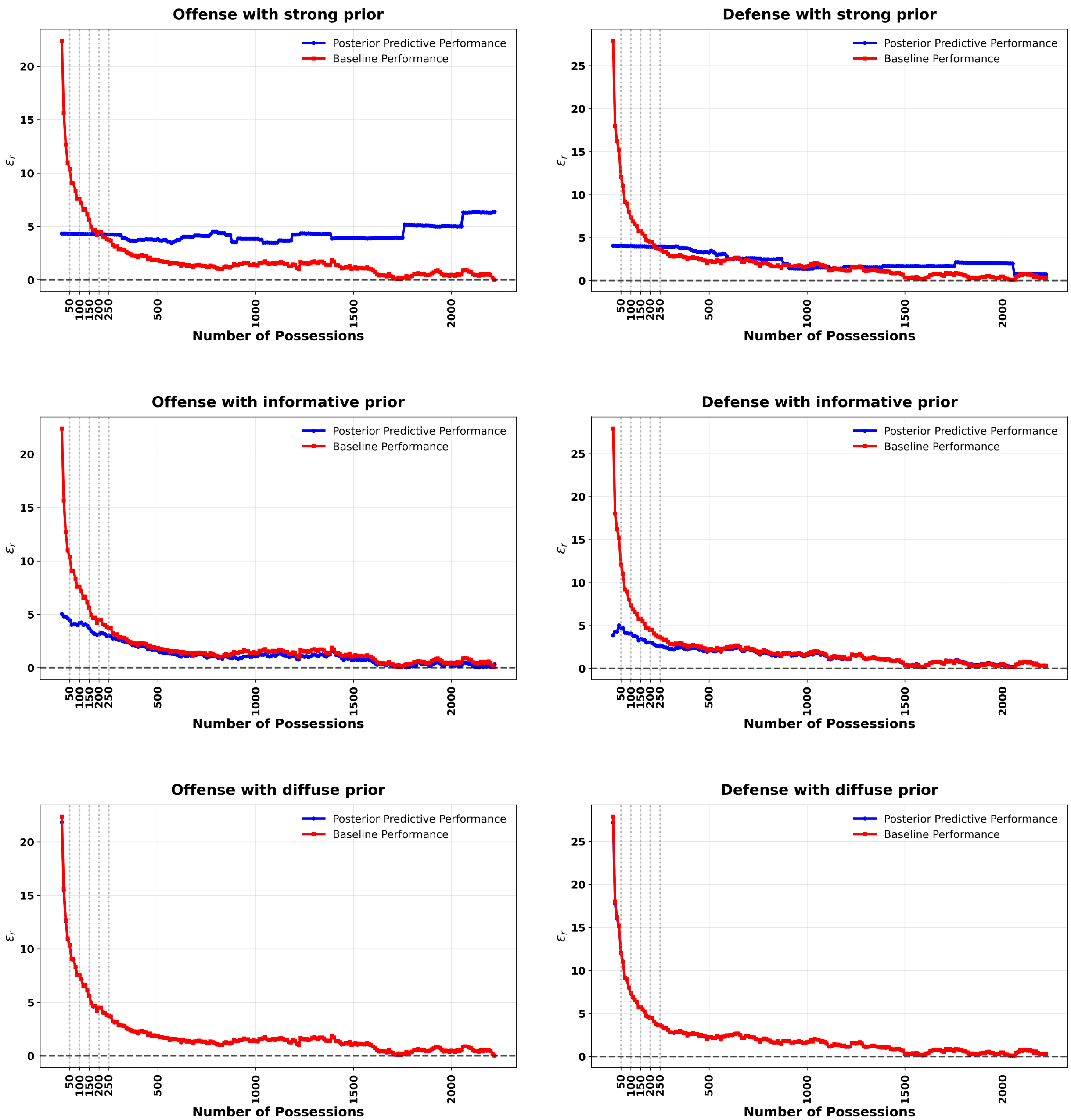

Figure 4. Informative priors can significantly improve the lineup performance forecast. When we use a very strong prior ($0.01s$ - top row), we essentially predict the league average regardless of the amount of data we have for the lineup. This results in significant error. A weak/diffuse prior ($2s$ - bottom row) on the other hand, will lead to a posterior dominated exclusively from the data regardless of the sample size. An informative prior ($0.1s$ - middle row) will provide the *right* amount of regularization and provide significant forecast improvement when the sample sizes are small.

When a very strong prior is used, it effectively dominates the posterior even when a fairly large amount of data is available. As a result, the posterior mean remains very close to the prior mean regardless of the amount of observations available. This behavior is expected from the conjugate normal–normal model, where the posterior mean is a precision-weighted average of

Average possessions per lineup	16.5
Median possessions per lineup	6.0
90^{th} percentile of possessions per lineup	32.0
Average points scored per possession (ppp)	1.138
Standard deviation of ppp	1.19

Table 1. Basic statistics from our lineup dataset. Half of the lineups have less than 6 possessions in our data, while only for 10% of the lineups we have more than 32 observations.

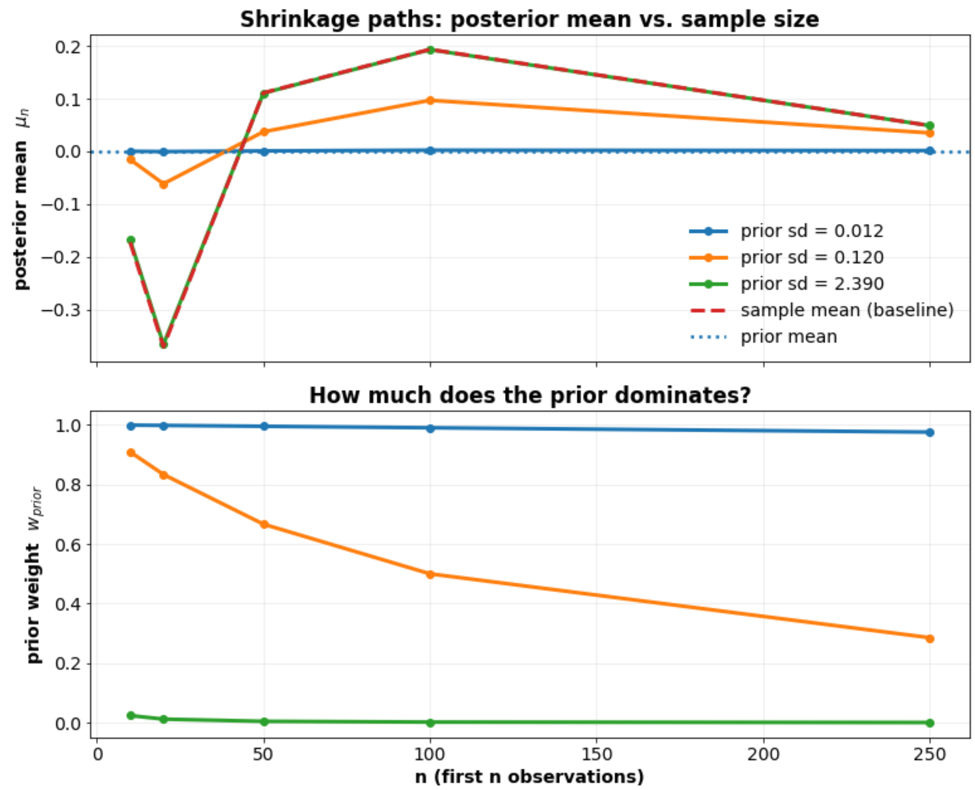

Figure 5. Shrinkage behavior of the posterior mean under different prior strengths. **Top panel:** Posterior mean as a function of sample size for three priors, compared to the prior mean and the sample mean. Strong priors induce persistent shrinkage, while diffuse priors quickly converge to the baseline estimator. **Bottom panel:** Posterior weight on the prior mean as a function of sample size. This weight governs the shrinkage behavior observed in the top panel, with stronger priors retaining influence longer and weaker priors deferring rapidly to the data.

the prior mean and the sample mean. In this case the forecasts from the posterior predictive distribution will basically represent the predictions of the prior and the error will be relatively flat as a function of the number of possessions used. When the prior variance is only 1% of the observation variance, an extremely large number of observations is required before the data meaningfully influence the posterior. For small sample sizes, such a strong prior improves the predictive performance in our case, but this improvement does not stem from more accurate inference. Instead, it arises from the model's reluctance to react to noise in limited data. As the sample size grows, however, the posterior continues to underweight genuinely informative data. Consequently, its predictive performance deteriorates and can become significantly worse than that of even a simple baseline model. At the opposite extreme, using a highly diffuse prior allows the data to dominate the posterior, even when the observations largely reflect noise (as is often the case with small sample sizes). As a result, the posterior effectively ignores the prior and closely tracks the sample mean, which coincides with our baseline estimator. Consequently, in this setting the framework offers little to no improvement over the baseline in terms of forecast ability.

When we use a prior in between these 2 extremes, we allow the Bayesian framework to strike the right balance between our prior belief and the observations. More specifically, when we have a small amount of observations, the prior still heavily regularizes the data likelihood, but allows the observations to gradually contribute more information in the posterior calculation. This results in much improved predictive performance especially for sample sizes smaller than 50 observations, which is the case for more than 95% of the lineups!

Figure 5 further illustrates how prior strength and sample size jointly determine the behavior of the posterior mean in the normal–normal model. The data correspond to a randomly selected lineup from those with at least 300 possessions. The plot at the top shows the posterior mean for the offense as a function of the number of observations for three different priors, alongside the sample mean and the prior mean. For simplicity, we have expressed the points scored as above/below the league average, so the prior is set to 0 (i.e., league average). Strong priors induce substantial shrinkage toward the prior mean, causing the posterior mean to move slowly - if at all - even as additional data are observed. In contrast, diffuse priors allow the posterior mean to rapidly align with the sample mean, providing little regularization beyond the baseline estimator. Note also that the informative prior is the one that gives a posterior mean that is *closer* to the final observed rating (that is, the final point of the sample mean line), evne when the amount of data is small (e.g., 50 possessions). The curve for the informative prior (standard deviation of 0.12) tracks the one for the sample mean pretty well, but it also provides the appropriate shrinkage.

The bottom panel explains this behavior by plotting the posterior weight assigned to the prior mean as the sample size increases. As aforementioned, in the normal-normal model, the posterior mean is a precision-weighted average of the prior mean and the sample mean. The posterior weight on the prior quantifies how strongly the prior continues to influence inference. From the conjugate formula we saw earlier, we can obtain the prior weight as:

$$w_{prior} = \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2} \tag{7}$$

Strong priors retain a high weight for a large number of observations, whereas diffuse priors relinquish influence quickly. The point at which the posterior's weight on the prior becomes negligible corresponds directly to the point at which the posterior mean in the top panel begins to closely track the sample mean.

The shrinkage paths presented suggest that the central modeling choice is not the prior mean itself, but the rate at which prior influence should diminish as data accumulate. This raises a natural question: given a specific amount of evidence, how strong should the prior be? As we elaborated in the previous section, instead of choosing a single prior among a finite number of options, we learn a regression model that estimates the prior variance as a function of sample size. This approach allows the model to adaptively calibrate prior strength, balancing robustness to small-sample noise against responsiveness to informative data. Figure 6 shows the models fitted for offense and defense. Under our parameterization $\log \tau(n) = \alpha + \beta \cdot \log n$, the offensive curve increases rapidly with n ($\alpha_{off} = -4.9$, $\beta_{off} = 0.35$), whereas the defensive curve is nearly flat ($\alpha_{off} = -3.3$, $\beta_{off} = 0.02$). This implies markedly different stabilization dynamics across the two sides of the ball.

For offense, $\tau_o(n)$ is very small at low n ($\alpha_{off} = -4.9$), which corresponds to strong pooling toward the prior mean when evidence is scarce. As n increases, $\tau_o(n)$ grows quickly, allowing offensive lineup ratings to separate from each another and become increasingly data-driven. In contrast, defense exhibits a much larger baseline prior dispersion ($\alpha_{def} = -3.3$) but minimal growth with additional evidence. Consequently, defensive lineup abilities remain more tightly clustered, even at moderate to large n, and continue to require relatively stronger regularization. The fitted models imply that the rate at which prior influence decays is substantially faster for offense than for defense. This aligns with the empirical observation that offensive lineup performance becomes lineup-specific more quickly, while defensive performance remains more context-dependent and noisy, which warrants sustained shrinkage.

While the error related to the lineup ratings per se are certainly important and of interest, equally important is the ranking we obtain from each lineup. In some cases, we are not interested about the exact performance of the lineup but rather on whether lineup A is *better* than lineup B. Even if our Bayesian ratings are more accurate when it comes to predicting the actual value of offensive and defensive ratings, they might still provide the same ranking with the raw ratings. Figure 7 shows the rank correlation between raw and Bayesian lineup ratings as a function of the minimum number of observed possessions. As expected, rank agreement increases rapidly with sample size, particularly for offense, indicating that Bayesian shrinkage primarily affects sparsely observed lineups while leaving rankings among well-observed lineups largely unchanged. Defensive rankings converge more slowly, consistent with greater uncertainty and sustained regularization on the defensive side of the ball. Given our earlier results demonstrating improved predictive accuracy of the proposed model, these differences in rankings are expected to reflect more reliable assessments of relative lineup performance.

Although ranking lineups by their posterior predictive means provides a deterministic ordering, indicating whether lineup A or lineup B is better, the availability of the full posterior predictive distribution allows us to further quantify the uncertainty in this comparison by estimating the probability that lineup A outperforms lineup B. For example, in our data there are two lineups with posterior predictive means for their offensive ratings of $\mu_A = 0.124$ and and $\mu_B = 0.032$, and corresponding variances of $\sigma_A^2 = 1.41$ and $\sigma_B^2 = 1.43$ respectively (means are again expressed as points per possession above/below league average). Figure 8 presents the two posterior predictive distributions (left) and the distribution of their difference (right). Based on posterior predictive means alone, lineup B is estimated to be 0.092 points per possession worse on offense than lineup A, corresponding to 9.2 points per 100 possessions. Taken at face value, this difference is comparable to the gap between a top-tier offense and a lower-tier offense in the league. However, this interpretation ignores the substantial uncertainty associated with each lineup's estimated performance. By explicitly accounting for uncertainty, we can compute the probability that lineup B actually outperforms lineup A:

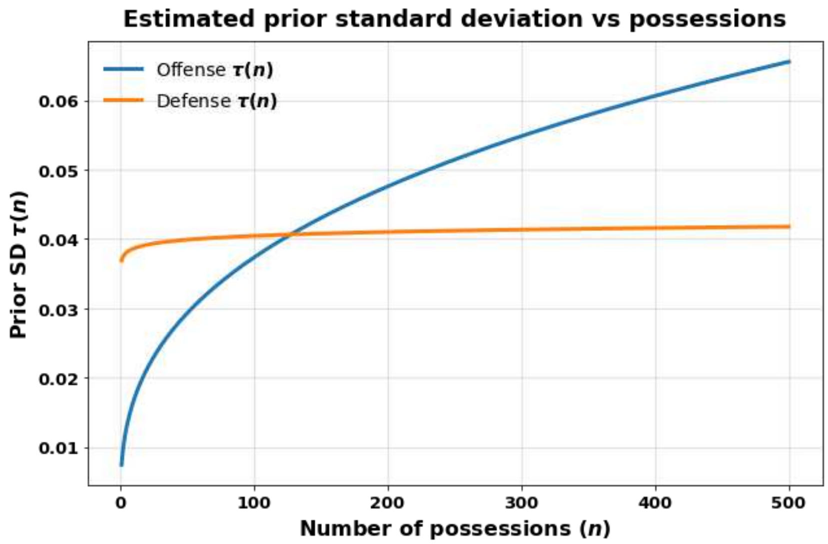

Figure 6. Prior heterogeneity as a function of possessions played. Estimated prior standard deviations $\tau(n)$ for offensive and defensive lineup ratings highlight distinct stabilization dynamics, with offensive performance separating more rapidly across lineups than defensive performance.

$$\Pr[B > A] = \Pr[B - A > 0] = 1 - \Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) \tag{8}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Thus, while posterior means alone suggest that lineup A is better than lineup B, incorporating posterior predictive uncertainty reveals that the two lineups are much closer in performance than a point estimate would indicate. This ability to produce uncertainty-aware, probabilistic comparisons is a key advantage of the proposed framework relative to existing lineup rating approaches[7,9].

Discussion

In this study, we introduced a Bayesian framework for estimating posterior predictive distributions of the offensive and defensive ratings of basketball lineups. Our approach is based on a Normal–Normal conjugate model, in which lineup ratings are assigned a Normal prior and updated using possession-level data. Traditionally, lineup evaluation relies on raw offensive and defensive ratings, defined as points scored and allowed per 100 possessions, and these quantities are often treated as forward-looking performance estimates. Our results demonstrate that this practice can be highly misleading, particularly for lineups with a limited number of possessions. By contrast, Bayesian shrinkage yields more stable and better-calibrated estimates, especially in the low-sample regime, which applies to the vast majority of observed lineups.

In our empirical analysis, we adopted a league-average prior mean for lineup ratings. However, more informative prior means could further enhance predictive performance. For example, a lineup consisting of five All-Star players should reasonably be expected to outperform a lineup of replacement-level players even before any lineup-specific data are observed. Encoding such information through player-based prior means, rather than assuming all lineups are initially league-average, represents a natural extension of our framework and could further improve predictive accuracy.

Our evaluation also revealed that the strength of the prior plays a critical role in predictive performance and that its optimal choice depends on the amount of data available for each lineup. Rather than selecting a single prior specification for all lineups, we introduced a simple regression-based approach that adapts the prior variance as a function of the number of observed possessions. This allows the model to apply stronger regularization when data are scarce and to relax prior influence as evidence accumulates. As an alternative, one may specify a desired level of prior influence on the posterior, expressed through the effective prior weight w_{prior}, and then determine the corresponding prior variance to be used via the analytical relationship governing the posterior shrinkage. Both approaches provide principled and flexible mechanisms for defining the prior strength of our model.

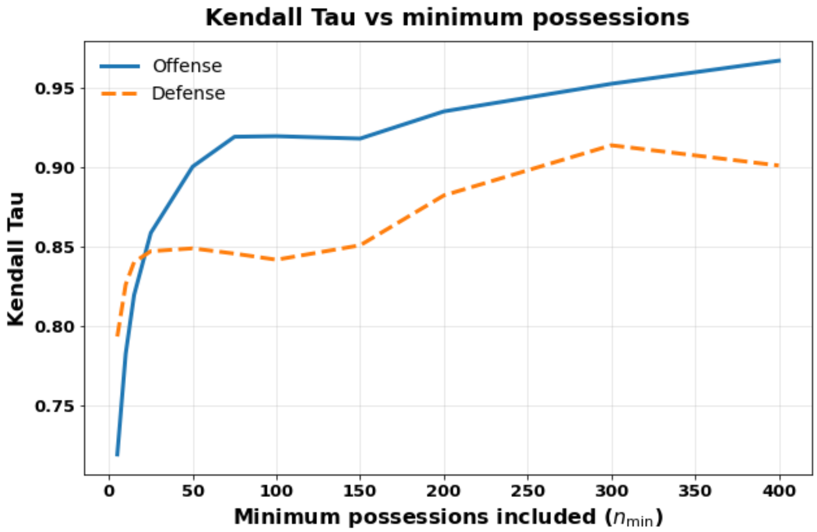

Figure 7. Rank correlation between raw and Bayesian lineup ratings as a function of sample size. Kendall's τ between raw (data-only) and Bayesian posterior mean rankings is shown as a function of the minimum number of possessions included, separately for offense and defense. Rank agreement increases rapidly with sample size, indicating that Bayesian shrinkage primarily affects sparsely observed lineups, while rankings among well-observed lineups remain largely unchanged.

Part of the noise for lineups that have played a small number of possessions comes from the opponent strength. When a lineup has been on the court only for 10 possessions, there will be 2 main sources of noise: (a) the possession to possession variance that we have incorporated in our model through the data likelihood, and, (b) the opponent strength. In a small sample size the strength of competition faces might be very different than the league average. One could adjust the points scored during a possession taking into account the player ratings of the opposing lineup[1]. For example, if a team gave up 0 points against an elite offensive unit, the adjusted points allowed should be < 0. If they gave up 3 points to a poor offensive unit, the points allowed for that possessions should be > 3. This adjustment can use the same player ratings mentioned above. Another option would be to learn the posterior distribution for the ratings of a lineup through a Bayesian regression, similar to the one used for learning player ratings[11]. This will automatically adjusts for opponent strength.

Finally, we demonstrate that posterior predictive means provide a more reliable basis for comparing lineup performance than raw ratings. Crucially, access to the full posterior predictive distributions allows us to formalize *pairwise lineup dominance* through the probability that lineup A is superior to lineup B.

References

1. Winston, W. L. *Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football* (Princeton University Press, 2012).
2. Illardi, S. Adjusted plus-minus: An idea whose time has come (2007). https://www.82games.com/ilardi1.htm.
3. Sill, J. Improved nba adjusted+/-using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT sloan sports analytics conference* (2010).
4. Hastie, T., Tibshirani, R., Friedman, J. *et al. The elements of statistical learning* (Springer series in statistics New-York, 2009).
5. Engelman, S. The next big thing: real plus-minus (2014). https://www.espn.com/nba/story/_/id/10740818/introducing-real-plus-minus.
6. Deshpande, S. K. & Jensen, S. T. Estimating an nba player's impact on his team's chances of winning. *J. Quant. Analysis Sports* **12**, 51–72 (2016).
7. Pelechrinis, K. Linnet: Probabilistic lineup evaluation through network embedding. In *Joint European conference on machine learning and knowledge discovery in databases*, 20–36 (Springer, 2018).
8. Kalman, S. & Bosch, J. Nba lineup analysis on clustered player tendencies: A new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates. In *14th annual MIT sloan sports analytics conference* (2020).

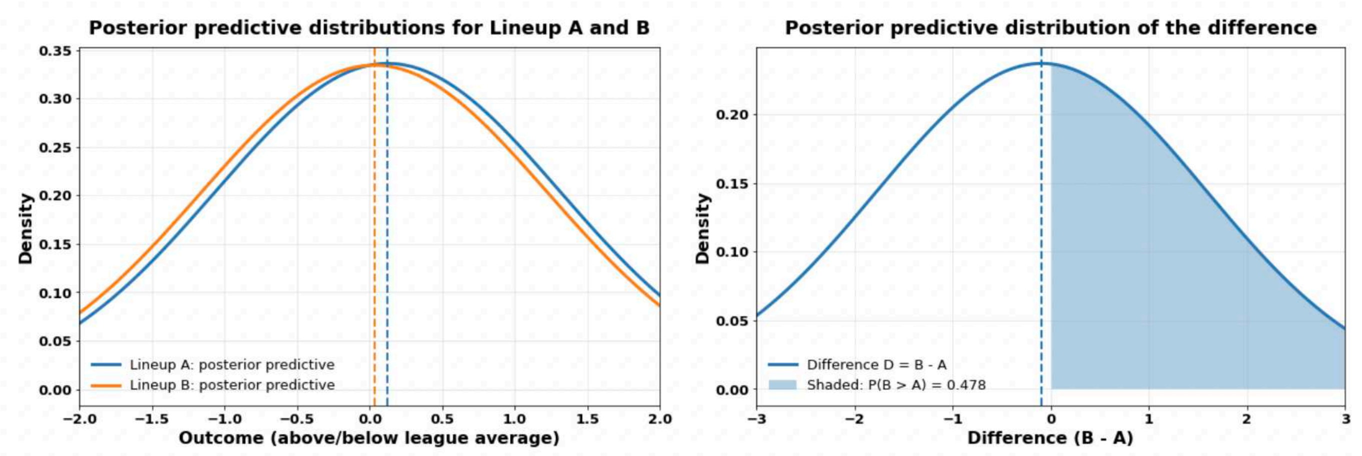

Figure 8. Posterior predictive comparison of two lineups. *Left:* Posterior predictive distributions for Lineup A and Lineup B, illustrating substantial overlap despite different posterior means. *Right:* Posterior predictive distribution of the difference $B-A$, with the shaded region corresponding to $\Pr(B>A)$.

9. Pelechrinis, K. Lineup evaluation through win-probability models and bayesian adjustment. In *Cascadia Symposium on Statistics in Sports* (2018).

10. Murphy, K. P. Conjugate bayesian analysis of the gaussian distribution. *def* **1**, 16 (2007).

11. Petridis, C. & Pelechrinis, K. Lineup regularized adjusted plus-minus (l-rapm): Basketball lineup ratings with informed priors. *arXiv preprint arXiv:2601.15000* (2026).

Author contributions statement

K.P. conceived of the presented idea. All the authors developed the theory and performed the empirical analysis, including visualizations. K.P. and C.P analyzed the results. All the authors contributed to the presentation of the results and the final manuscript.

Competing interests

The authors do not have to declare any competing interests.

Funding Information

Dr. Pelechrinis' work was partially supported by a grant from the Mark Cuban Foundation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SIBLR.pdf