An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data

Luca Grassetti†, Ruggero Bellio†,
Luca Di Gaspero††, Giovanni Fonseca†,
Paolo Vidoni†*

†Department of Economics and Statistics,
University of Udine, Italy
Via Tomadini, 30/A – 33100 Udine (UD).

††Polytechnic Department of Engineering and Architecture,
University of Udine, Italy
Via delle Scienze, 206 – 33100 Udine (UD).

Abstract

In this work we analyse basketball play-by-play data in order to evaluate the efficiency of different five-man lineups employed by teams. Starting from the adjusted plus-minus framework, we present a model-based strategy for the analysis of the result of partial match outcomes, extending the current literature in two main directions. The first extension replaces the classical response variable (scored points) with a comprehensive score that combines a set of box score statistics. This allows various aspects of the game to be separated. The second extension focuses on entire lineups rather than individual players, using a suitable extended model specification. The model fitting procedure is Bayesian and provides the necessary regularisation. An advantage of this approach is the use of posterior distributions to rank players and lineups, providing an effective tool for team managers. For the empirical analysis, we use the 2018/2019 regular season of the Turkish Airlines Euroleague Championship, with play-by-play and box scores for 240 matches, which are made available by the league website. The results of the model fitting can be used for several investigations as, for instance, the comparative analysis of the effects of single players and the estimation of total and synergic effects of lineups monitoring. Moreover, the behaviour of players and lineups during the season, updating the estimation results after each gameday, can represent a rather useful tool. basketball analytics, statistical model, play-by-play data, web-crawling, data-driven decision process.

1 Introduction

In every system that employs human capital, each participant has an assigned task to complete and the combination of these tasks contributes to the final result. One of the main goals in this framework is to obtain the best possible outcome more efficiently. To improve the work process, for example, van de Water and Bukman (2010) propose a method to obtain a balanced team, that is, a team where each member performs a specific task which maximises the expected output. Team sports competitions are clear examples of this situation, and therefore the analysis of players' efficiency is widely studied

*Corresponding author: paolo.vidoni@uniud.it

(see, for example, Carmichael, Thomas, and Ward, 2001; Hollinger, 2005; Hughes and Bartlett, 2002; Yesilyurt, 2014).

From an external viewpoint, the aim is typically the prediction of team performances and there is an extensive literature on final ranking or on single match outcome predictions; see Haigh (2009) for a general introduction to quantitative methods in sports and Karlis and Ntzoufras (2009) for an application to soccer match results. In this respect, the assumption of Poisson distribution for each team match scoring is adopted in many sports, with the resulting implications of this modelling approach. Concerning basketball, Shen (2014) and Ruiz and Perez-Cruz (2015) compute match-winning probabilities under such a model. Manner (2016) introduces time-varying team strengths in the model and Kvam and Sokol (2006) consider a logistic regression model in a Markov chain setup. Machine learning algorithms for classification are also used for predicting match results, as done by Shi et al. (2013) and Yang and Lu (2012). All these results consider the team as a single entity. Instead, Deshpande and Jensen (2016) consider individual player's contributions to match-winning probability at a given time of the game. From an internal perspective, the aim is to determine the best management of players in order to succeed. Single player's performance is described by box scores data that are regularly provided by media. For example, assists, rebounds, points are used to measure the contribution of each player to the final match result. Masoumzadeh et al. (2016) model the outputs of a production system and present an application to basketball players. Sport's literature focuses mainly on individual efficiency measures.

The most straightforward efficiency measure for a single player is based on the points scored by her/his team and by the opponent during the time that specific player is on the pitch. This is the starting point of the Plus/Minus Rating (PM). In the early 2000s, Rosenbaum (2004) introduces a regression-based version of the PM metric, called Adjusted PM (APM), with application to National Basketball Association (NBA) data. His proposal considers the computation of the index based on play-by-play data aggregated over periods of playing time without any substitution for either team, called *shifts*. The APM for each player is defined as the estimated regression coefficient in a linear model where the point differential during each shift (averaged with respect to the number of possessions for each team) is the response variable. The predictors are the signed dummy variables for all the players involved in the shift. The model is defined from the perspective of the home team.

In the same period, several authors (e.g. Kubatko et al., 2007; Ilardi, 2007) propose extensions to the model specification including also other player's game statistics in order to obtain a more comprehensive view of the impact of a player during a match or a whole season. Such APM measures have a noteworthy impact on the field of professional basketball since they represent player's efficiency measures adjusted for the other players on the field. However, these methods are prone to some technical criticism since they entail sparse design matrices and multicollinearity. Therefore, Sill (2010) and Engelmann (2011) propose regularised versions of APM based on ridge regression, called RAPM. The improvement in accuracy and robustness of RAPM led to its use also for the analysis of the performance of players in Major League Soccer (MLS) (Kharrat et al., 2017) and in National Hockey League (NHL) (Macdonald, 2012). For a more exhaustive account on the subject of PM-based indexes see Engelmann (2017) and Hvattum (2019).

Nowadays, play-by-play data are recorded in real-time during the match and made available on the web to practitioners and supporters. This information is used by team managers to build effective lineups and hence the necessity for more specific efficiency measures naturally arises. This work tries to tackle this issue, remaining entirely within the realm of the PM measures, which are defined for the evaluation of performances and not for the prediction of future results. Starting from the RAPM setting, the main target of the paper is to extend the analysis of play-by-play data in order to provide a suitable efficiency measure for the performance of entire lineups, defined as five-player units on the field for a given team. Indeed, the choice of the best lineup cannot take into account just the compound of the single player's capabilities, but should also be based on balancing different aspects of the game, as, for example, the offensive and defensive capabilities of the team as a whole. Therefore a central aspect in evaluating the efficiency of a given lineup is played by the interaction among teammates and by the counteraction of the opponent's lineup on the field. The specification of a suitable model, encompassing contributions of both lineup and player, is used to study these effects. Due to the high dimensionality of the model, regularisation is essential for the estimation task in our proposal, which replaces ridge regression with a more structured method. To this end, both an empirical Bayes approach and a full Bayesian treatment are proposed in what follows; see Gelman et al. (2014); Efron and Hastie (2016) for an introduction.

An innovative feature of our model is the adoption as a response variable of a more comprehensive performance index rating, rather than the simple point differential; in the sequel, such performance index rating is simply called *overall score*. This measure is still based on an aggregation of play-by-play data at the shift level but it combines a wide set of game statistics for each shift. As a side result, our score measure can also be viewed as the composition of four sources, providing further insight on the effective strength of a given lineup: the inside scoring skill (lay-ups, dunks and free shots), the mid-range shooting skill, the three-pointers shooting skill and the set of complementary skills (assists, rebounds, blocks and steals), hereafter called *other skills*.

Finally, a further aim of our proposal is giving to team managers an effective tool for choosing the best lineup depending on opponent's lineup, time of the game, current score of the match, and so on (see Lechner and Gudmundsson, 2012). For team owners and general managers, the proposed approach can be useful to assess the contribution of a head coach to the performance of the team (see Berri et al., 2009; Berry and Fowler, 2019, and references therein). From the players' side, the trust in the coach improves if the management choices are supported by an objective analysis of past performances (see, for example, Kao et al., 2017; Zhang and Chelladurai, 2013). To this end, we propose how to rank players or lineups using the model estimated effects, so as to provide some support to managers.

An empirical application is developed to illustrate the proposed method. In particular, we apply the model to the Euroleague Championship data (regular season 2018/2019). The dataset is obtained by aggregating the play-by-play data for the 240 games of the entire regular season. The raw data are collected for each play of the game, but the final observational unit is the shift.

The paper is organised as follows. In Section 2, we present the data used for the analysis and some details about the data handling process. The model used for the estimation of the lineup's efficiency is introduced in Section 3. In Section 4 we report the results obtained applying our model to Euroleague Championship data. Particular attention is devoted to the managerial implications of the proposed methodology, and Section 5 presents some illustrative examples. Finally, Section 6 contains some concluding remarks.

2 Data wrangling and data exploration

In this section, the data wrangling process is briefly described, and the results of the exploratory data analysis are reported. The data concern all the matches of the 2018/2019 regular season of the Euroleague Championship (240 matches). The league website provides play-by-play data integrated with the box scores and further interesting information.

2.1 Data wrangling

The methodology for getting the data is rather sophisticated. The raw data have been obtained through web data extraction from the Game Center pages of the Euroleague website (`https://www.euroleague.net/`) using the Apache Selenium web automation tool. The major difficulty of the data extraction is due to the fact that those web pages are interactive in nature, making extensive use of modern Asynchronous Communication (i.e., AJAX) to provide the requested information dynamically. For this reason, a complex page interaction is needed to display (and, consequently, retrieve) all the needed information. Specifically, the data in the web pages are extracted only after the page content has been modified and rendered by the browser JavaScript engine. This process is achieved by querying the page dynamic Document Object Model (DOM), that allows to programmatically access and inspect the visual objects in the page. On overall, this automated web scraping task is performed in parallel and all the machinery for this task has been written in Python 3 (Van Rossum and Drake, 2009). The scraped content is collected in the form of 'pandas' dataframes and stored in tabular format.

The R statistical software (R Core Team, 2020) and, in particular the `stringi` (Gagolewski, 2019) and `stringr` packages (Wickham, 2019), have been used for the data reorganisation. Both the box scores data and the play-by-play information are collected for all the considered matches. A play is defined as an event during the possession involving a positive or negative value for the attacking team, and considered as relevant for the result of the game. In particular, as introduced in Grassetti et al. (2019a,b), we compute a new score measure assigned to the offensive team and to the defensive team with opposite signs. The more customary outcome given by the points scored in each play is also gathered. Moreover, other features are collected, such as information concerning the time of the event and game status. Finally, the box score data have been used to identify the five-man unit involved in each play. This information is essential to organise the plays in shifts, as required for the analyses.

2.2 The score variable

The score measure is defined as in Grassetti et al. (2019a). In particular, the scores reported in Table 1 are assigned with opposite signs to the offensive team and the defensive team, respectively. The events considered in the definition of the overall scores are those deemed as important for the outcome of the play.

Table 1: Scores of the events used in the computation of the outcome measure for each play.

Value	Events
-1	missed free-throw, turnover or offensive foul
-0.5	missed shot (two points or three points shots)
0.5	assist
1	steal, offensive or defensive rebound, block, scored free-throw or received foul
2	scored shot
3	scored three-pointer

Some records in the raw data are replicated because some events are reported for both teams. For instance, steals and personal defensive fouls always correspond to turnovers and received fouls, respectively. The duplications have been removed. The resulting score measure is more comprehensive, but it is not too far from the final result of the match.

A further interesting analysis regards the disentanglement of the overall score into four components. As introduced in Section 1, these four aspects are related to the close and mid-range shooting capability, the three-pointers performance and the sum of the other skills entering the overall score computation: turnovers, fouls, assists, steals, rebounds and blocks. Indeed, these other skills strongly affect both the offensive and defensive performances of a team and hence the observed score. Figure 1 displays the relationship between the overall score and its four components. While the linear relationship between these score contributions and the total measure is apparent, the pairwise relationship among the components is weak. This suggests that the different kinds of contributions may actually correspond to different aspects of the player's and lineup's performences.

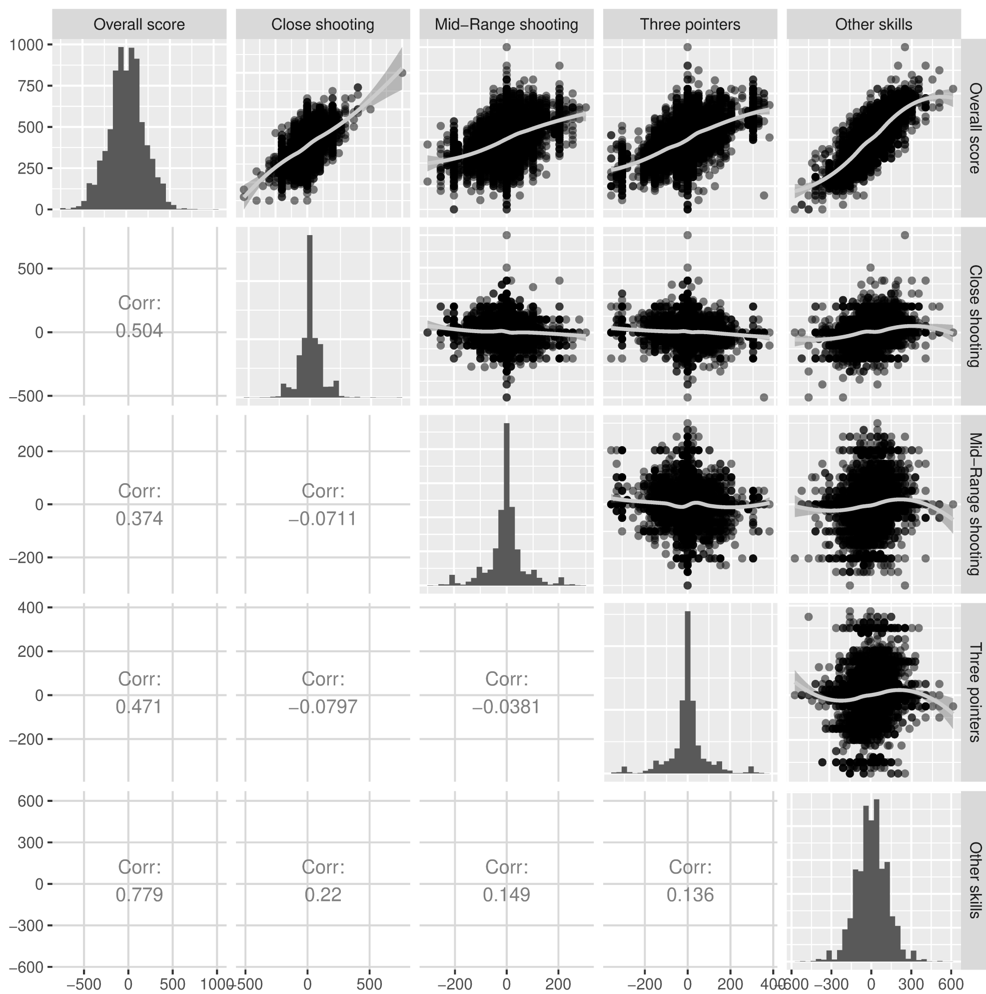

Figure 1: Relationship between the overall score and its four components; pairwise correlations below the main diagonal. As the sample size is very large all the correlations are significantly different from zero.

The analysis that follows has been developed considering the overall score, which seems more comprehensive and informative than the number of points used in the original APM and RAPM methodology. Figure 2 visualises the relationship between the two variables. There is a strong linear relationship (correlation 0.83), but the marginal distribution of the overall score is better suited for the linear regression analyses that follow, as it is smoother and without apparent mass points.

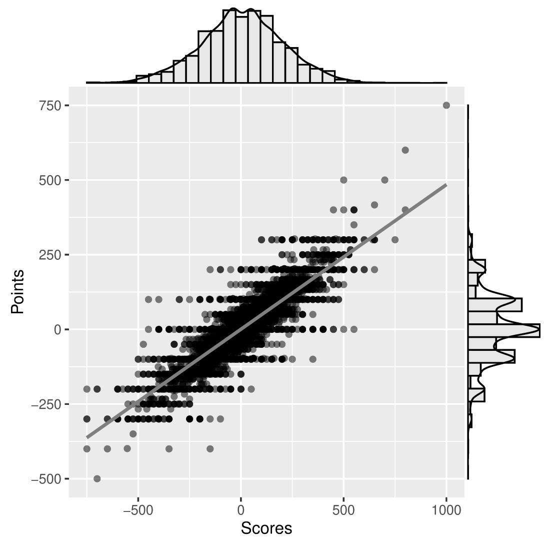

Figure 2: Observed points and overall scores for the shift data, with marginal distributions.

2.3 Data exploration

The cleansed dataset consists of 7923 shifts and 37683 possessions, from 240 matches played by the 16 teams of Euroleague Championship. The total number of lineups in the dataset is 3894 and the players involved in the games are 247.

The following preliminary analyses aim at describing the characteristics of the different teams in terms of lineups' and players' usage. Table 2 shows that a few teams present a small variability in the number of possessions at the lineup level, such as Darussafaka Istanbul and Gran Canaria. For other teams, such as Anadolu Istanbul and Olimpia Milan, the number of possessions varies substantially across different lineups. An interesting result is that, in this framework, the differences observed among teams are less pronounced than in the case of the Italian Lega A Championship (the national league considered in Grassetti et al., 2019a,b). Moreover, the behaviour of Olimpia Milan team, the one taking part in both leagues, is rather different in the two championships. This evidence can be interpreted as the result of a different approach to match management in the two frameworks.

A similar analysis is developed from the point of view of the players. The results of this further investigation (see Table 3) show that in general there are some differences among teams in the player's management approach, and the patterns of this differentiation are similar to the ones observed in the analysis of lineups.

The plots in Figure 3 visualise the choices of four selected team coaches for the management of lineups and players. The left panels summarise the distribution of the number of possessions for lineups. By observing the cumulative probability functions, some different patterns can be found. For instance, Olimpia Milan main lineups play a quite higher number of possessions than the corresponding units of Real Madrid. The four teams employ a roughly similar number of players, while the Olimpia Milan team fields a much smaller number of lineups than the other three teams. The cumulative frequency distributions regarding the number of possessions is similar across the four teams, with the values of the Gini concentration coefficient equal to 0.67 for Olimpia Milan, 0.63 for Real Madrid, 0.59 for Buducnost Podgorica and 0.56 for Panathinaikos Athens. More variation is found by extending the analysis to the full set of 16 teams, with the Gini concentration coefficient varying in the range

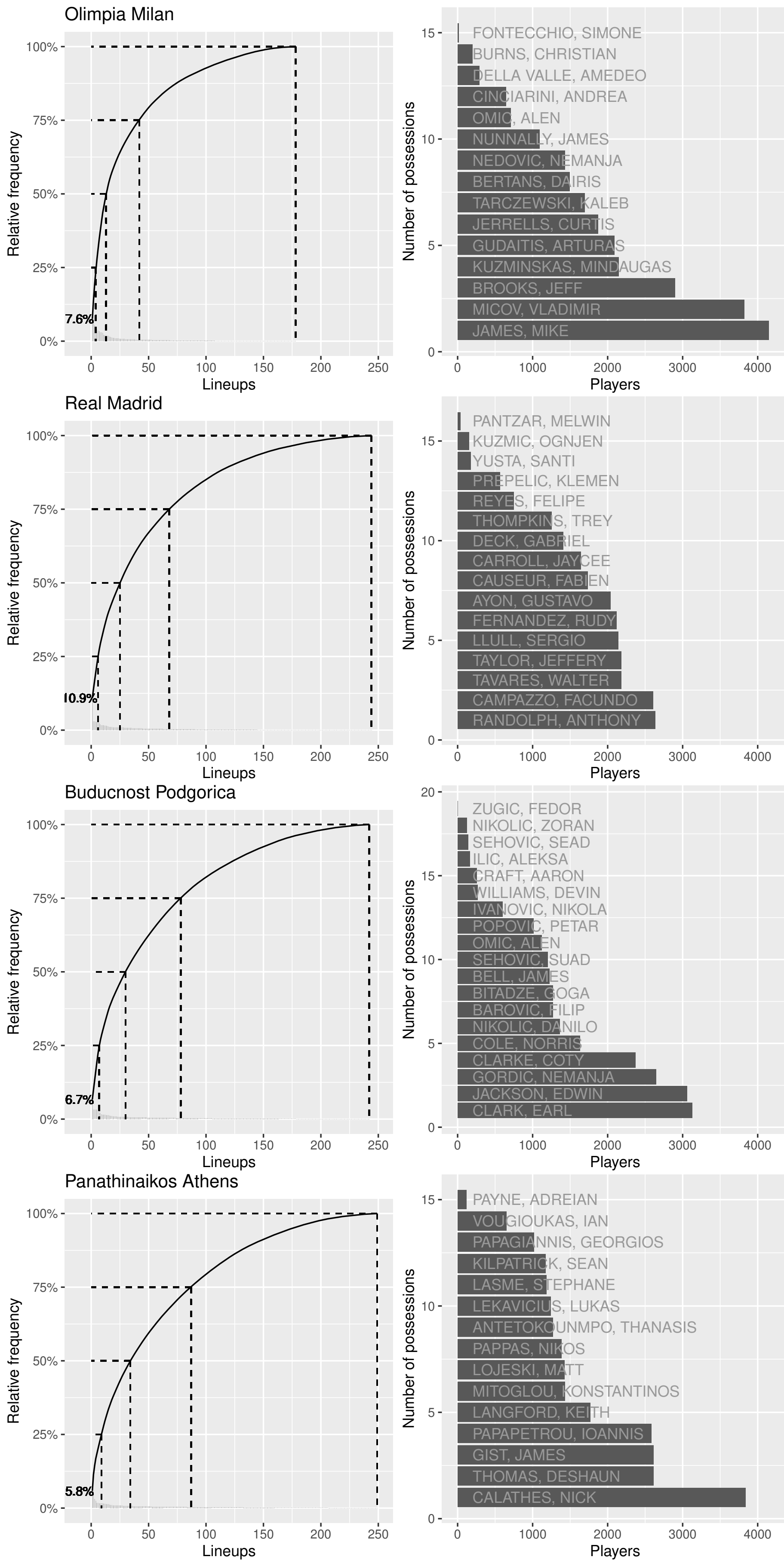
Olimpia Milan
Relative frequency
Lineups
7.6%
Number of possessions
Players
FONTECCHIO, SIMONE
BURNS, CHRISTIAN
DELLA VALLE, AMEDEO
CINCIARINI, ANDREA
OMIC, ALEN
NUNNALLY, JAMES
NEDOVIC, NEMANJA
BERTANS, DAIRIS
TARCZEWSKI, KALEB
JERRELLS, CURTIS
GUDAITIS, ARTURAS
KUZMINSKAS, MINDAUGAS
BROOKS, JEFF
MICOV, VLADIMIR
JAMES, MIKE
Real Madrid
10.9%
PANTZAR, MELWIN
KUZMIC, OGNJEN
YUSTA, SANTI
PREPELIC, KLEMEN
REYES, FELIPE
THOMPKINS, TREY
DECK, GABRIEL
CARROLL, JAYCEE
CAUSEUR, FABIEN
AYON, GUSTAVO
FERNANDEZ, RUDY
LLULL, SERGIO
TAYLOR, JEFFERY
TAVARES, WALTER
CAMPAZZO, FACUNDO
RANDOLPH, ANTHONY
Buducnost Podgorica
6.7%
ZUGIC, FEDOR
NIKOLIC, ZORAN
SEHOVIC, SEAD
ILIC, ALEKSA
CRAFT, AARON
WILLIAMS, DEVIN
IVANOVIC, NIKOLA
POPOVIC, PETAR
OMIC, ALEN
SEHOVIC, SUAD
BELL, JAMES
BITADZE, GOGA
BAROVIC, FILIP
NIKOLIC, DANILO
COLE, NORRIS
CLARKE, COTY
GORDIC, NEMANJA
JACKSON, EDWIN
CLARK, EARL
Panathinaikos Athens
5.8%
PAYNE, ADREIAN
VOUGIOUKAS, IAN
PAPAGIANNIS, GEORGIOS
KILPATRICK, SEAN
LASME, STEPHANE
LEKAVICIUS, LUKAS
ANTETOKOUNMPO, THANASIS
PAPPAS, NIKOS
LOJESKI, MATT
MITOGLOU, KONSTANTINOS
LANGFORD, KEITH
PAPAPETROU, IOANNIS
GIST, JAMES
THOMAS, DESHAUN
CALATHES, NICK

Table 2: Summary statistics for the number of possessions by lineup.

Team	No. of Lineups	Mean	Median	S.D.	Max.	C.V.
Anadolu Istanbul	140	33.1	9.5	62.9	529	1.9
Barcelona	192	24.2	13.5	38.0	372	1.6
Baskonia Vitoria	183	26.5	14.0	39.9	278	1.5
Bayern Munich	225	20.8	8.0	38.6	292	1.9
Buducnost Podgorica	242	18.9	10.0	30.5	304	1.6
CSKA Moscow	225	21.6	9.0	30.9	198	1.4
Darussafaka Istanbul	375	12.2	7.0	14.2	109	1.2
Fenerbahce Istanbul	214	21.3	11.0	33.5	331	1.6
Gran Canaria	294	15.7	11.0	18.2	148	1.2
Khimki Moscow Region	248	18.9	11.0	25.1	216	1.3
Maccabi Tel Aviv	286	16.5	8.0	35.1	470	2.1
Olimpia Milan	178	27.6	9.0	52.3	371	1.9
Olympiacos Piraeus	289	16.1	8.0	28.5	269	1.8
Panathinaikos Athens	249	19.5	12.0	30.4	283	1.6
Real Madrid	244	19.4	8.0	40.1	516	2.1
Zalgiris Kaunas	310	15.6	7.0	27.2	227	1.7

Table 3: Summary statistics for the number of possessions by player.

Team	No. of Players	Mean	Median	S.D.	Max.	C.V.
Anadolu Istanbul	13	1783.5	1747.0	1161.4	3284	0.7
Barcelona	13	1786.2	2011.0	800.8	2860	0.4
Baskonia Vitoria	15	1614.7	1674.0	1168.8	3103	0.7
Bayern Munich	14	1675.4	1811.5	1012.1	3019	0.6
Buducnost Podgorica	19	1201.8	1199.0	994.8	3128	0.8
CSKA Moscow	15	1616.7	1955.0	1084.6	3218	0.6
Darussafaka Istanbul	15	1519.3	1365.0	703.6	2698	0.5
Fenerbahce Istanbul	16	1427.5	1296.5	1126.3	2945	0.9
Gran Canaria	20	1154.2	1205.5	794.2	2426	0.7
Khimki Moscow Region	16	1463.8	1432.0	946.9	2746	0.7
Maccabi Tel Aviv	15	1571.7	1806.0	1052.5	2892	0.6
Olimpia Milan	15	1636.0	1491.0	1250.4	4146	0.8
Olympiacos Piraeus	14	1665.4	1487.0	986.9	3082	0.7
Panathinaikos Athens	15	1620.7	1385.0	932.3	3839	0.7
Real Madrid	16	1476.2	1688.5	887.5	2634	0.5
Zalgiris Kaunas	17	1421.8	1078.0	1160.8	3055	1.1

$[0.51, 0.70]$. For what concerns the coach management of players, visualised by the right panels, the different choices made by teams are evident. Such variation is quantified by the Gini concentration index for the distribution of the number of possessions per player, ranging between 0.26 and 0.48.

Figure 4 displays the performance of players and lineups for the four selected teams, measured as the average score differential between home and away teams. The averages are represented as a function of the number of possessions, and they are multiplied by 100. This is customary in the APM literature since an NBA game is roughly made of 100 possessions. These plots highlight that, for the lineups, the variability of the average score is inversely related to the number of possessions played. The number of possessions ranges from 1 to more than 500. The plots on the right panel represent the average scores for each player in the four selected teams. From these plots, it is clear that the average outcomes for Olimpia Milan players are in general slightly negative, but for one player presenting a large average score, though related to a small number of possessions. On the contrary, most of Real Madrid players present a positive and non-negligible effect. The opposite thing can be observed for Buducnost Podgorica players. Finally, the plot for Panathinaikos Athens shows a balanced number of positive and negative average scores.

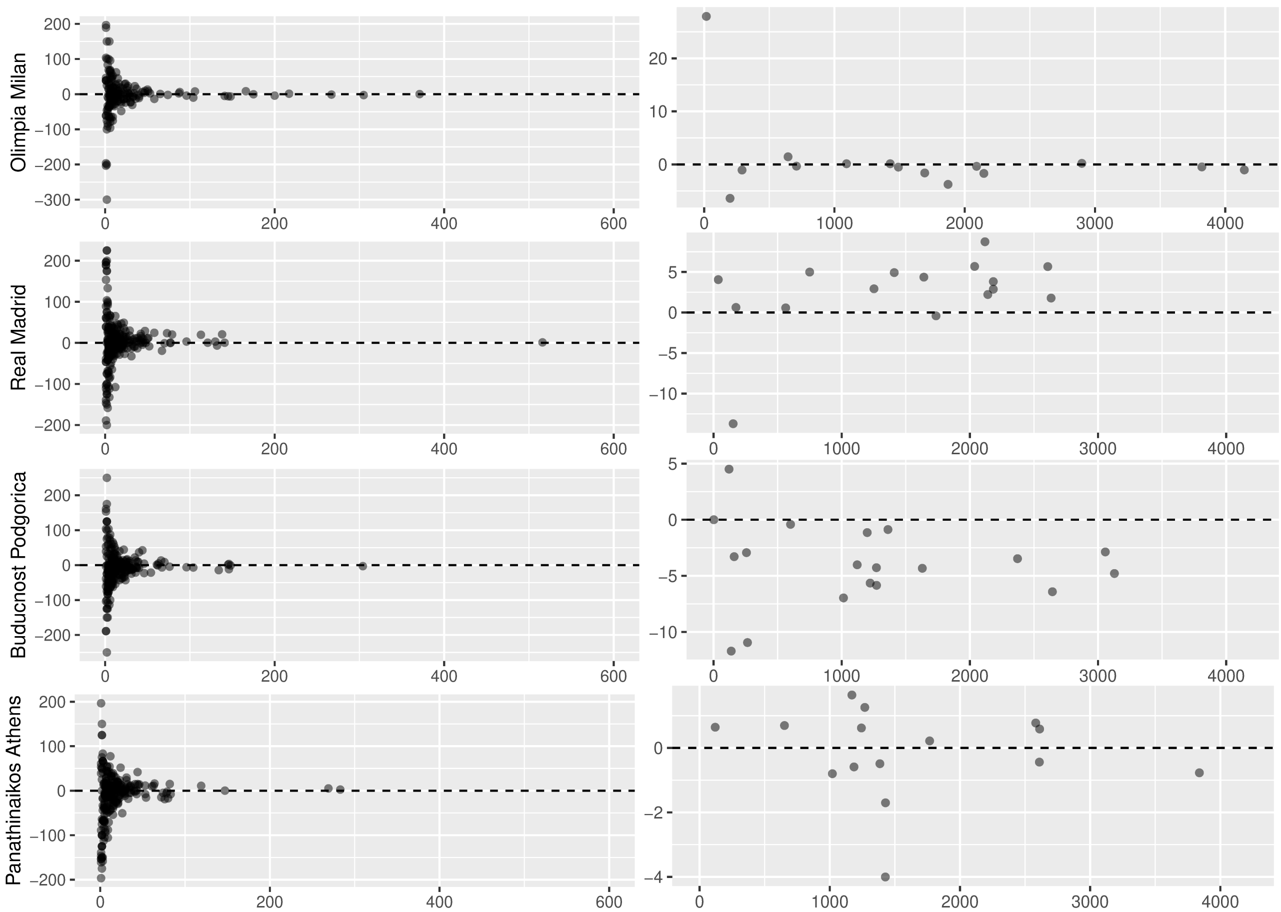

Figure 4: Average score of each lineup (left panels) and each player (right panel) for four selected teams as a function of the number of possessions.

In the subsequent analyses, the data corresponding to the so-called *garbage time* have been discarded since the information content is very low, if not detrimental. We defined this condition as the period of the match concentrated in the last 5 minutes with 20 or more points differential. The total number of excluded shifts is 170, and about 100 lineups are excluded from the model estimation procedure. All the lineups removed from the analysis have played three shifts at most.

3 Model-based analysis

The regression models for lineup's and player's effects are defined as follows. For the case of lineups, the starting point inspired from the APM literature is a linear model for the response y_t of shift t, with $t = 1, \ldots, T$,

$$y_t = \beta_0 + \mu_{h[t]} - \mu_{a[t]} + \varepsilon_t \,, \tag{3.1}$$

where, y_t is the measure of the outcome of each shift and ε_t is a normal error term. More precisely, the response in (3.1) is defined as the difference between the mean score of the home and of the away teams for each shift. Where only one team produces a score in a shift, the mean outcome of the other team is replaced by the grand mean for the entire sample, as customary in the APM analysis. Similar considerations would apply in case the number of points, or any other component of the score measure, is taken as the response variable. As just introduced, we consider all the matches of the regular season, for which $T = 7752$ (after excluding the shifts played in garbage time). The vector of lineup's effects $\boldsymbol{\mu}$ has length N, equal to the total number of lineups ($N = 3837$ in the data at hand). The notation $h[t]$ and $a[t]$ defines the lineup for the home and away teams for shift t, respectively, i.e. $h[t]$ and $a[t]$ take a value in the set $\{1, \ldots, N\}$.

The model for player's effects is very similar, with the difference that each shift entails ten different players, five for each team, rather than just two teams. Equation (3.1) is then replaced by

$$y_t = \beta_0 + \sum_{j=1}^{5} \gamma_{h_j[t]} - \sum_{j=1}^{5} \gamma_{a_j[t]} + \varepsilon_t \,, \tag{3.2}$$

where $\boldsymbol{\gamma}$ is the vector of player's effects, with length M (with $M = 247$ for the data at hand). Here the two functions $h_j[t]$ and $a_j[t]$ identify the j-th player involved in shift t, for home and away team respectively, so that each of these functions takes value in the set $\{1, \ldots, M\}$. This model specification is very similar to the one adopted by Deshpande and Jensen (2016) in a different framework.

3.1 Model extensions

Models (3.1) and (3.2) can be extended in various directions, and we propose two possibilities. The first one is the inclusion of some additional covariates, that may be of some interest. A further modification entails considering the effects of both player and lineup in the same model specification. The underlying logic is the same driving the inclusion of main and interaction effects in an analysis of variance model since lineup's effects could be viewed as a sort of high-order interaction among players. The resulting model, expressed in matrix form, is given as follows

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{Z}^{(l)}\,\boldsymbol{\mu} + \mathbf{Z}^{(p)}\,\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \,. \tag{3.3}$$

Here $\mathbf{y}$ is the vector collecting all the response variable values, and $\mathbf{X}$ is the design matrix for additional covariates, with an associated vector of coefficients $\boldsymbol{\beta}$ that includes the intercept. Furthermore $\mathbf{Z}^{(l)}$ and $\mathbf{Z}^{(p)}$ are the two design matrices for lineup's and player's effects, respectively, as defined in equations (3.1) and (3.2). Similarly, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ collect all the lineup's and player's effects. Notice that the two design matrices $\mathbf{Z}^{(l)}$ and $\mathbf{Z}^{(p)}$ are rather sparse since they have only two non-zero elements in each row in the case of lineups and ten non-zero elements in each row in the case of players. For the Euroleague data, this results in a sparsity around 99.5% for $\mathbf{Z}^{(l)}$ and around 96% for $\mathbf{Z}^{(p)}$.

It should be noted that in all the models the response variable for each shift is averaged over the observed number of possessions, so that suitable weights must be employed in the estimation procedure. This is achieved by assuming the following normal distribution for each element of $\boldsymbol{\varepsilon}$

$$\varepsilon_t \sim \mathcal{N}\left(0,\ \frac{\sigma^2}{n_t}\right) \qquad t = 1, \dots, T\,, \tag{3.4}$$

where n_t is the number of possessions for shift t. In the data at hand, the number of possessions per shift varies between 1 and 30, with a median equal to 4.

3.2 Estimation of the effects of lineup and player

The estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ requires a regularisation technique, due to their large dimensions. The RAPM method of Sill (2010) estimates $\boldsymbol{\gamma}$ by means of ridge regression, but a more flexible route is given by empirical or full Bayesian approaches that achieve regularisation by treating the effects of lineup (or player) as normal random variables.

The estimation based on empirical Bayes and that based on ridge regression are indeed related (e.g. Ruppert et al., 2003; Efron and Hastie, 2016) and essentially differ only in the approach used to select the regularisation parameter. Whereas for ridge regression the tuning parameter is usually estimated by cross-validation, in the empirical Bayes approach the variance of random effects is estimated by (Restricted) Maximum Likelihood Estimation (MLE). An important advantage of the empirical Bayes approach lies in the simple treatment of the extended model (3.3), whose estimation via ridge regression requires instead the tuning of two parameters.

Another possibility is given by a full Bayes approach, which offers some further theoretical and practical advantages (Gelman et al., 2014). The method supplements the likelihood function with prior distributions for the variance components and fixed effects. Due to its flexibility and smooth implementation, this is our main estimation proposal.

To wind up, the full set of distributional assumptions for model (3.3) adopted here is given as follows. For the lineup's and player's effects, independent normal distributions are specified

$$\mu_h \sim \mathcal{N}\left(0, \sigma_\mu^2\right) \quad h = 1, \dots, N\,, \qquad \gamma_j \sim \mathcal{N}\left(0, \sigma_\gamma^2\right) \quad j = 1, \dots, M\,, \tag{3.5}$$

and these assumptions plus that on the error term (3.4) are shared with the Bayesian approach. Furthermore, prior distributions are assumed for $\boldsymbol{\beta}$ and the three standard deviations $\sigma_\mu, \sigma_\gamma, \sigma$. Here we adopt weakly informative priors (Gelman et al., 2014, §5.7), resulting in independent zero-mean normal distribution with a (moderately) large standard deviation σ_0 for the components of $\boldsymbol{\beta}$, and Half-Cauchy distributions with a (moderately) large scale parameter A for the three standard deviations. Due to the range of the response values, which have been inflated by multiplication by 100, we set the values of σ_0 and A to 100 and 10, respectively. The former choice implies a very diffuse prior for the regression coefficients, whereas the latter corresponds to prior distributions that "even in the tail, they have a gentle slope (unlike, for example, a half-normal distribution) and can let the data dominate if the likelihood is strong in that region", as stated by Gelman (2006, §4.4).

On the computational side, the log likelihood function plus the log prior have been coded using the `R` package `TMB` (Kristensen et al., 2016), which allows for highly-efficient implementation using `C++` templates of both empirical Bayes and modal estimation (MAP) of model (3.3). The package employs automatic differentiation to obtain the gradient of the objective function, and it handles sparse matrices efficiently. The `R` package `tmbstan` (Monnahan and Kristensen, 2018) allows to sample from the posterior distribution of the model parameters given the data using the `R` interface to the `Stan` probabilistic programming language (Stan Development Team, 2020), resulting in a very smooth and reliable implementation of the full Bayesian approach.

Figure 5 represents scatterplots of estimates of lineup's effects $\boldsymbol{\mu}$ and player's effects $\boldsymbol{\gamma}$ obtained by the described route for the empirical Bayes approach (MLE label), modal estimation (MAP label) and posterior means based on MCMC sampling (Full Bayes label). It is apparent that the MLE and MAP results are nearly the same, yet adding the log prior to the log likelihood function facilitates the numerical optimisation. Indeed, MAP estimation takes just a few seconds, whereas the log likelihood maximisation for the empirical Bayes approach may be stuck into local maxima, thus requiring close monitoring of the maximisation process. A few minutes are instead required for posterior sampling, but again there is an almost perfect positive correlation between posterior modes and posterior means, as shown in Figure 5. The shrinkage of posterior means towards zero is more pronounced, and the full Bayes estimates are actually attenuated compared to the MAP or empirical Bayes estimates. We note that the inferences resulting from MAP estimation or Bayesian sampling are essentially the same, and the fact that the former may be obtained almost instantaneously may be appealing for some on-line applications. On the other hand, the availability of posterior samples offers some advantages, as it will be illustrated in what follows.

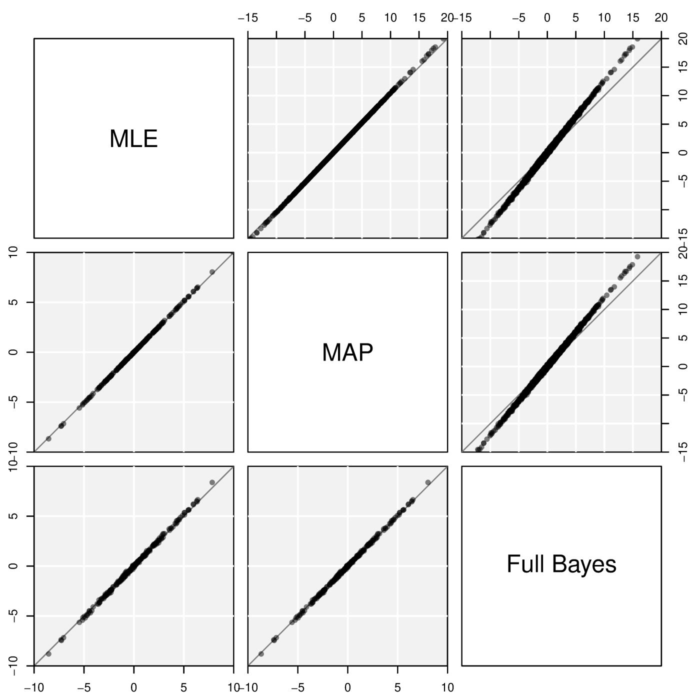

Figure 5: Player's (lower panels) and lineup's (upper panels) estimated effects obtained with different estimation methods. Bisectors added to each plot.

We then present some parameter estimates. In particular, the MAP estimates and the posterior means for model (3.3) are reported in Table 4. Some different explicative variables were considered, but the only one that seems to matter is the factor identifying the quarter in which the shifts are played. Such factor is included in the model specification by using the dummy variables for the match quarters. It is apparent that the overall advantage for home teams represented by the positive estimated intercept is compensated by negative values of the estimated coefficients for the dummy variable indicating the quarters, and the fourth one in particular. Furthermore, MAP estimates, when the response variable is given by the four components of the overall score, are reported in Table 5. Remarkably, the estimated coefficients and variance components differ across the various responses, supporting the fact that they summarise different aspects of the game.

Table 4: Overall score model estimates: MAP and full Bayes estimated fixed effects coefficients, with standard errors in brackets and 95% posterior intervals. Last quarter includes overtimes.

	MAP	Full Bayes		
Variable	Estimate (S.E.)	Estimate	95% Posterior Int.	
Intercept (Home)	10.69 (3.89)	10.75	4.10	17.49
Second Quarter	-1.53 (5.39)	-1.54	-10.97	7.76
Third Quarter	-4.35 (5.66)	-4.33	-13.95	5.21
Last Quarter	-9.84 (5.52)	-9.95	-19.79	-0.48
$\log \sigma_\mu$	2.85 (0.22)	2.64	1.79	3.11
$\log \sigma_\gamma$	1.86 (0.16)	1.85	1.56	2.11
$\log \sigma$	5.95 (0.01)	5.95	5.94	5.97

Table 5: Estimated fixed effects coefficients (via MAP) for different choice of the response variable with standard errors in brackets. Last quarter includes overtimes.

Variable	Close shooting	Mid-Range shooting	Three pointers	Other Skills
Intercept (Home)	7.50 (1.68)	0.05 (1.35)	-0.84 (1.69)	4.11 (2.17)
Second Quarter	-0.89 (2.24)	-0.88 (1.88)	-0.72 (2.32)	0.86 (3.00)
Third Quarter	-2.50 (2.37)	-3.67 (1.93)	4.61 (2.41)	-2.68 (3.06)
Last Quarter	-3.99 (2.32)	-4.35 (1.90)	0.72 (2.57)	-2.44 (3.08)
$\log \sigma_\mu$	1.73 (0.29)	2.06 (0.15)	2.60 (0.09)	1.44 (0.50)
$\log \sigma_\gamma$	1.24 (0.12)	0.27 (0.42)	1.01 (0.21)	1.04 (0.17)
$\log \sigma$	5.08 (0.01)	4.88 (0.01)	5.14 (0.01)	5.36 (0.01)

The assumptions of model (3.3) can be checked in various way. For a preliminary inspection, we obtained some plots of the residuals based on the empirical Bayes estimate. The plots do not display any shortcomings. More incisive evaluations are possible for assessing the quality of the Bayesian model fit by applying the methodology of posterior predictive checking (Gelman et al., 2014, §6.3), which is readily applied using the posterior samples. Figure 6 reports some plots obtained by the `bayesplot` package (Gabry and Mahr, 2020). The plots compare the observed score standardised by the number of possessions (i.e. $y_t \sqrt{n_t}$) with some response vectors simulated from the predictive distribution. The overall message is that the model provides a good fit, despite a somewhat attenuated tendency in reproducing scores very close to zero, and some occasional outliers.

In closing this part, we compare the empirical Bayes estimates with ones given by a penalised procedure based on ridge regression as endorsed by RAPM. The implementation of ridge regression for model (3.3) requires the writing of some ad-hoc code, for evaluating the Generalised Cross-Validation (GCV) score (Golub et al., 1979) over a two-dimensional grid of values for the tuning parameters of lineup's and player's effects, respectively. Our code making use of sparse linear algebra requires considerably longer computational time than any other method employed. Furthermore, ridge regression provides only point estimates rather than a full set of inferential outcomes. At any rate, the ridge

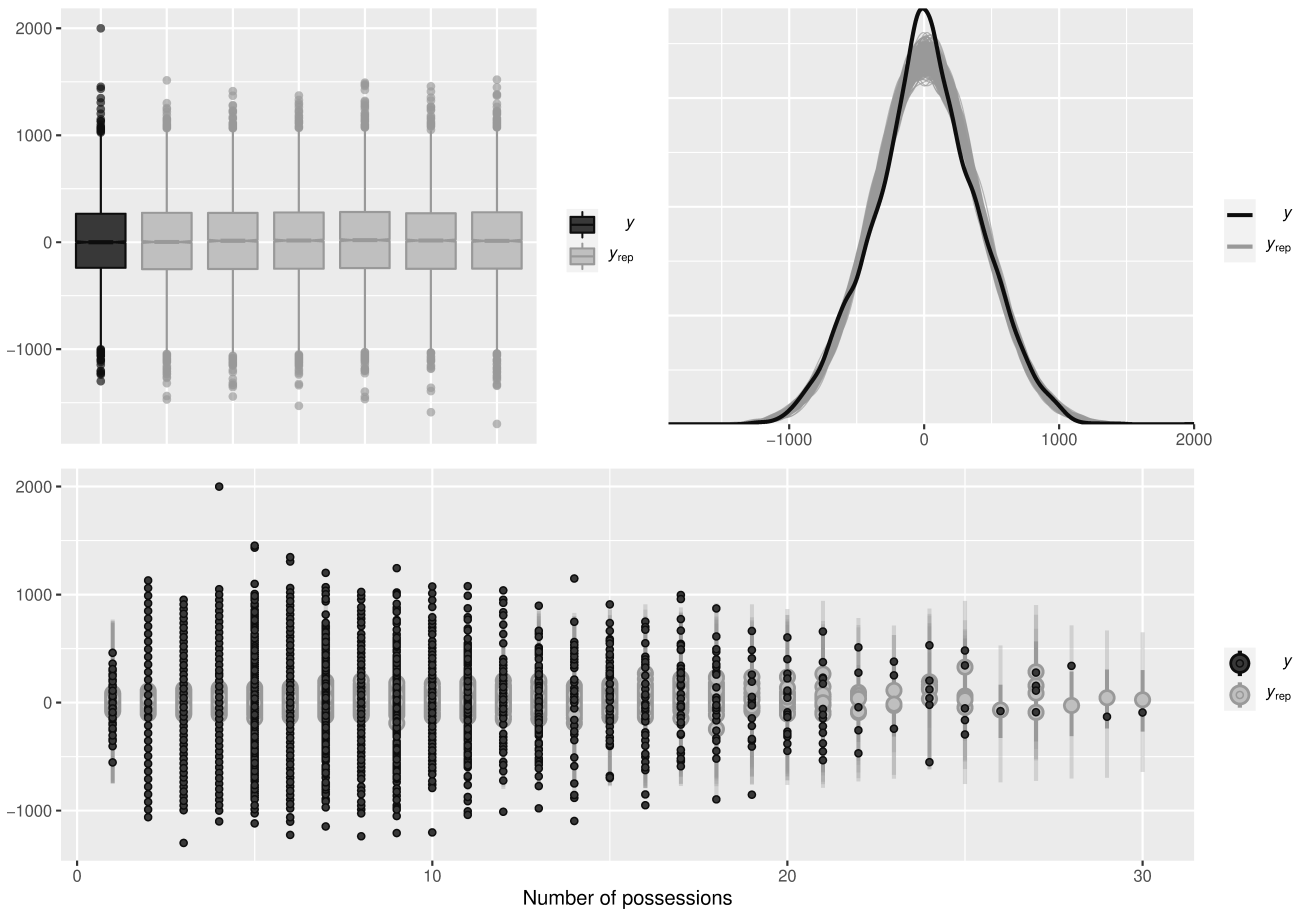

Figure 6: Bayesian predictive model checking: comparison of the standardised observed score with some samples simulated from the predictive distribution. Top panel, left: boxplots of observed response and of six simulated samples. Top panel, right: density plots of observed variable and of 500 simulated samples. Bottom panel: 50% and 95% intervals computed using the predictive distribution, with actual values of the observed score superimposed, sorted by the number of possessions.

regression results for lineup's effects and player's effects turned out to be very close to the empirical Bayes estimates, as displayed in Figure 7. Moreover, classic theory on Best Linear Unbiased Predictor estimation of random effects allows to map the tuning parameters of ridge regression into the parameter space of model (3.3) (e.g. Robinson, 1991). In the case under study, the ridge regression estimates correspond to $\log \widehat{\sigma}_\mu = 2.92$, $\log \widehat{\sigma}_\gamma = 1.75$ and $\log \widehat{\sigma} = 5.95$, respectively, suggesting that ridge regression provides less smoothing for lineup's effects and more for player's effects compared to the full Bayes approach. Yet, the three estimated variance components are well inside the 95% posterior intervals reported in Table 4.

4 Analysis of estimated effects of lineup and player

The results of the model fitting allow for several investigations and, to this end, some observations are in order. Considering the model specification (3.3), the outcome of the estimation procedure can be

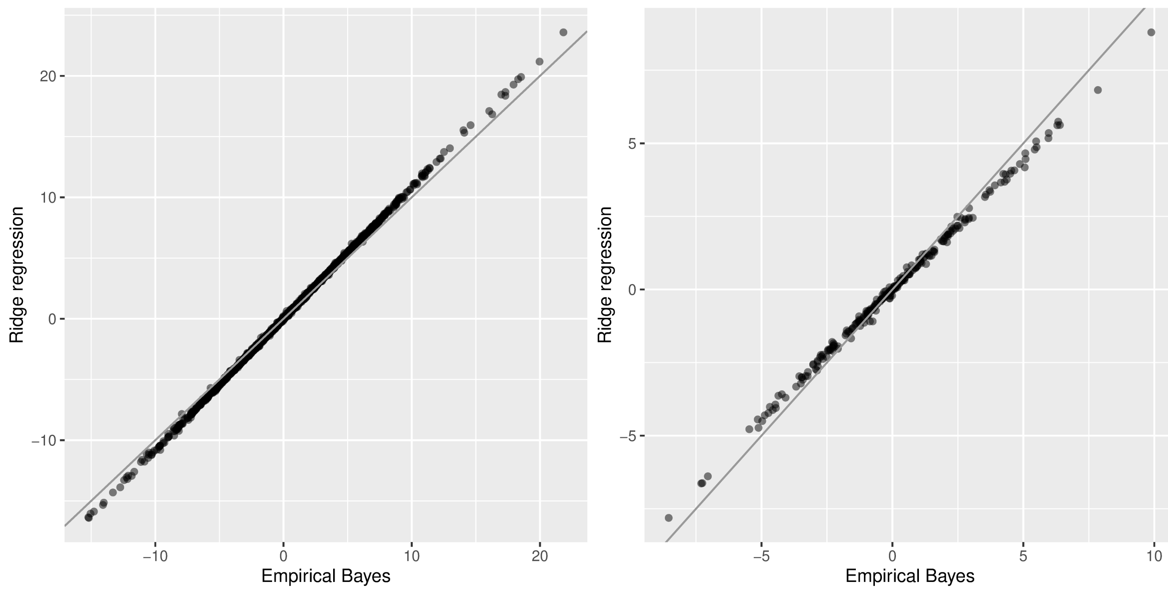

Figure 7: The comparison between empirical Bayes and ridge regression results: estimated lineup's effects (left panel) and player's effects (right panel). Bisectors added to each plot.

used to define different measures for each lineup. The first one is given by the sum of the estimated effects $\widehat{\boldsymbol{\gamma}}$ (which can be positive or negative) of the players entering the lineup, that is defined without considering any interaction among the five components. The interaction effect is caught instead by the estimated $\widehat{\boldsymbol{\mu}}$, which could be labelled as a *pure lineup effect.* Note that we can also define the *total lineup effect* as the sum of the two aforementioned effects. The latter is employed in Figure 8, which is the estimated counterpart of Figure 4 since it reports the estimated effects for lineup and players for the same four teams. Comparing the two figures, some adjustments are apparent since the estimates take into account the different lineups and players which are simultaneously present on the field. The shrinkage provided by regularisation is also noteworthy. The most relevant finding concerns the location of the estimated total lineup effects, which depends on the quality of the overall performance. For example, the majority of the Real Madrid lineups have a positive estimated effect, whereas the opposite holds for the Buducnost Podgorica team.

Furthermore, it is of interest to investigate the connection between the two sets of estimates, $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\gamma}}$. To this end, Figure 9 displays the relationship between the pure lineup effects and the sum of player's effects, for the model estimated using the overall score as a response variable. The observed correlation is positive, as suggested by the smoother, yet the kind of information captured by the pure lineup effect is only partially observed in the sum of the player's effects. In other words, the effect of interaction among players of the same lineup is real, and it is far from being captured by the simple sum of individual player's effects. Similar plots, not shown, are obtained when the four components of the overall score are considered as the response variable, extending the same conclusion to the various aspects of the game.

To enhance the preliminary analysis on the overall score reported in Section 2.3, further considerations could be made on the separate estimated lineup's effects associated with the four components of the score shown in Figure 1. Here we are referring to the total lineup effect, as defined above. The relationship between the lineup's effects estimated when the response variable equals the overall score

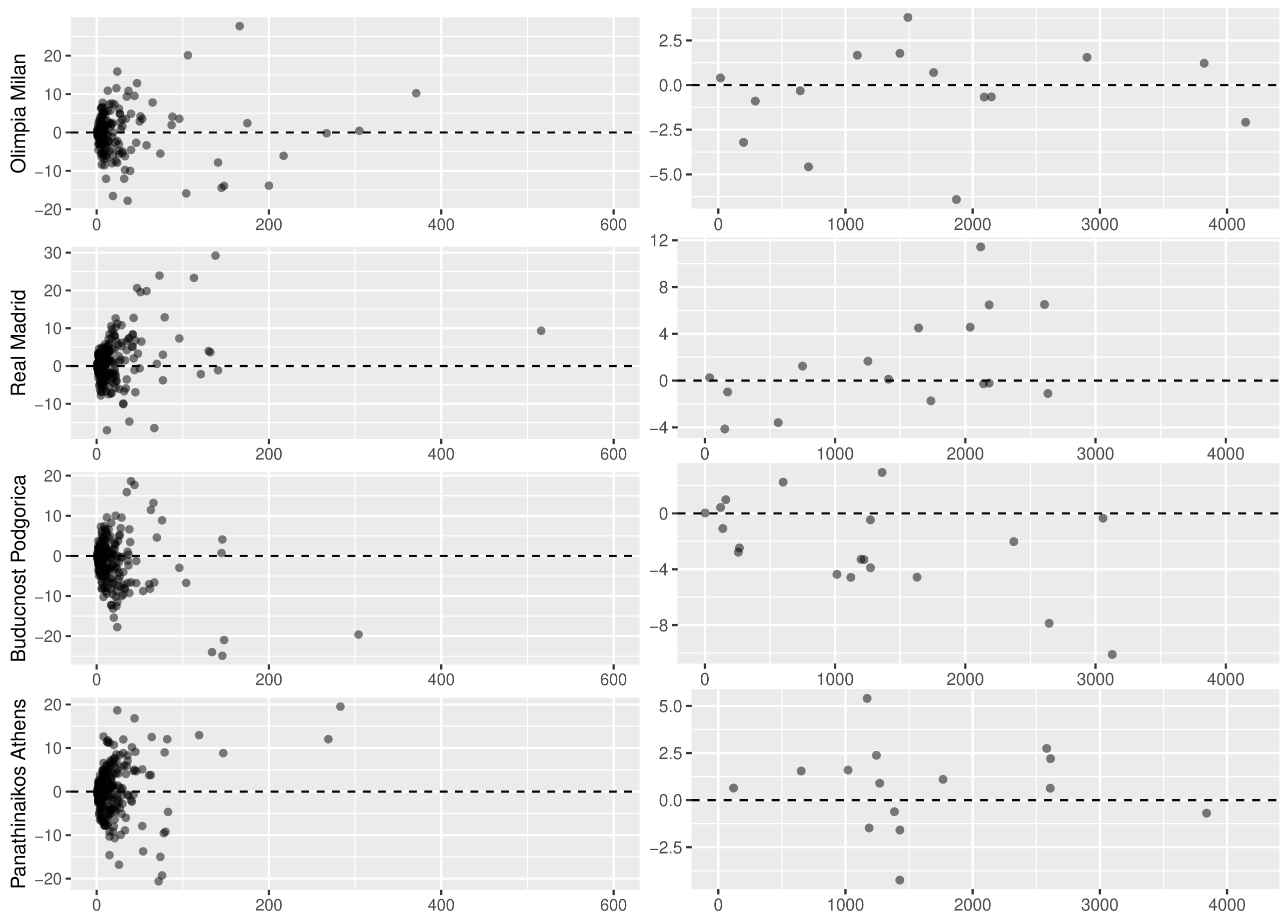

Figure 8: Estimated total effects of each lineup (left panels) and of each player (right panel) for four selected teams as a function of the number of possessions.

and the corresponding measures obtained, when the response variable is replaced in turn by the four components, lead to some plots very similar to those in Figure 1. Again, the observed relationships among different components are rather diverse, testifying the intrinsic multi-dimensionality of the shift outcomes.

The estimation of total lineup effects for the different responses is summarised at the team level in Table 6. Notice that the sum of the disentangled estimated components is close but not equal to the results based on the overall score since the models related to the single components are independently estimated. The table reports the averages of the different estimated total lineup effects, weighted by the number of possessions, along with the corresponding team ranking obtained from the averages. For instance, the Real Madrid team presents a positive global effect which is split into the four components, suggesting that all but the mid-range shooting contribute to the positive score. Instead, Khimki Moscow Region presents a negative global evaluation, but it has the fifth score across all teams for what concerns three-pointer efficiency. This is rather abridged, and it mirrors the team statistics about the different aspects, but it hints at the possibility for a single team to evaluate its different lineups in terms of different efficiency measures. This has some potential for team management, as further illustrated in the following section.

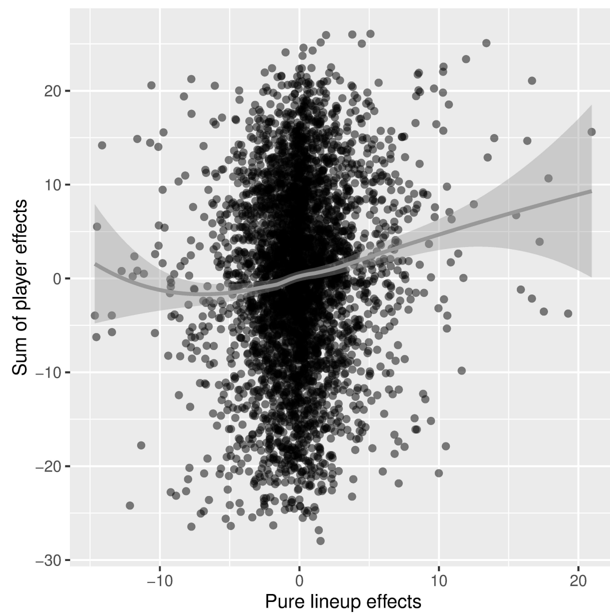

Figure 9: Estimated lineup's effects and sum of the player's effects for the overall score, with a robust smoother added.

Table 6: Averages of estimated total lineup effects for the overall score and its four components and the corresponding team rankings (in brackets). Teams are ordered according to the official standings at the end of the regular season.

Teams	Overall score	Close shooting	Mid-Range shooting	Three pointers	Other skills
Fenerbahce Istanbul	16.17 (1)	3.18 (6)	0.88 (4)	7.17 (1)	4.19 (3)
CSKA Moscow	12.62 (3)	4.65 (4)	-0.10 (8)	3.83 (4)	4.38 (2)
Real Madrid	14.51 (2)	5.25 (3)	-0.88 (14)	5.32 (3)	4.81 (1)
Anadolu Istanbul	8.85 (5)	1.96 (8)	1.40 (3)	6.24 (2)	-0.60 (11)
Barcelona Lassa	5.06 (7)	5.67 (2)	-0.147 (9)	-3.85 (16)	3.24 (7)
Panathinaikos Athens	3.96 (9)	2.18 (7)	2.60 (1)	-2.76 (13)	1.20 (8)
Baskonia Vitoria	7.22 (6)	3.44 (5)	-0.17 (11)	-0.03 (9)	3.86 (4)
Zalgiris Kaunas	9.68 (4)	9.25 (1)	0.60 (6)	-3.60 (15)	3.60 (5)
Olympiacos Piraeus	4.43 (8)	-0.51 (10)	1.73 (2)	-1.37 (11)	3.53 (6)
Maccabi Tel Aviv	1.24 (11)	-1.28 (11)	0.83 (5)	0.98 (7)	0.61 (9)
Bayern Munich	1.34 (10)	0.36 (9)	-0.16 (10)	0.88 (8)	0.44 (10)
Olimpia Milan	-0.77 (12)	-1.54 (12)	-0.80 (13)	2.82 (6)	-0.92 (12)
Khimki Moscow Region	-5.77 (13)	-5.67 (15)	-0.04 (7)	2.95 (5)	-2.99 (14)
Gran Canaria	-8.38 (14)	-4.06 (14)	-2.03 (16)	-0.14 (10)	-1.32 (13)
Buducnost Podgorica	-17.04 (16)	-6.54 (16)	-0.35 (12)	-1.70 (12)	-7.53 (16)
Darussafaka Istanbul	-15.33 (15)	-1.65 (13)	-1.75 (15)	-2.86 (14)	-7.30 (15)

5 Team management and ranking issues

The management of a sports team can be viewed as a subjective decision process which can be supported by advanced statistical analyses based on the available data on past matches. In this section, we discuss the possibility of using the estimated random effects defined in Section 3 in order to compare players and lineups by means of standardised measures.

There is a substantial agreement in stating that the quality of a sports team is not simply the sum of the individual abilities of the players since a team is a collection of interdependent individuals who share responsibilities for match outcomes (Lechner and Gudmundsson, 2012). Thus, for every coach or manager of professional basketball teams, it is important to assess both the quality of the lineup's interplay and the performance of the individual players on the court, accounting for the value of the opponent's lineups and players and for further concomitant conditions, as the time of the match and the home-field advantage. The data-driven evaluation of the effects $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$, specified in the model (3.3), makes it possible to analyse the net performances of lineups and players. The combination of these two effects produces a realistic assessment of the team performance, given as the composition of the individual skills and of the effectiveness of interaction among players. These conclusions may foster better management of the available human resources and they can provide additional information also to experienced coaches for selecting the best lineups and improving the interaction among the players (Berri, 1999).

In the present section, we select one particular team, and we show how it is possible to use the inferential results presented in Section 3.1 in order to describe and, possibly, improve the team performance. We focus on the Olimpia Milan team and analyse the corresponding lineup's and player's effects, by considering the posterior means obtained in the full Bayes approach by posterior sampling. In Table 7, we consider nine lineups with at least fifty possessions and report the estimated total lineup effects related to both the overall score and the associated four components. According to the overall score, the chosen lineups correspond to the three best and the three worst lineups, and to three average ones. Note that the summarised measures are the total lineup effects, as defined above. As in Table 6, the sum of the effects estimated on the disentangled components is not equal to the effect estimated on the overall score. As noted before, the models are estimated separately.

In addition, Table 8 summarises the estimated player's effects, related to the overall score and the associated four components, for all the players of the Olimpia Milan team. By comparing Tables 7 and 8, we note that the best and the worst-performing lineups do not always involve the best and the worst-performing players, respectively. This is clearly related to the fact that the lineup's composition depends on several different aspects, such as the tactics, the role of the players, the composition of the opposing lineup, the time of the match. However, the possibility of measuring the net contribution of every single player is an important piece of information for effective team management. Some care is needed in reading the results of Table 8: for example, the positive three-pointer effects estimated for players such as Tarczewski and Gudaitis are likely to derive from being part of good-performing lineups in three-pointers, rather than to their own good performance in three-point shooting.

Since an effective lineup should have a good interaction among the individuals, as measured by the pure lineup effect, and should involve players with high performance, we represent in Figure 10 the lineups of the Olimpia Milan team according to these two key aspects. Additional information concerns the number of possessions of each lineup. This results in a clear graphical representation of the overall quality of a lineup, emphasising that the best five-man units are in the upper right corner. These findings may be useful for comparing the performance of all the observed lineups and for identifying the best ones.

A further application concerns the possibility of using the estimates of the effects $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ for

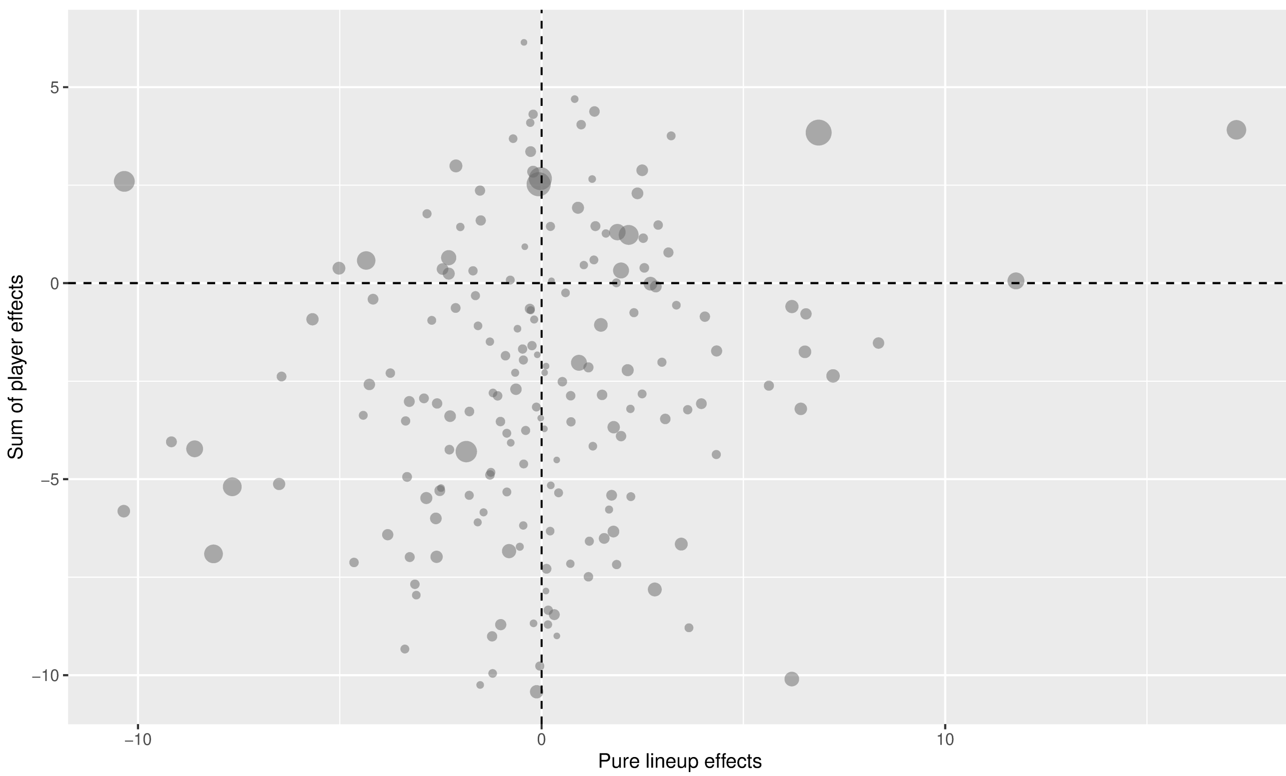

Figure 10: Lineups of the Olimpia Milan team classified according to the pure lineup effect and the sum of player's effects, related to the score response. The size of the circles is proportional to the number of possessions.

Table 7: Estimated total lineup effects for 9 selected lineups of the Olimpia Milan team with at least 50 possessions; both the overall score and the four components are considered.

Lineup	No. poss.	Overall score	Close shoot.	Mid-Range shoot.	Three pointers	Other skills
Bertans, Brooks, James, Micov, Tarczewski	166	21.13	6.31	-1.04	6.09	2.29
James, Kuzminskas, Micov, Nedovic, Tarczewski	106	11.81	0.97	9.29	4.62	-1.56
Bertans, Brooks, Gudaitis, James, Micov	371	10.71	-2.17	-0.43	9.10	3.34
Cinciarini, Gudaitis, James, Kuzminskas, Micov	87	-1.11	2.16	-5.81	9.68	-1.07
Brooks, Cinciarini, James, Micov, Tarczewski	74	-1.65	-1.88	-1.60	11.91	-1.70
Brooks, Cinciarini, Gudaitis, James, Micov	141	-3.77	-3.14	-2.57	0.61	0.40
Brooks, James, Jerrells, Micov, Tarczewski	104	-12.82	-6.14	-0.22	8.17	-4.67
Bertans, Gudaitis, James, Jerrells, Kuzminskas	148	-12.86	-4.14	-8.52	-1.66	-0.61
Gudaitis, James, Jerrels, Kuzminskas, Micov,	145	-15.04	-5.84	-4.56	-0.15	-2.80

ranking the lineups or the players according to their specific net contribution to the overall score. Ranking methods have a long and important tradition. A natural application is to sport data, where they can be used to compare the value of teams or players, and also for predicting the results of future matches; see, for example, Govan et al. (2009), and Langville and Meyer (2012).

A ranking procedure is a method for assigning a rank, that is an integer from the set $\{1,\ldots,K\}$, to the $K>1$ objects of a given set. The rank specifies the relative importance of an object and the assignment is ideally one-to-one. Usually, the objects are ranked according to the (estimated) value of an unknown parameter θ_k, $k=1,\ldots,K$, typically a (numerical) rating which measures a performance value or an ability level. In our framework, the parameter θ_k corresponds to a specific lineup's or player's effect, namely, a single component of the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$, respectively, or also their composition which defines the total lineup effect. Since we apply a full Bayes approach, the ranking procedure (Laird and Louis, 1989) is based on the joint posterior distribution of the effects $\theta_1,\ldots,\theta_K$, where the components are supposed to be marginally independent and normally distributed, at least approximatively. Then, a first possibility is to rank the units according to the posterior distribution of the effects, i.e. by employing their posterior mean or median values.

Table 8: Players of the Olimpia Milan team: estimated player's effects related to the overall score and the associated four components.

Player	No. poss.	Overall score	Close shoot.	Mid-Range shoot.	Three pointers	Other skills
Bertans	1491	2.86	1.25	-0.34	0.56	1.63
Brooks	2899	1.64	-0.79	0.36	0.70	0.64
Nedovic	1428	1.62	-1.53	0.09	2.07	-0.19
Nunnally	1093	1.47	2.03	0.14	-0.87	0.47
Micov	3820	1.15	0.96	0.18	0.16	-0.16
Fontecchio	17	0.35	0.29	0.02	-0.05	0.10
Tarczewski	1693	0.10	0.39	-0.05	0.94	-0.90
Gudaitis	2091	0.02	-0.40	-0.87	1.56	0.32
Cinciarini	644	-0.40	-0.02	-0.77	0.91	0.30
Della Valle	291	-0.74	-1.02	0.09	0.17	-0.38
Kuzminskas	2147	-0.97	1.20	-1.01	0.17	-0.09
James	4146	-1.83	-0.79	-0.03	-0.14	-0.93
Burns	199	-2.84	-2.03	0.24	-0.66	-1.02
Omic	709	-3.49	-2.90	0.68	-1.58	-1.01
Jerrells	1872	-5.28	-4.15	-0.79	1.20	-1.50

An alternative approach is based on ranks defined as

$$R_k = \sum_{j=1}^{K} \mathrm{I}(\theta_k \leq \theta_j) \quad k = 1, \ldots, K, \tag{5.1}$$

with $\mathrm{I}(\cdot)$ the indicator function. In this case, the units are ranked by considering the joint posterior distribution of $R_1, \ldots, R_K$, employing their posterior mean or median values. As noted by Laird and Louis (1989), when the marginal posterior variances of the θ_ks are equal, the two approaches will tend to produce the same final ranking. In our setting, this is not always the case. In particular, the estimated effects exhibit some heterogeneity in their variance since the variance will be smaller with a higher number of possessions, and this is coupled by possible asymmetric distribution for some of the ranks.

The use of the posterior ranks also permits the specification of more effective confidence statements, useful for possible evaluation of the significance in rank differences. Starting from the MCMC simulated values for the posterior distribution of the effects of the lineup and of the player, it is quite immediate to obtain suitable probability intervals based on posterior quantiles of the effects and their corresponding ranks, as defined in equation (5.1).

As an application, we consider the subset of the 24 lineups presenting more than 200 possessions and we compute 50% and 80% posterior intervals for the total lineup effect and the associated ranks. These results, reported in Figure 11, are obtained by considering the overall score response in model (3.3). The plots confirm that the final outcome depends on the approach followed. In particular, the rank-based probability intervals are usually smaller and, consequently, they represent a more effective route for judging the significance of the observed mean rank differences. The lineups of Olimpia Milan team are highlighted in the plots. The comparative analysis of their rankings may guide the choice of the best lineup for each opponent, thus supplying a further tool for the team management.

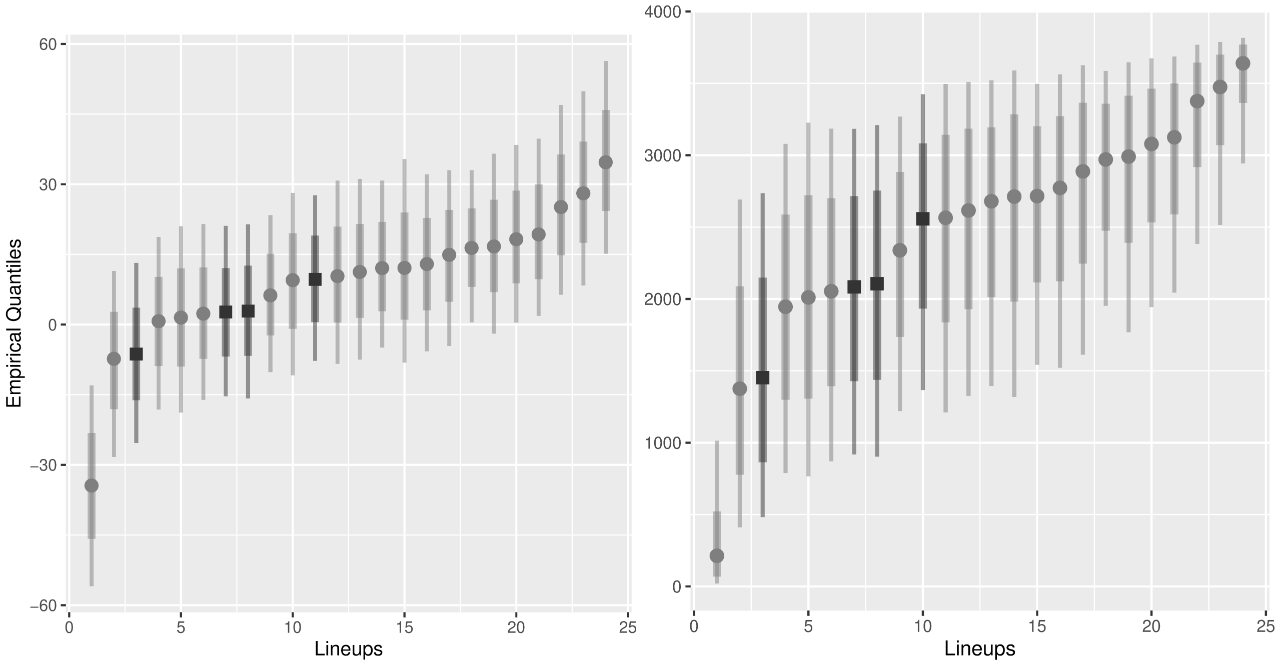

Figure 11: Lineups with more than 200 possession: posterior probability intervals with 80% level (thin line) and 50% level (bold line) for the total lineup effect (left panel) and the related rank (right panel), using results for the overall score model. Dark grey squares highlight the lineups of the Olimpia Milan team.

We have described two alternative approaches for ranking lineups and players: the first one based on the posterior distribution of the effects and the second one on the posterior distribution of their ranks. The two approaches may lead to different conclusions. In case of disagreement, we tend to prefer the rank-based approach, which usually provides straightforward comparative evaluations and more effective final rankings. Moreover, rankings can also be determined by considering the results obtained from the analysis of the four different components of the overall score, leading to a more comprehensive evaluation. Starting from the rankings built on the overall score we can better describe the characteristics of a player or a lineup by means of the analysis of the component-specific performances.

All the results reported in the present section are based on the analysis of the data about the complete regular season. From the team management point of view, the possibility to study the behaviour of players and lineups during the season can represent a great advantage. To this aim, the model can be applied recursively starting approximately after the first half of the season. This threshold is set up in order to collect a sufficient amount of information. In this framework, the estimated effects can be adopted in the decision-making process in order to select the best lineups and to plan the substitutions. The results of this kind of dynamic analysis can also be considered in the analysis of the trends of the efficiency of the lineups and players. A first glimpse of this kind of analysis is given in Figure 12, which represents the time trend of the estimated effects for the top five players of the entire league, namely the five players with the highest value of $\widehat{\gamma}$. Observing the time trends, it is clear that during the period of observation, some players improve their performances regularly, while some other display a less regular pattern. The on-line identification of these changes may help the management of the team. For instance, a decreasing trend can be a signal of bad physical condition

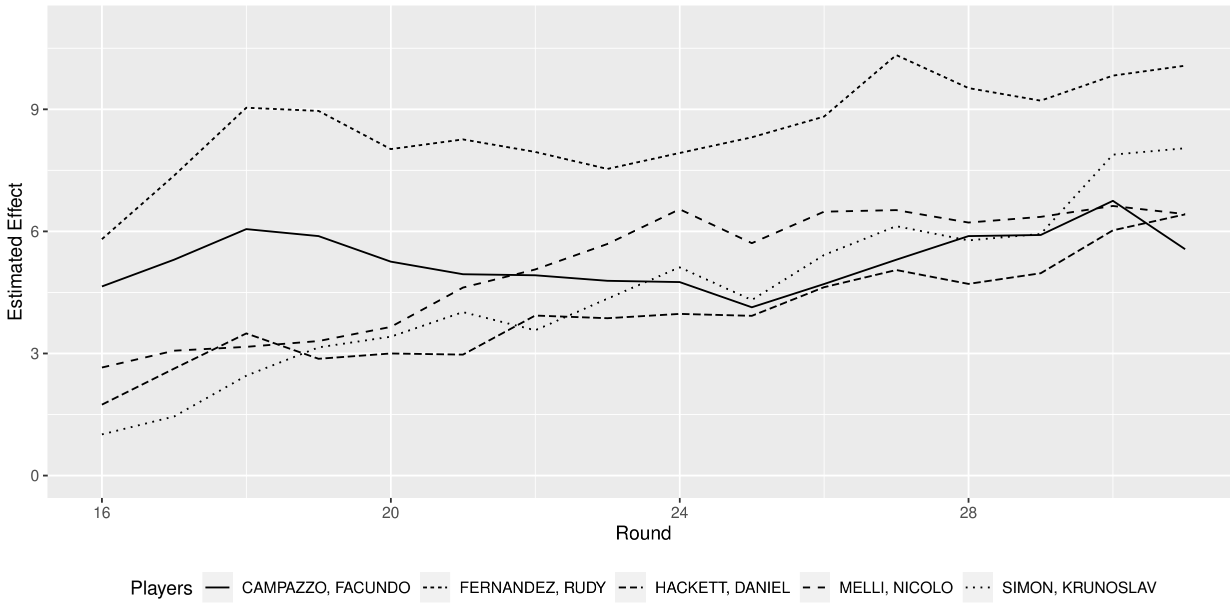

Figure 12: The trend of the estimated effects of the five best players.

and this information can be very useful in preventing injuries.

6 Conclusions

The literature on basketball analytics considers the RAPM measures as an important tool for the evaluation of single players. This article extends the RAPM methodology by considering in the model specification a response variable which is more comprehensive than the points scored, and includes further important features. Moreover, the estimation of lineup's effects is developed in addition to that of player's effects. The proposed model has been estimated following a Bayesian approach, which offers several advantages compared to the ridge regression approach adopted in the literature.

The novelties introduced in this paper allow for further exploitation of the information carried in the play-by-play data, defining new analyses potentially useful for team management. Most of the reported results are based on the model for the overall score, but similar analyses could be extended to the four components of the overall score, thus identifying various complementary aspects of the game. Despite the fact that the analyses are built on solid statistical methodology, the emphasis on the presentation of the results is on graphical tools, targeting the needs of the team managers.

Funding

This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant 2015EASZFS_003.

Acknowledgements

The authors thank the editors and the anonymous reviewers for suggestions and comments which served to improve the paper. Thanks also to Euroleague Basketball for making the play-by-play data available.

References

Berri, D. J.: 1999, Who is 'most valuable'? measuring the player's production of wins in the national basketball association, *Managerial and Decision Economics* **20**(8), 411–427.

Berri, D. J., Leeds, M., Leeds, E. M. and Mondello, M.: 2009, The role of managers in team performance, *International Journal of Sport Finance* **4**(2), 75–93.

Berry, C. R. and Fowler, A.: 2019, How much do coaches matter?, *Sports Analytic Conference, MIT Sloan, Boston, MA.* (accessed November 22, 2019).
URL: *http://www.sloansportsconference.com/wp-content/uploads/2019/02/Do-Coaches-Matter.eps*

Carmichael, F., Thomas, D. and Ward, R.: 2001, Production and efficiency in association football. *Journal of Sports Economics* **2**(3), 228–243.

Deshpande, S. and Jensen, S.: 2016, Estimating an NBA player's impact on his team's chances of winning, *Journal of Quantitative Analysis in Sports* **12**(2), 51–72.

Efron, B. and Hastie, T.: 2016, *Computer Age Statistical Inference*, Cambridge University Press.

Engelmann, J.: 2011, A new player evaluation technique for players of the National Basketball Association (NBA), *Proceedings of the MIT Sloan Sports Analytics Conference.*

Engelmann, J.: 2017, Possession-based player performance analysis in basketball (adjusted +/- and related concepts), *Handbook of Statistical Methods and Analyses in Sports*, Chapman and Hall/CRC, pp. 231–244.

Gabry, J. and Mahr, T.: 2020, *R Package `bayesplot`: Plotting for Bayesian models.* R package version 1.7.2.
URL: *https://mc-stan.org/bayesplot¿*

Gagolewski, M.: 2019, *R Package `stringi`: Character string processing facilities.* R package version 1.4.6.
URL: *http://www.gagolewski.com/software/stringi/*

Gelman, A.: 2006, Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis* **1**(3), 515–534.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D.: 2014, *Bayesian Data Analysis (Third Edition)*, Chapman and Hall/CRC.

Golub, G. H., Heath, M. and Wahba, G.: 1979, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21**(2), 215–223.

Govan, A. Y., Langville, A. N. and Meyer, C. D.: 2009, Offense-defense approach to ranking team sports, *Journal of Quantitative Analysis in Sports* **5**(1).

Grassetti, L., Bellio, R., Fonseca, G. and Vidoni, P.: 2019a, Estimation of lineup efficiency effects in basketball using play-by-play data, *in* G. Arbia, S. Peluso, A. Pini and G. Rivellini (eds), *Book of Short Papers SIS2019*, Pearson, , pp. 363–370.

Grassetti, L., Bellio, R., Fonseca, G. and Vidoni, P.: 2019b, Play-by-play data analysis for team managing in basketball, *in* K. Dimitris, N. Ioannis and D. Sotiris (eds), *Proceedings of Mathsport International 2019*, Propobos Publications, pp. 129–139.

Haigh, J.: 2009, Uses and limitations of mathematics in sport, *IMA Journal of Management Mathematics* **20**(2), 97–108.

Hollinger, J.: 2005, *Pro Basketball Forecast, 2005-06*. Dulles, VA: Potomac Books Inc.

Hughes, M. D. and Bartlett, R. M.: 2002, The use of performance indicators in performance analysis, *Journal of Sports Sciences* **20**(10), 739–754.

Hvattum, L. M.: 2019, A comprehensive review of plus-minus ratings for evaluating individual players in team sports, *International Journal of Computer Science in Sport* **18**(1), 1–23.

Ilardi, S.: 2007, Adjusted plus-minus: An idea whose time has come, *82games. com* .

Kao, S.-F., Hsieh, M.-H. and Lee, P.-L.: 2017, Coaching competency and trust in coach in sport teams, *International Journal of Sports Science & Coaching* **12**(3), 319–327.

Karlis, D. and Ntzoufras, I.: 2009, Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference, *IMA Journal of Management Mathematics* **20**(2), 133–145.

Kharrat, T., Peña López, J. and McHale, I.: 2017, Plus-minus player ratings for soccer, *arXiv preprint arXiv:1706.04943* .

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. and Bell, B. M.: 2016, TMB: Automatic differentiation and Laplace approximation, *Journal of Statistical Software* **70**(5), 1–21.

Kubatko, J., Oliver, D., Pelton, K. and Rosenbaum, D. T.: 2007, A starting point for analyzing basketball statistics, *Journal of Quantitative Analysis in Sports* **3**(3).

Kvam, P. and Sokol, J.: 2006, A logistic regression/markov chain model for NCAA basketball, *Naval Research Logistics (NrL)* **53**(8), 788–803.

Laird, N. M. and Louis, T. A.: 1989, Empirical Bayes ranking methods, *Journal of Educational Statistics* **14**(1), 29–46.

Langville, A. N. and Meyer, C. D.: 2012, *Who's # 1?: the Science of Rating and Ranking*, Princeton University Press.

Lechner, C. and Gudmundsson, S. V.: 2012, Superior value creation in sports teams: Resources and managerial experience, *Management* **15**(3), 284–312.

Macdonald, B.: 2012, Adjusted plus-minus for NHL players using ridge regression with goals, shots, Fenwick, and Corsi, *Journal of Quantitative Analysis in Sports* **8**(3).

Manner, H.: 2016, Modeling and forecasting the outcomes of NBA basketball games, *Journal of Quantitative Analysis in Sports* **12**(1), 31–41.

Masoumzadeh, A., Toloo, M. and Amirteimoori, A.: 2016, Performance assessment in production systems without explicit inputs: an application to basketball players, *IMA Journal of Management Mathematics* **27**(2), 143–156.

Monnahan, C. and Kristensen, K.: 2018, No-u-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages, *PLoS One* **13**(5).

R Core Team: 2020, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Robinson, G. K.: 1991, That BLUP is a good thing: the estimation of random effects (with discussion), *Statistical Science* **6**(1), 15–51.

Rosenbaum, D.: 2004, Measuring how NBA players help their teams win, *82Games.com (http://www.82games.com/comm30.htm)* .

Ruiz, F. and Perez-Cruz, F.: 2015, A generative model for predicting outcomes in college basketball, *Journal of Quantitative Analysis in Sports* **11**(1), 39–52.

Ruppert, D., Wand, M. P. and Carroll, R. J.: 2003, *Semiparametric Regression*, Cambridge University Press.

Shen, K.: 2014, Data analysis of basketball game performance based on bivariate poisson regression model, *Computer Modelling & New Technologies* **18**, 474–479.

Shi, Z., Moorthy, S. and Zimmermann, A.: 2013, Predicting NCAAB match outcomes using ML techniques-some results and lessons learned, *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*.

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

Stan Development Team: 2020, *R Package rstan: the R Interface to Stan*. R package version 2.21.1.
URL: *http://mc-stan.org/*

van de Water, H. and Bukman, C.: 2010, A balanced team-generating model for teams with less than nine persons, *IMA Journal of Management Mathematics* **21**(3), 281–302.

Van Rossum, G. and Drake, F. L.: 2009, *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.

Wickham, H.: 2019, *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.
URL: *https://CRAN.R-project.org/package=stringr*

Yang, J. and Lu, C.-H.: 2012, Predicting NBA championship by learning from history data, *Proceedings of Artificial Intelligence and Machine Learning for Engineering Design* .

Yesilyurt, E.: 2014, Determining the players' efficiency in NBA: some economic and managerial results, *International Journal of Sport Management, Recreation & Tourism*, **13**, 1–19.

Zhang, Z. and Chelladurai, P.: 2013, Antecedents and consequences of athlete's trust in the coach, *Journal of Sport and Health Science* **2**(2), 115–121.