**The Whole vs. the Sum of its Parts**

Basketball

1583

1. **Introduction**

Basketball is a team game, and it is a commonplace among NBA observers that sometimes teammates, like Stockton and Malone or Jordan and Pippen, mesh particularly well. On the other hand, even very talented players may in some cases be a poor fit for their teams, as in the recent cases of Josh Smith in his years with the Pistons, or Dwight Howard on the 2011-2012 Lakers. The question of team synergy, or when the whole exceeds (or falls short of) the sum of its parts, has been comparatively neglected in basketball analytics, even as individual metrics have become very sophisticated.

In this paper we will tackle these questions. The first goal of this paper is to quantify offensive and defensive synergy for pairs of teammates in the NBA. We do this by adding interaction terms to the regularized adjusted plus-minus model, as we will explain in detail below. The second goal of the paper is to refine Alagappan’s clustering analysis of NBA player types, and to examine whether or not certain player types, or playing styles, mesh very well or very poorly with each other on offense and defense.

* 1. **Data**

Our dataset consists of play-by-play data from NBAstuffer.com for NBA seasons beginning with 2004-2005 and ending with 2014-2015. For each possession of each season, this dataset includes the names of the offensive and defensive players on the court, the number of points scored, and certain event type classifications, such as “hook shot” or “blocking foul”.

* 1. **Reproducibility**

The NBAstuffer.com data our paper uses is proprietary. However, similar data for the period 2005-2012 is publicly available from BasketballValue.com. This dataset, along with code for replicating our results, is posted at:

(redacted to preserve anonymity)

1. **Relation to Previous Work**

Adjusted plus-minus (APM) was introduced in the context of the NBA by Dan Rosenbaum [3]. Regularized adjusted plus-minus (RAPM) was introduced by Joe Sill [4]. (ESPN’s Real Plus-Minus metric, according to ESPN, “reflects enhancements to RAPM” by Jeremias Engelmann, although the details are proprietary.)

Our clustering work is inspired primarily by Muthu Alagappan’s work on basketball positions [1]. However, we take the analysis further by using play-by-play data with about 120 unique event types, rather than the seven statistical categories considered by Alagappan.

1. **Regression**

**3.1 Methodology**

In this subsection we describe two models that we fit to the possession-by-possession scoring data described in Section 1.1, to learn the offensive and defensive skills of the players involved. The first model, which we designate the *additive model*, is simply regularized adjusted plus-minus (RAPM), which makes the assumption that players’ skills add together to determine the expected outcome of each possession. The second model, which we designate the *interactive model*, also includes an interaction term for each pair of teammates on each possession. Whereas the additive model uses 10 coefficients to estimate the expected outcome of each possession (one for each player on offense and defense), the interactive model uses 30 coefficients because there are 10 pairings of offensive teammates and 10 pairings of defensive teammates introduced into the model.

Before defining the models, we introduce some notation. Each player *j* has an offensive coefficient *βj* and a defensive coefficient *δj*. For example, *β*ChrisPaul represents Chris Paul’s offensive skill, and *δ*TonyAllen represents Tony Allen’s defensive skill. Additionally, for two players *j* and *k*, *β*{*j*,*k*} represents their chemistry on offensive while *δ*{*j*,*k*} represents their chemistry on defense. For example, *β*{StephenCurry,KlayThompson} represents the Splash Brothers’ chemistry on offense. We use *i* = 1, …, *n* to index the possessions within each season, and *Yi* denotes the number of points scored on the *i*th possession. The set *O*i comprises the five players on the floor for the team with possession of the ball, and the set *Di* comprises the five players on the floor for the team without possession. With that out of the way, we can move on to the models.

**3.1.1 Additive model (RAPM)**

The additive model assumes that *ηi*, the expected number of points scored on the *i*th possession, is the sum of the offensive and defensive skills of the players on the floor:

where *α* is an intercept term representing the league average for points scored per possession.

To estimate the coefficients in the above model, we use a technique called ridge regression, as is used in RAPM. Using boldface ***β*** to denote the vector containing all *βj*’s and boldface ***δ*** to denote the vector containing all *δj*’s, we fit the regression model by solving the optimization problem:

for some *λ* > 0. We choose *λ* to have minimum prediction error on out-of-sample possessions, through a procedure called cross validation. By solving the above optimization problem, we find the skill coefficients which best explain the data, subject to the restriction that the coefficients cannot be too large. This restriction leads to more robust estimates than ordinary least squares and is the difference between RAPM and adjusted plus-minus.

We solve the regression outlined above using the cv.glmnet() function in the R package glmnet. To significantly speed up the computation, we use the SparseMatrix construction from the R package Matrix for our design matrix, leveraging the fact that each row of the design matrix has only 10 nonzero entries.

Note that our assumption is *not* that the number of points scored on a possession follows a normal distribution, which of course is not the case. We only make an assumption on the expected number of points scored on each possession having a specific linear form. To be clear, the optimization problem that we solve corresponds to optimizing the likelihood under an implicit assumption of Gaussian noise, but we do not make any claim of optimizing over any likelihood. We only aim to make a close approximation to the expected number of points scored on each possession.

**3.1.2 Interactive model**

Our novel extension of RAPM introduces an interaction term for each pair of teammates on the floor together. The expected number of points scored on the *i*th possession is given by:

Once the linear model for the expected number of points scored on each possession is specified, the rest of the procedure is the same as for the additive model. Now boldface ***β*** represents the vector containing all individual *βj*’s *and* all pairwise *β*{*j*,*k*}’s, and boldface ***δ*** represents the vector containing all individual *δj*’s and all pairwise *δ*{*j*,*k*}’s. As before, the fitted regression coefficients are defined by:

As before, *λ* > 0 is chosen via cross validation. Again we use glmnet in R with the SparseMatrix representation. Now each row has 30 nonzero entries. The computation would not be feasible without the sparse representation.

**3.1.3 Interpretation**

It is trivial to prove that, due to the regularization that we use in ridge regression, by construction the average offensive coefficient is zero and the average defensive coefficient is zero. Hence if any player *j* were to be taken off the floor and replaced by an average player on offense, the expected points per possession would decrease by *βj*. So we can interpret 100 times *βj* as the offensive points added above average per 100 possessions for player *j*. This is how we will report the results in the following section. Similarly, 100 times *δj* is the defensive points added above average per 100 possessions for player j, so positive numbers correspond to bad defense and negative numbers correspond to good defense.

The interpretation of the interaction coefficients is similar. If players *j* and *k* have average offensive chemistry (*β*{*j*,*k*} = 0), then the value that they contribute together to the offense (in terms of points per possession) is *βj* + *βk*. However, if their chemistry is different from the average, then they contribute *βj* + *βk* + *β*{*j*,*k*} points per possession. So *β*{*j*,*k*} represents the offensive value that two teammates offer, *beyond the sum of their skills*. We interpret 100 times *β*{*j*,*k*} as the offensive points added above average per 100 possessions *attributable to the offensive chemistry* between players *j* and *k*. Similarly we interpret 100 times *δ*{*j*,*k*} as the defensive points added above average per 100 possessions attributable to the defensive chemistry between players *j* and *k*.

**3.2 Results**

The tables below show the largest coefficients we obtained, for both offense and defense, and for both positive and negative sign. The units are points per possession. The largest interaction term coefficients, which are around 0.03 points per possession in magnitude, represent approximately the difference between an average player and an elite player.

Good offense

|  |  |  |  |
| --- | --- | --- | --- |
| Player 1 | Player 2 | Season | Coefficient |
| Dirk Nowitzki | Jason Terry | 2010-2011 | 0.0330 |
| Damian Lillard | Nicolas Batum | 2013-2014 | 0.0327 |
| Damon Jones | Shaquille O’Neal | 2004-2005 | 0.0316 |
| Carmelo Anthony | Iman Shumpert | 2011-2012 | 0.0313 |
| Kevin Durant | Russell Westbrook | 2011-2012 | 0.0313 |

Bad offense

|  |  |  |  |
| --- | --- | --- | --- |
| Player 1 | Player 2 | Season | Coefficient |
| Draymond Green | Marreese Speights | 2013-2014 | -0.0322 |
| Josh Childress | Shelden Williams | 2006-2007 | -0.0279 |
| Bryon Russell | Nene Hilario | 2004-2005 | -0.0267 |
| Kenny Thomas | Willie Green | 2004-2005 | -0.0266 |
| Rodney Buford | Travis Best | 2004-2005 | -0.0263 |

Good defense

|  |  |  |  |
| --- | --- | --- | --- |
| Player 1 | Player 2 | Season | Coefficient |
| Rasho Nesterovic | Tim Duncan | 2005-2006 | -0.0396 |
| Gerald Wallace | Wesley Matthews | 2011-2012 | -0.0327 |
| Danny Granger | David Harrison | 2005-2006 | -0.0314 |
| Kawhi Leonard | Tiago Splitter | 2013-2014 | -0.0313 |
| Kawhi Leonard | Tim Duncan | 2012-2013 | -0.0261 |

Bad defense

|  |  |  |  |
| --- | --- | --- | --- |
| Player 1 | Player 2 | Season | Coefficient |
| Martell Webster | Trevor Booker | 2013-2014 | 0.0412 |
| Channing Frye | Gerald Green | 2013-2014 | 0.0306 |
| Leandro Barbosa | Steve Nash | 2007-2008 | 0.0274 |
| Javale McGee | Jordan Crawford | 2011-2012 | 0.0267 |
| Charlie Villanueva | Michael Redd | 2007-2008 | 0.0255 |

1. **Clustering**
   1. **Methodology**

Our goal was to cluster players into positions according to their playing styles, disregarding, in particular, their respective ability levels and nominal positions. We used the same play-by-play dataset for this task, and the total player events, including e.g. 3-point attempts, defensive rebounds, and dunks, over the course of a season were used as features. In order to account for differing playing time and skill levels, only per-minute numbers were considered, and only shot attempts, rather than makes, were included. Different shot types, e.g. 2-point jump shots, 3-point shots, hook shots, and lay-ups, were treated as distinct features.

Some events among the feature set, such as illegal assists and hanging technical fouls, occur with very low frequency during a season of basketball. In order to reduce the dimensionality of the problem and reduce the noise in the dataset, we limited the feature set to those that occur, on average, more than once per game (i.e. more than once per 480 player-minutes).

The second pre-processing step was Principal Component Analysis. The dimensionality of the problem was further reduced by restricting to the first seven principal components, which together account for 90% of the variance in the dataset.

The clustering was performed using K-medioids. In order to determine the optimal number of clusters, we considered two metrics. The first metric is the “split pair percentage” (SP%), defined as , where P is the total number of player pairs existing in season N and N+1, and CP is the total number of player pairs that are assigned to the same cluster both in season N and in season N+1. The SP% is a measure of temporal persistence of clusters, and by minimizing it we can expect to maximize the interpretability of clusters over several seasons. The second metric is the “degenerate cluster percentage” (DC%), defined as the percentage of clusters which have two or fewer players. Minimizing this metric helps us ensure that most clusters will be large enough to be meaningful. The K-medioids algorithm was run for K varying between 5 and 30, and the value K=18 was accepted since it minimized the sum of SP% and DC%. Two of the resulting eighteen clusters were degenerate.

* 1. **Output**

The output of the clustering algorithm is summarized in the following table:

|  |  |  |
| --- | --- | --- |
| Cluster No. | Qualitative Description | Notable Players |
| 1 | Scoring Combo Guards | Mo Williams, Jason Terry, Joe Johnson, Kirk Hinrich |
| 2 | Long Wings | Tayshaun Prince, Beno Udrih, Shaun Livingston, Luke Ridnour |
| 3 | Degenerate (Damion James) | |
| 4 | 3-and-D | Steve Blake, Chris Duhon, Martell Webster, Shane Battier |
| 5 | Shooters | Steve Novak, J.R. Smith, Kyle Korver, Ryan Anderson |
| 6 | Offensive First Options | LeBron James, Carmelo Anthony, Dwyane Wade, Kobe Bryant |
| 7 | Rim Protectors | Samuel Dalembert, Tyson Chandler, Andrew Bogut, Joakim Noah |
| 8 | Degenerate (Hassan Whiteside) | |
| 9 | Rebounders | Udonis Haslem, Brandon Bass, Juwan Howard, Elton Brand |
| 10 | Traditional Big Men | Zaza Pachulia, Kwame Brown, Nene, Chris Andersen |
| 11 | Versatile Perimeter Players | Boris Diaw, Luke Walton, Josh Childress, Mike Conley |
| 12 | Polished Big Men | Carlos Boozer, Zach Randolph, David West, Tim Duncan |
| 13 | Interior Players | DeSagana Diop, Jason Collins, Joel Anthony, Nick Collison |
| 14 | Defensive Wings | Tony Allen, Andrei Kirilenko, Andre Miller, Ramon Sessions |
| 15 | Scoring Bigs | Dwight Howard, Shaquille O’Neal, Kevin Love, Eddy Curry |
| 16 | Volume Scorers | Richard Hamilton, Caron Butler, Tracy McGrady, Andrea Bargnani |
| 17 | Offensive Wings | Manu Ginobili, Chauncey Billups, Jamal Crawford, Deron Williams |
| 18 | Shooting Forwards | Matt Barnes, Rasheed Wallace, Ersan Ilyasova, Mike Dunleavy |

The “Notable Players” are selected as the players with most seasons in each cluster, while the “Qualitative Description” is established by looking at the players that appear most frequently in each cluster and extracting common trait based on common basketball judgment.

**4.3 Clusters and Regression Coefficients**

In order to evaluate whether certain player types mesh well with another specific player type, the average interaction coefficients were calculated for each pair of clusters and are shown in the Figure below. In both cases blue color indicates poor synergy and red color indicates good. The cluster numbering corresponds to the numbering in the Table shown in Section 4.2. Perhaps the most important takeaways revolve around Cluster #6 which includes most of the league’s superstars. It can be observed that this cluster tends to have poor defensive synergy with Scoring Combo Guards (#1), Offensive Wings (#17) and Scoring Forwards (#18). This conclusion is also fairly intuitive, as offensive specialists tend to be less good on defense. On the other hand, the league’s superstars justify their reputation on offense where they make their teammates better regardless of the other players’ style of play.

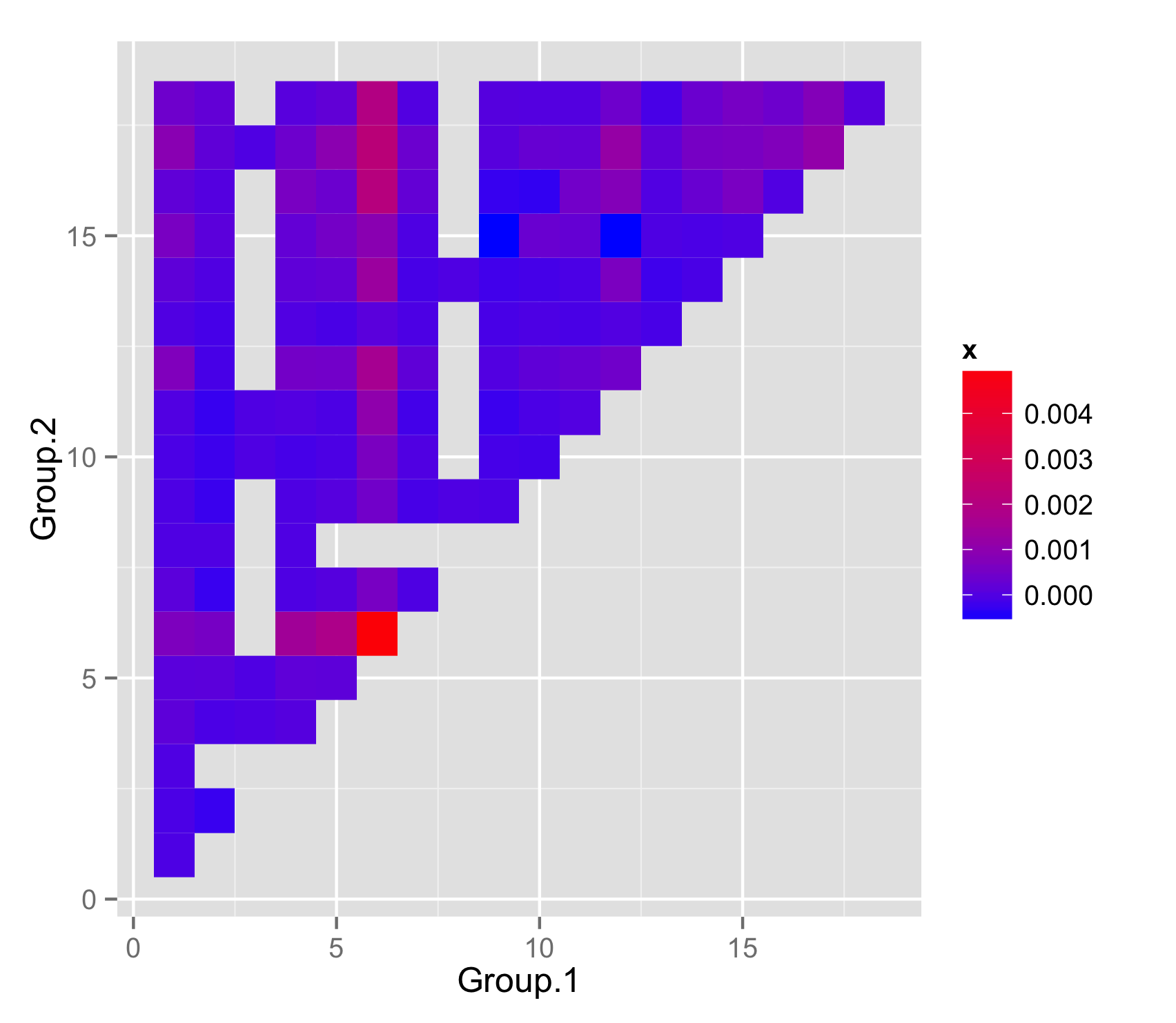
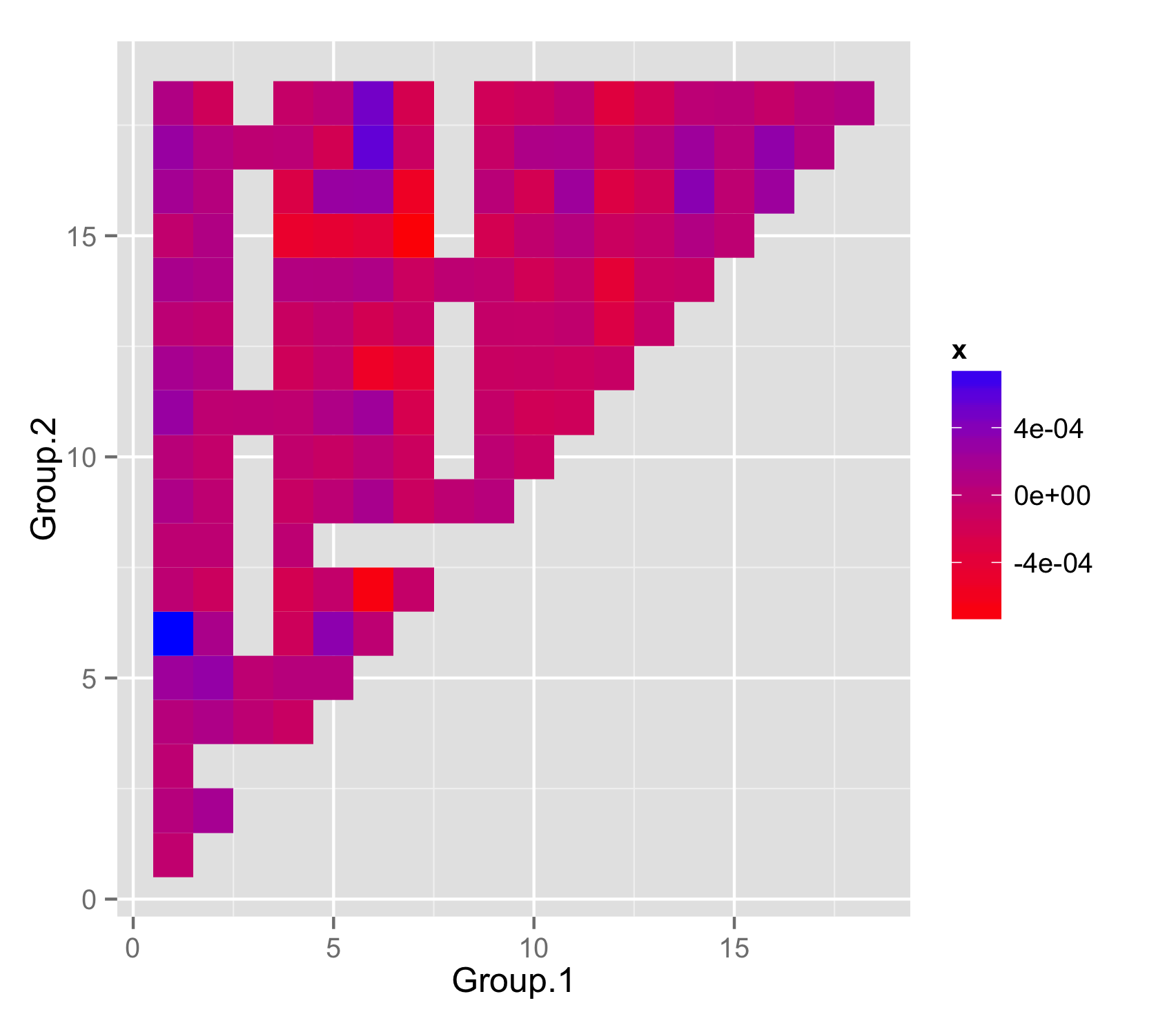


Figure . Average interaction coefficients between clusters on defense (left) and offense (right)

1. **Conclusion**

**5.1 Predictive Power**

A natural question is whether or not introducing interaction terms increases the predictive power of the model. Unfortunately, we determined that the answer is no, using the following experiment. Within each season we conducted an experiment by splitting the data into a training set (games played between November and February, inclusive) and a test set (games played in March). We fit each model to the training data, using 10-fold cross-validation to select the regularization parameter. The results in the table below give the percentage increase in predictive power over the naïve strategy of simply predicting the overall average for the number of points scored in each possession.

|  |  |  |
| --- | --- | --- |
| Year | RAPM | RAPM with interaction terms |
| 2004-2005 | 0.179% | 0.160% |
| 2005-2006 | 0.066% | 0.078% |
| 2006-2007 | 0.123% | 0.131% |
| 2007-2008 | 0.061% | 0.056% |
| 2008-2009 | 0.148% | 0.145% |
| 2009-2010 | 0.105% | 0.088% |
| 2010-2011 | 0.128% | 0.124% |
| 2011-2012 | 0.125% | 0.128% |
| 2012-2013 | 0.088% | 0.082% |
| 2013-2014 | 0.123% | 0.137% |
| 2014-2015 | 0.119% | 0.126% |

So, RAPM with interaction terms outperforms RAPM in 5 of 11 seasons, and there is no evidence for a significant difference in predictive power between the two models.

**5.2 Applications**

However, we believe that our techniques have great inherent descriptive interest. They could very realistically be useful for NBA teams attempting to identify sub-optimal lineups, or to optimize rotations. More tentatively, they could also be useful in the creation of strategy for roster building, or in identifying free agents who might mesh well or poorly with a given roster.

**References**

[1] Alagappan, Muthu. From 5 to 13: Redefining the Positions in Basketball. Sloan 2012.

[2] Omidiran, Dapo. A New Look at Adjusted Plus/Minus for Basketball Analysis. Sloan 2011.

[3] Rosenbaum, Dan. Measuring How NBA Players Help Their Teams Win. www.82games.com/comm30.htm

[4] Sill, Joe. Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing.