Inferring Latent Representations of Professional Basketball Players

Sameer Lal
December 21, 2015

1 Introduction

Last year the National Basketball Association (NBA) experienced one of the most popular and prosperous seasons in the league's history. Significantly outgrowing its humble origins, the sport has become a world-wide phenomenon. As competition in the NBA has intensified over the years, teams are exploring all options to gain an edge. As a result, the strategy of the sport (how to construct a team, which plays to run, what shots to take, etc.) has also evolved drastically.

Traditionally, tactical decisions were made with intuition and knowledge of the game. Many coaches and team managers were once players themselves. Yet the use of technology to track players' movements and actions on the court has generated rich data sets to leverage. With the deluge of data, many teams are using players' statistics to facilitate decision making and analysis.

A team is not just the sum of its parts. There are many intricacies to the on court dynamics that can dictate a team's success. Usually, a player's *individual* statistics have been used to analyze their ability and predict performance with potential teammates. Yet this information may be limited in capturing features which may indicate how well players fit with each other. Instead, I have modified matrix factorization to model line up data, which only records how a cumulative group of players performed over the course of a season. This highly accurate model can better infer how well players perform together, resulting in a novel method to develop strategy in the sport.

2 Related Works

There have been many attempts to isolate the true value of a player and how well he fits with their team. An advanced set of metrics, such as player efficiency rating (PER) or wins above replacement player (WARP), were invented with the intention of determining just how much a player was affecting their team's performance. At a higher level, these metrics are mostly averages for positive and negative events players record during the course of a season. More complex statistical models have been applied to predicting outcomes in basketball but fail to infer player skill and team dynamics. [1] use a dependent probabilistic matrix factorization model to attempt predicting game outcomes using side information about game venue and time. [2] utilizes spatial location player shot attempt data to summarize shooting habits with a non-negative matrix factorization. My proposed model effectively leverages a different data set to extract information which could ultimately help team personnel decisions and on court strategy.

3 Data

The data set provided at `basketball-reference.com`, has line-ups of size 2 (dyadic), 3, 4, and 5 players for all teams in the NBA over the past 10+ years. For each data point, there exists a unique set of players,$\{a, b, c, d, e\}$, and a set of metrics which describe their performance when playing together: points, minutes played, shooting, etc. In order to collect all this data from the website, a crawler was written in *JavaScript* which can be run using *Node.js*. Line ups of all sizes have been collected for the previous NBA season (2014-2015). The advantage of the crawler is that it can easily scrape data for even earlier seasons, theoretically reducing variance in inference estimates. But the league is constantly changing. Players develop, experience injuries, trades are made, and coaches are replaced. There could be several hidden factors which are not accounted for that affect the data season to season. Therefore only a single season was used for this analysis.

4 Models

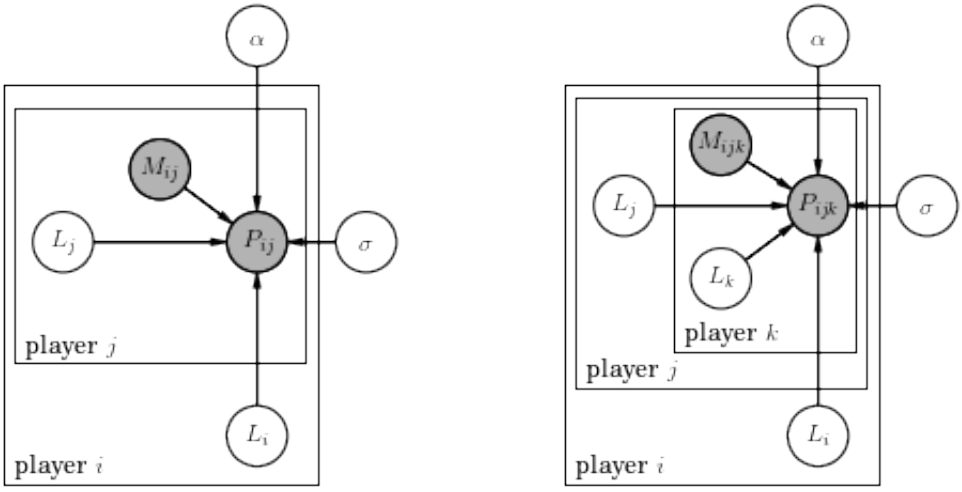

Figure 1: Graphical Model diagrams for 2 man (*left*) and 3 man (*right*) line up models.

4.1 Probabilistic Matrix Factorization

To model the dyadic data (2-man line ups), I used probabilistic matrix factorization (PMF) with uniform priors (Figure 1). The model is described in [3], with some minor adjustments for this particular problem. Instead of users and movies, suppose there are T players in the data set and their hidden representations, L_t, are R-dimensional feature vectors. P_{ij} is the observed points differential (how many more points this line up scores over its opposition per 100 possessions) for the two players (players i and j) in the line up. M_{ij} is the number of minutes (normalized) that line up i, j played together during

the course of the season. The minutes observations are used to scale variance to more accurately model noisy observations of P_{ij} as a result of very little playing time. Due to the nature of the game, $P_{ij} = P_{ji}\ and\ M_{ij} = M_{ji},\ i \neq j$. The conditional distribution of the observed data is modeled as follows:

$$p(P_{ij}|\alpha, \sigma, L, M_{ij}) = \mathcal{N}(\mu_{ij}, \lambda_{ij}) \tag{1}$$

$$\mu_{ij} = L_i L_j^T + \alpha \tag{2}$$

$$\lambda_{ij} = \frac{\sigma}{M_{ij}} \tag{3}$$

4.2 Larger Models

To accommodate the larger line ups (3-man, 4-man, and 5-man), the 2-man PMF model was further modified. As displayed in Figure 1, the observations P_{ijk} are dependent on L_i,L_j,L_k (latent player representations), and M_{ijk}. As expected, M_{ijk} is the amount of time players i,j, and k played together over the course of the season. Similar to the dyadic case, $P_{ijk} = P_{jik} = P_{kji}\ and\ M_{ijk} = M_{jik} = M_{ikj},\ i \neq j \neq k$. Considering that $L_{i,r}$ is the r^{th} component of player i feature vector, the conditional distribution of the observed data is modeled as follows:

$$p(P_{ijk}|\alpha, \sigma, L, M_{ijk}) = \mathcal{N}(\mu_{ijk}, \lambda_{ijk}) \tag{4}$$

$$\mu_{ijk} = \sum_{r}^{R} \left(L_{i,r} L_{j,r} L_{k,r}\right) + \alpha \tag{5}$$

$$\lambda_{ijk} = \frac{\sigma}{M_{ijk}} \tag{6}$$

The models for 4-man and 5-man line ups were extended in the same fashion. See Appendix for exact descriptions of conditional distributions.

4.3 Implementation in *Stan*

For inference, the dyadic model was implemented in the probabilistic programming language, *Stan*. Initially, since most of the data was pre-processed using `Python`, the model and inference were computed in *PyStan*. Upon attempting to sample the posterior distribution using HMC, the inference algorithm failed. HMC struggles with matrix factorization due to there being multiple posterior modes. In order to overcome this issue the models were implemented in *CmdStan* and successful inference was accomplished using ADVI. In order to ensure proper inference was occurring some testing was done with simulated data (discussed further in the Appendix). The code for this work can be found at `https://github.com/sl3368/NBAPlayerVectors`.

2 Man Line Up Model (N=4000)				
R-Value	*25*	*50*	*75*	*100*
ME	2070	76038	207355	290778
MSE	2.9	117.3	489.7	839
MPE	144250	4162750	8424425	13798450
MPSE	142	6020	24018	40800

3 Man Line Up Model (N=15000)				
R-Value	*25*	*50*	*75*	*100*
ME	88.6	158.2	591.2	2241.4
MSE	0.053	0.0599	0.151	0.533
MPE	7860	15384	58710	232300
MPSE	1.5	1.84	5.4	19.4

4 Man Line Up Model (N=26000)				
R-Value	*25*	*50*	*75*	*100*
ME	107.26	124.8	172.7	291.6
MSE	0.0188	0.021	0.022	0.027
MPE	8280	9980	14890	25530
MPSE	0.5	0.5	0.6	0.8

5 Man Line Up Model (N=16000)				
R-Value	*25*	*50*	*75*	*100*
ME	140	323.2	1150.5	3736.8
MSE	0.009	0.013	0.032	0.099
MPE	12440	35720	126080	420842
MPSE	0.17	0.4	1.1	3.4

Figure 2: Cross validation results for all models of different line up size and vector length (R). ME is the mean error. MSE refers to mean scaled error-$ME * (normalized\ minutes\ played)^2$. MPE is mean percent error. MPSE is mean percent scaled error-$MPE * (normalized\ minutes\ played)^2$.

5 Results

5.1 Evaluation of Models

Models were evaluated by prediction accuracy on held out line ups. 10-fold cross validation was performed for all models at varying representation vector lengths (values of R). For all sizes of line ups, a large majority of the data was line ups which played very few minutes. Since the aim is to provide a model which can predict the performance of a line up which might play significant amounts throughout the season (a likely strategical question), scaled mean error and scaled mean percent error metrics were measured.

5.2 Empirical Findings

The cross validation results, displayed in Figure 2, show a highly accurate model when considering scaled error (and percent scaled error). Although for 2 man line ups, model accuracy is quite low, the larger models exhibit very good performance on held out prediction. The various models improve in accuracy as line ups increase in the number of minutes play (Figure 4), a desirable trend. Furthermore there is no bias towards predicting different types of line ups, poor or strong, evident in Figure 6 (Appendix).

One of the reasons the larger line up models are much more accurate at prediction on held out data than the 2 man line up model is the nature of a held out instance. In a held out dyadic instance, the combination of the players has not been "seen" by the model. For 4 or 5 man line ups, although the exact combination of players has not been trained upon, the model has learned from how subsets of the players in the line up have interacted with other players. Thus the training data for the larger line ups is a better estimator for latent player representations. In any case, teams usually are looking to make changes of only one or two players in a 5-man line up, corroborating the effectiveness of these models.

Rank	Player	Euclidean Distance of t-SNE Vectors
3	T. Chandler	1.65
7	R. Hibbert	2.19
9	B. Wright	2.61
10	A. Davis	2.62

Figure 3: Similarity to Dwight Howard using Euclidean Distance.

5.3 Interpretation of Latent Space

In addition to accurately predicting line ups, the inferred latent representations of players should intuitively mimic their profiles and playing styles observed on the court. The representation space was visualized in a two dimensional space using t-SNE. As seen in Figure 8, there is slight clustering around groups of players that perform well together (for instance the starting line up of the Cleveland Cavaliers in Figure 7). To further

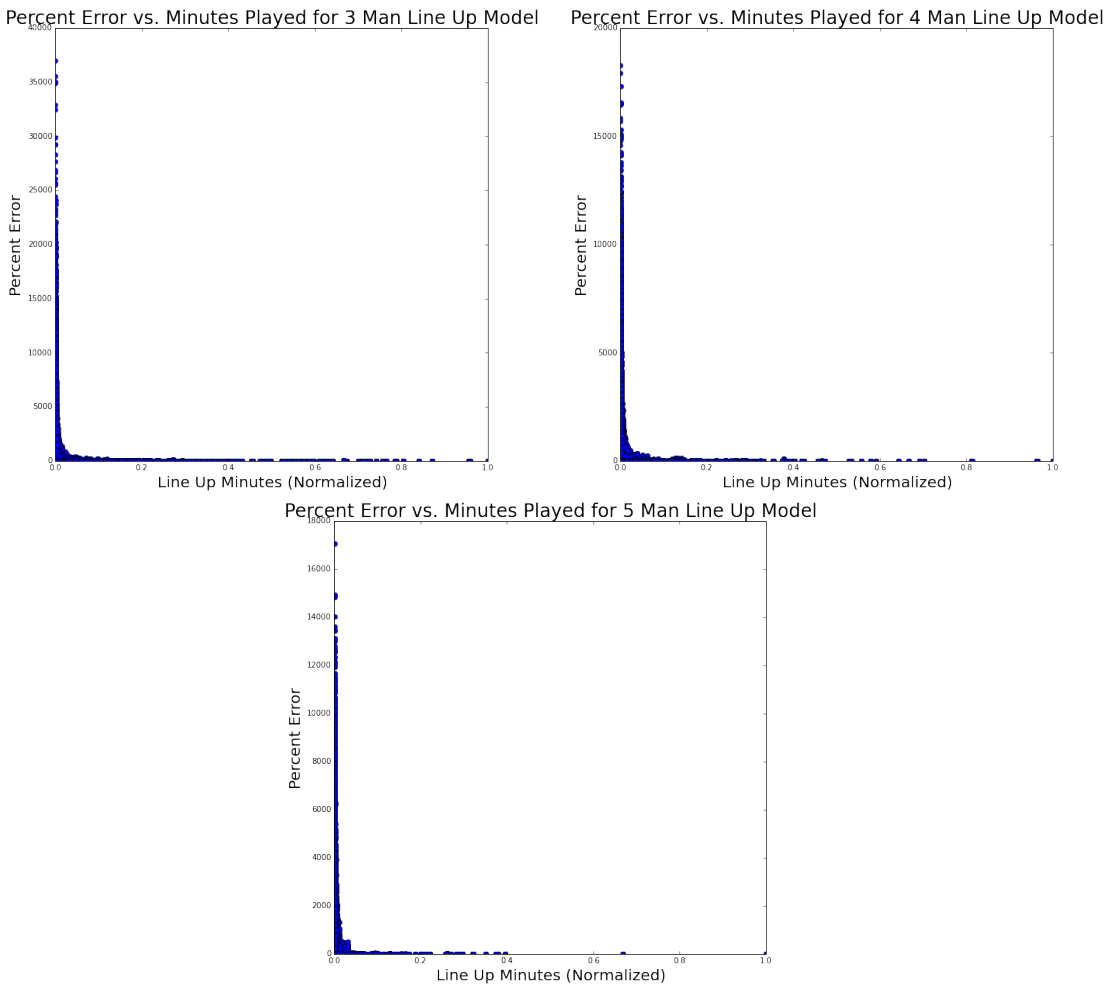

Figure 4: The accuracy of models improves for held out line ups which play more minutes.

validate the model's inference, an inspection of nearest neighbors for well known players using euclidean distance results in highlighting players of similar type or playing style, which would fit well with line ups that would include that player. For example, Dwight Howard is a defensive minded center, using euclidean distance we find some other defensive minded players centers shown in Figure 3.

6 Conclusion

Predicting how well groups of professional basketball players will perform together is a challenge many teams are facing. Constantly searching for how to improve themselves through different tactics and personnel moves, NBA franchises are seeking to develop statistical models to better inform their strategic decisions. Alternative to modeling *individual* player data, latent representations of players were inferred using an extension of PMF. The resulting models for line ups of size 3, 4, and 5 were highly accurate

in predicting performance of held out line ups of players. Furthermore, the inferred representations of these players exhibit some knowledge of player type and skill set.

An interesting extension of this work is to incorporate other information about line up performance, such as rebounds, pace, and assists and estimate these metrics as a multivariate Gaussian. This could result in an even more informative model, that could help teams improve on a single objective (defensive or rebounding), a common occurrence.

References

[1] Ryan P. Adams, George E. Dahl, and Iain Murray. "Incorporating Side Information into Probabilistic Matrix Factorization Using Gaussian Processes". In: *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*. Catalina Island, California, July 2010. URL: http://hips.seas.harvard.edu/files/w/papers/adams-dpmf-uai-2010.pdf.

[2] Andrew Miller et al. "Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball". In: *Thirty-First International Conference on Machine Learning (ICML)*. June 2014.

[3] Ruslan Salakhutdinov and Andriy Mnih. "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo". In: *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*. 2008, pp. 880–887. DOI: 10.1145/1390156.1390267. URL: http://doi.acm.org/10.1145/1390156.1390267.

Appendix

4 Man Model Description

P_{ijkl} is the point differential of players i,j,k, and l over the course of the season. M_{ijkl} is the normalized amount of time players played together over the course of the season. Similar to the 3 man case, $i \neq j \neq k \neq l$. Considering that $L_{i,r}$ is the r^{th} component of player i feature vector, the conditional distribution of the observed data is modeled as follows:

$$p(P_{ijkl}|\alpha, \sigma, L, M_{ijkl}) = \mathcal{N}(\mu_{ijkl}, \lambda_{ijkl}) \tag{7}$$

$$\mu_{ijkl} = \sum_{r}^{R} \left(L_{i,r}L_{j,r}L_{k,r}L_{l,r}\right) + \alpha \tag{8}$$

$$\lambda_{ijkl} = \frac{\sigma}{M_{ijkl}} \tag{9}$$

5 Man Model Description

P_{ijklm} is the point differential of players i,j,k, l, and m over the course of the season. M_{ijklm} is the normalized amount of time players played together over the course of the

season. Similar to the 4 man case, $i \neq j \neq k \neq l \neq m$. Considering that $L_{i,r}$ is the r^{th} component of player i feature vector, the conditional distribution of the observed data is modeled as follows:

$$p(P_{ijklm}|\alpha, \sigma, L, M_{ijklm}) = \mathcal{N}(\mu_{ijklm}, \lambda_{ijklm}) \tag{10}$$

$$\mu_{ijklm} = \sum_{r}^{R} \left(L_{i,r} L_{j,r} L_{k,r} L_{l,r} L_{m,r}\right) + \alpha \tag{11}$$

$$\lambda_{ijklm} = \frac{\sigma}{M_{ijklm}} \tag{12}$$

Simulations

In order to ensure that variational inference was performing with decent accuracy, tests were conducted with simulated data. Code was written to create data using the generative process described by the model. Several tests seem to indicate good accuracy. As an example, Figure 5 displays the accuracy in inferring the hidden representation of 100 fictional players with simulated data.

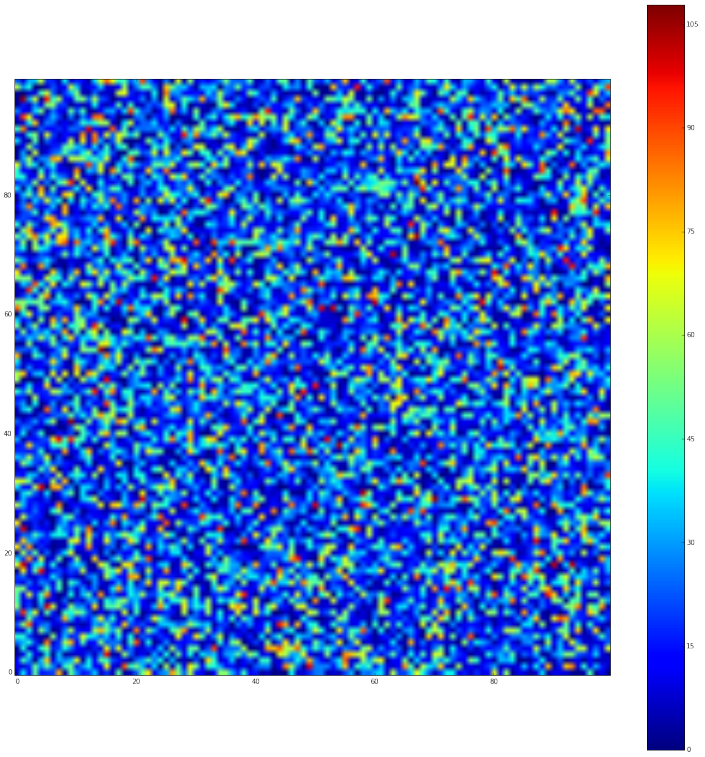

Figure 5: Absolute value of difference between the real values for hidden variables L and inferred values $\hat{L}$, ($|L - \hat{L}|$) using variational inference in *Stan*. Dark blue is zero.

Additional Figures

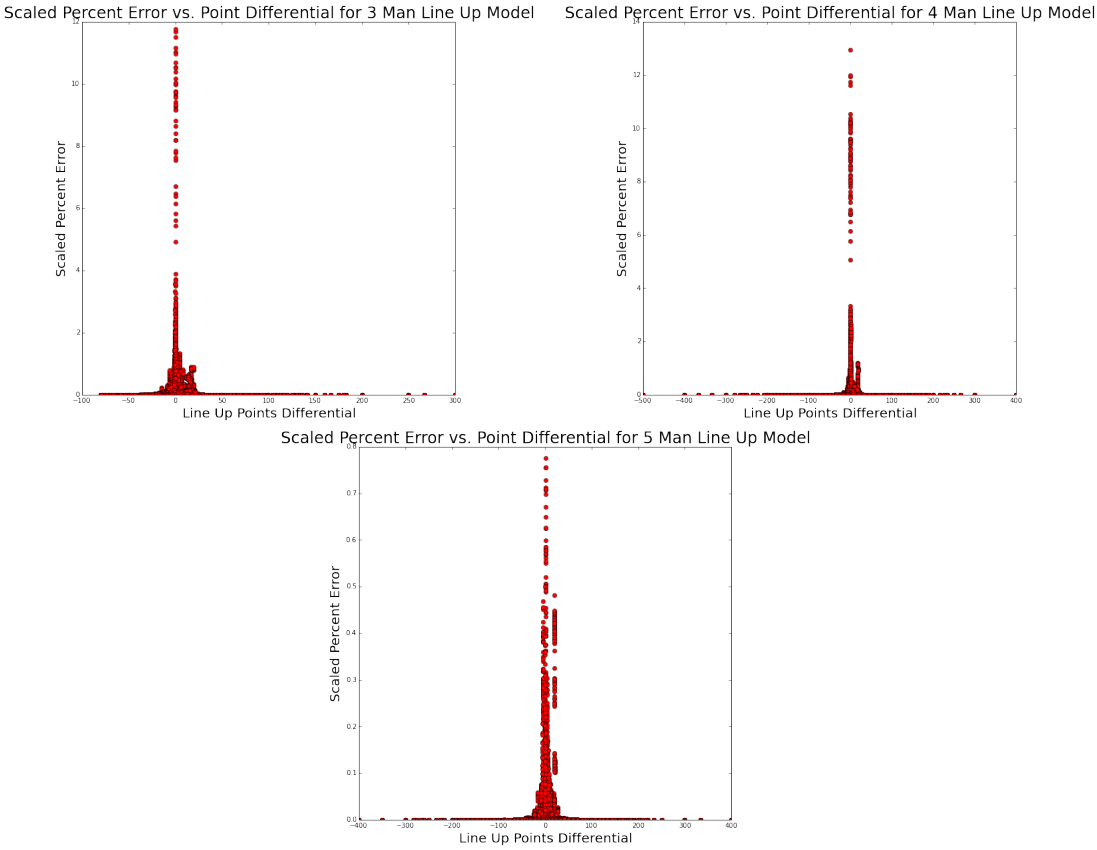

Figure 6: The accuracy of models is normally distributed over strong and weak line ups.

Rank	Player	Euclidean Distance of t-SNE Vectors
1	S. Marion	0.37
2	K. Love	1.21
3	I. Shumpert	1.34
4	K. Irving	1.45

Figure 7: Shortest Euclidean Distance of t-SNE vectors to LeBron James. This is more or less the starting rotation for the team.

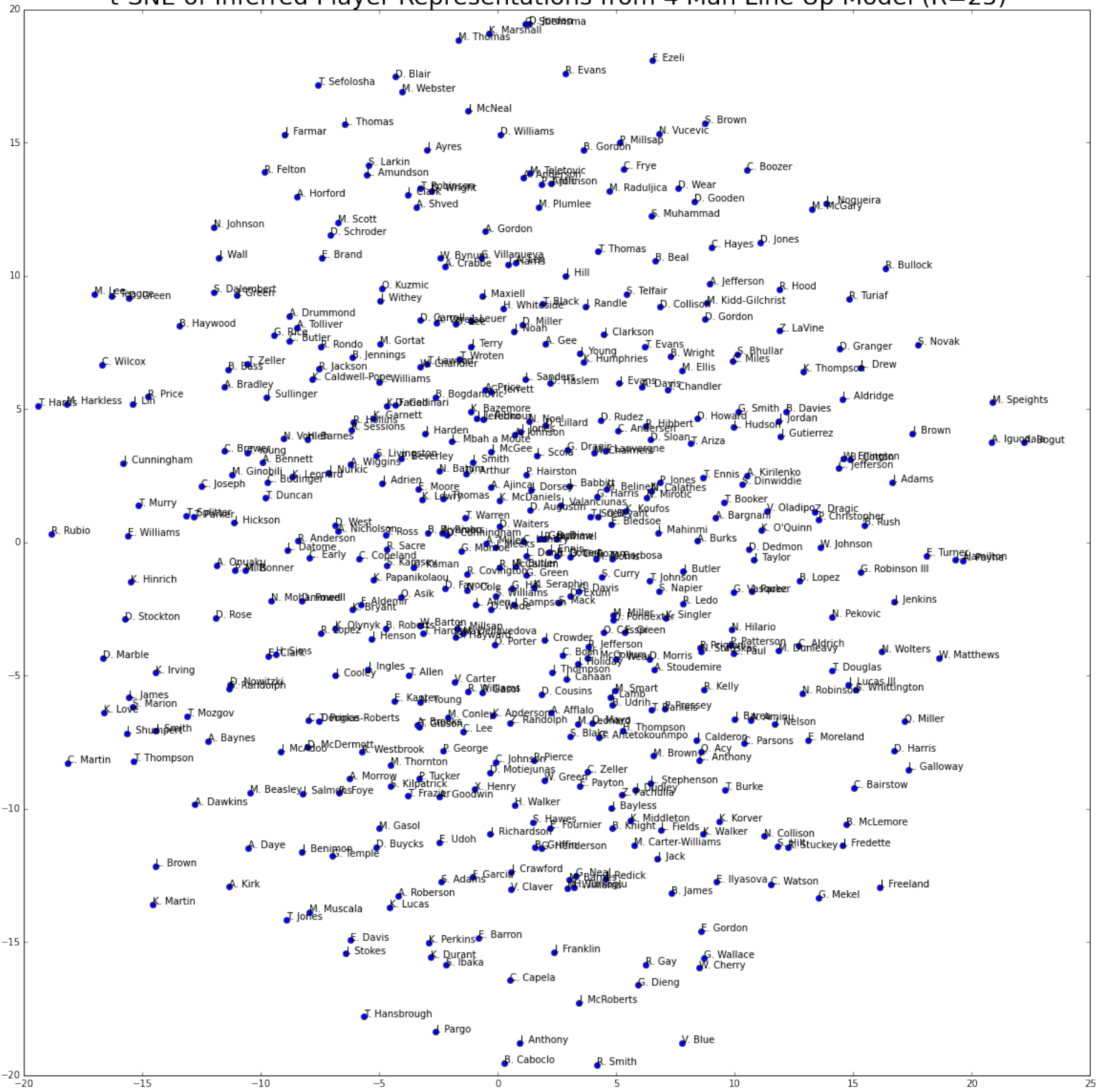

Figure 8: The t-SNE of the latent representations for players exhibits some interesting clustering.