Spatial Analysis of Basketball Game Using Machine Learning methods

Jicheng Lu, TEXAS A&M UNIVERSITY

Instructor: Xiaoning Qian

ABSTRACT: Machine learning techniques are widely needed in analyzing the basketball data. In this project, we apply an unsupervised protocol to evaluate the shooting habits of different players. We first model the spatial shot data as a point process, Log-Cox Gaussian Process, and use non-negative matrix factorization to decompose the intensity surface into two low-rank matrices: weight matrix and basis matrix. The results clearly summarize the shooting habits of different NBA players.

KEYWORDS

Gaussian process, Poisson process, LGCP, Non-negative matrix factorization.

1 INTRODUCTION

NBA (National Basketball Association) is the most popular men's professional basketball league in the world. With the growing influence of the league, the basketball sports has become not only an athletic event but also a comprehensive research area which involves data mining, data analysis, etc. Thus, numerous studies have been carried out to analyse the game data and further obtain a better understanding of the sports.

Given the large quantity of the games played in the past a few decades, a very rich and detailed dataset has been established and several advanced indexes have been calculated (such as the Four Factors [1]) based on the raw data. However, it is still a great challenge to evaluate the basketball games with such a large dataset, and therefore, this is where the machine learning methods should be implemented over the data.

Many of the related researches lie in the game prediction. Shi et al. [1] predicted the result of each individual NCAA games using various machine learning algorithms, such as C4.5, Ripper, Multi-Layer Perceptron, Naive Bayes and Random Forest. Puranmalka [2] established a model with different feature selection algorithms for the prediction of a single match with the involvement of both team-level and player-level data. Torres [3] compared three methods (i.e. linear regression, maximum likelihood and multi-layer perceptron) in the prediction accuracy of NBA games. Another research interest lies in the unsupervised learning of the player behaviors. Miller et al. [4] used point process modelling and dimension reduction method to evaluate the shot habits of different basketball players. Chen et al. [5] applied and

proved the effectiveness of a spatio-temporal model learning method for studying the group behaviors from basketball videos. Other studies focused on the improvement of the defensive effectiveness by identifying the different features [6] and prediction of the basketball trajectory using deep learning [7].

In this project, the objective is to apply an unsupervised protocol (i.e. a combination of point process modelling and dimensionality reduction method) to better evaluate the shot selection of different players.

2 METHODOLOGY

2.1 Gaussian Process

A Gaussian process (GP) is a stochastic process whose sample path is normally distributed. GP is frequently used as a probabilistic model over functions: $f : \chi \to R$, where the function value $f_n = f\left(x_n\right)$ corresponds to a function evaluation at some point $x_n \in \chi$. A key property of the Gaussian process is that it can be completely defined by its second-order statistics. Thus, the Gaussian process behavior is completely determined by the covariance function if it is assumed to have zero mean. The covariance function can define the basic aspects of the process, including smoothness, periodicity, and stationarity.

2.2 Poisson Process

A Poisson process (PP) is a completely spatially random point process on some space, χ, for which the number of points that end up in some set $A \subseteq \chi$ is Poisson distributed. In this project, a set of spatial points, $x_1, x_2, \ldots, x_n$, is modeled as a Poisson process with a non-negative intensity function $\lambda\left(x\right)$. Moreover, the positive property of intensity functions can lead to the non-negative decomposition of a global Poisson process (i.e. non-negative matrix factorization) into simpler weighted sub-processes.

2.3 Log-Gaussian Cox Process

A log-Gaussian Cox process (LGCP) is a doubly-stochastic Poisson process with a spatially varying intensity function modeled as an exponential Gaussian process (GP):

$$\begin{aligned} Z\left(\cdot\right) &\sim GP\left(0, k\right) \\ \lambda\left(\cdot\right) &\sim \exp\left(Z\left(\cdot\right)\right) \\ \mathrm{x}_1, \ldots, \mathrm{x}_N &\sim PP\left(\lambda\left(\cdot\right)\right) \end{aligned} \tag{1}$$

where doubly-stochastic refers two levels of randomness: the random function Z and the random point process with intensity λ.

2.4 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a dimensionality reduction technique that decomposes a non-negative matrix of the product of two low-rank matrices:

$$\Lambda \approx W \cdot B \tag{2}$$

where the target matrix $\Lambda_{N\times V}$ is composed of N data points of length V, the basis matrix $B_{K\times V}$ is composed of K basis vectors, and the weight matrix $W_{N\times K}$ is composed of the N non-negative weight vectors that linearly combine the basis vectors to approximate Λ. Each vector of the intensity function can be reconstructed from the weights and the bases:

$$\lambda_n = \sum_{k=1}^{K} W_{n,k} \cdot B_k \tag{3}$$

The optimal matrices W^{opt} and B^{opt} are determined by an optimization procedure that minimizes the loss function ℓ, which is a measure of reconstruction error or divergence between Λ and WB with the constraint that all elements in these matrices are non-negative:

$$W^{opt}, B^{opt} = \underset{W,B\geq 0}{\arg\min} \left\{ \ell\left(\Lambda, WB\right) \right\} \tag{4}$$

where ℓ is the Kullback-Leibler (KL) divergence when $\sum_{i,j} X_{i,j} = \sum_{i,j} Y_{i,j} = 1$.

$$\ell_{KL}\left(X, Y\right) = \sum_{i,j} X_{\mathrm{i},j} \log \frac{X_{i,j}}{Y_{i,j}} \tag{5}$$

Note that the KL loss function includes a log ratio term. This tends to disallow large ratios between the original and reconstructed matrices.

3 PROPOSED APPROACH

3.1 Proposed procedure

In this project, I combine the point process (LGCP) and dimensionality reduction (NMF) methods to analyse the shooting data of different players. The proposed approach is presented in Figure 1.

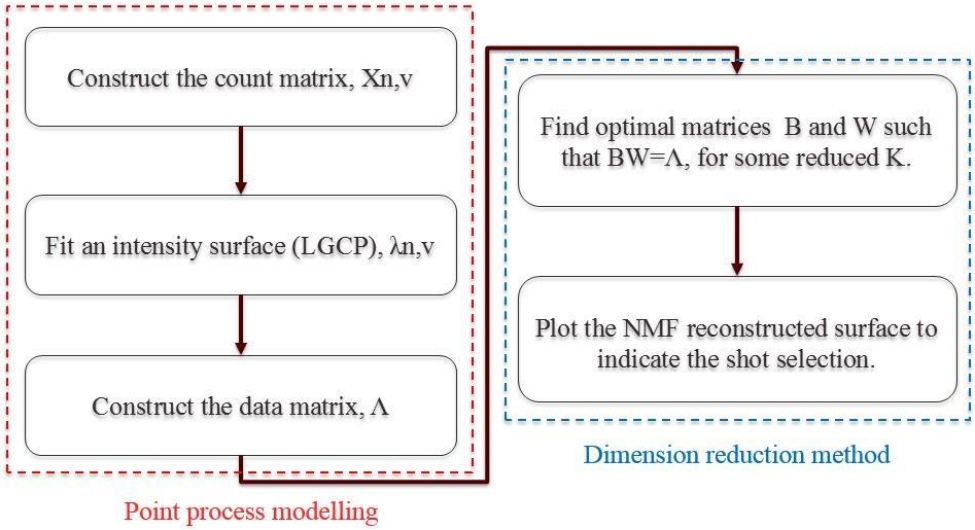


Figure 1. Proposed workflow in this project.

Given a set of shooting data of N players, the procedure is as follows:

a) Discretize the basketball court into multiple rectangular elements, and construct the count matrix $X_{n,v}$, which is the number of shots (both made and missed shots) by player n in element v;

b) Fit the intensity surface $\lambda = \left(\lambda_{n,1}, \lambda_{n,2}, \cdots, \lambda_{n,v}\right)$ for each player over the discretized court (LGCP).

c) Normalize the intensity surface to obtain the target matrix $\Lambda = \left(\bar{\lambda}_1, \bar{\lambda}_2, \cdots, \bar{\lambda}_N\right)$, where $\sum \bar{\lambda}_i = 1.$

d) Minimize the KL loss function and find optimal W and B for a reduced rank K such that $\Lambda \approx W \cdot B$, with non-negative constraint (NMF).

3.2 Fit the intensity surface

For each player's set of data points, x_n, the likelihood of the point process is discretely approximated as Poisson distribution:

$$p\left(x_n \mid \lambda_n\right)=\prod_{v=1}^{V} p\left(X_{n,v} \mid \lambda_{n,v}\right)=\prod_{v=1}^{V} e^{-\lambda_{n,v}} \frac{{\lambda_{n,v}}^{X_{n,v}}}{X_{n,v}!} \tag{6}$$

where λ_n is the discretized intensity function. This approximation comes from the spatially random property of the Poisson process, allowing us to treat each element independently. The probability of the count present in each element is Poisson with an intensity ($\lambda_{n,v}$).

According to the Log-Cox Gaussian Process (LGCP), the intensity λ_n is an exponential function of a Gaussian-distributed random parameter z_n. Thus, the posterior becomes

$$p\left(z_n \mid x_n\right) \propto p\left(x_n \mid \mathrm{z}_n\right) p\left(z_n\right)=\prod_{v=1}^{V} e^{-\lambda_{n,v}} \frac{{\lambda_{n,v}}^{X_{n,v}}}{X_{n,v}!} N\left(z_n \mid 0, K\right) \tag{7}$$

$$\lambda_n=\exp\left(\mathrm{z}_n+\mathrm{z}_0\right) \tag{8}$$

where z_0 is the bias term and the prior $p\left(z_n\right)$ is a mean zero normal with covariance K, determined by the squared exponential covariance function:

$$K_{\mathrm{i,j}}=k\left(x_i, x_j\right)=\sigma^2 \exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\phi^2}\right)+\sigma_{noise}^2 \delta_{ij} \tag{9}$$

where σ^2 is the marginal variation, $\left(x_i, x_j\right)$ are the two spatial locations on the court, σ_{noise}^2 is the noise covariance, δ_{ij} equals to one when i = j, ϕ is the length scale, which determines the smoothness of the surface.

When maximizing the posterior and fitting the intensity function, we will discover that there is a high correlation induced by the court's spatial structure. Hence, in order to

overcome the strong dependency between variables, we use the elliptical slice sampling method [8] to approximate the posterior of λ_n for each player. The elliptical slice sampling method is a novel Markov chain Monte Carlo algorithm with characteristics of simplicity, robustness and no free parameters. We iteratively slice this model until the convergence criterion is satisfied:

$$\frac{\left\| z^{t+1} - z^{t} \right\|}{\left\| z^{t+1} \right\|} < \varepsilon \tag{10}$$

where ε is a small value (e.g. 10^{-5}).

3.3 NMF Optimization

Once we obtain the intensity surface, we need to find the optimal matrices W^{opt} and B^{opt} by minimizing the KL loss function. In this project, we use the iterative NMF algorithm to search the two optimal lower-rank matrices [9,10]. The iteration procedure is given as follows:

$$\begin{aligned} W_{ia} &\leftarrow W_{ia} \sum_{m} \frac{\Lambda_{im}}{(WB)_{im}} B_{am} \\ W_{ia} &\leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}} \\ B_{am} &\leftarrow B_{am} \sum_{i} W_{ia} \frac{\Lambda_{im}}{(WB)_{im}} \end{aligned} \tag{11}$$

4 RESULTS AND DISCUSSION

4.1 Data resource

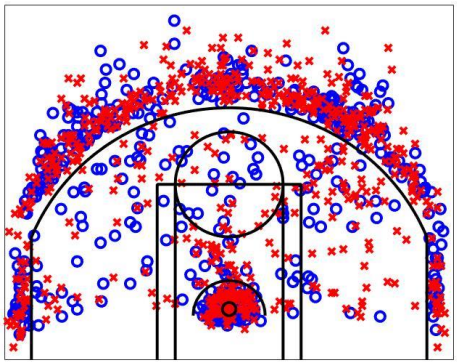
The shooting data consists of made and missed field goal attempt locations from the games in 2016-2017 NBA regular season (http://nbasavant.com/shot_search.php). In this project, we select three excellent players (Stephen Curry, James Harden and Kawhi Leonard) to analyze their shooting habits. Some raw data is presented in Figure 2.

4.2 Results

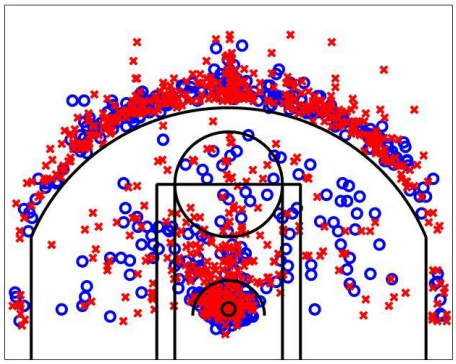
We first present the original shot location points on a half court among three different players in Figure 2. The blue circles represent the made shots while the red crosses represent the missed shots.

Based on the original data, we discretize the half court into 49×49 elements and calculate the number of shots in each rectangular element. Figure 3 depicts the discretized court, while Figure 4 shows the count maps for each player.

Figure 5 presents the LGCP surface (i.e. intensity surface) of each player. Note that compared to the count maps in Figure 4, the LGCP surfaces can smooth the discretized count data and better evaluate the shot selection of different players. Setting the reduced rank $K = 10$ and using the iterative algorithm for minimizing the KL loss function, we reconstruct the intensity surface and present in Figure 6. Although all the three players frequently shoot at the paint area close to the basket, we can still discover that there is a variation in shot selection among the three basketball players. Stephen Curry specializes the three points at all locations, including the top, two wings and two corners. James Harden spends many shot attempts at the top area. However, Compared to the other two players, Kawhi Leonard can shot from both the three point area and the mid-range area. Table 1 gives the errors between the LGCP surfaces and the NMF reconstructed surface.

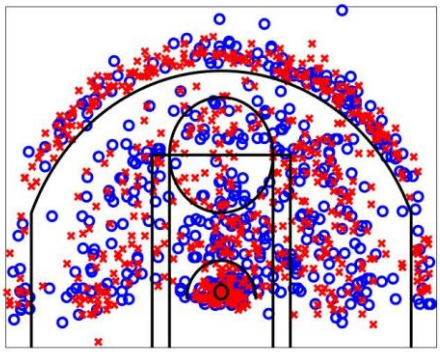


(a) Stephen Curry (1238 shots)

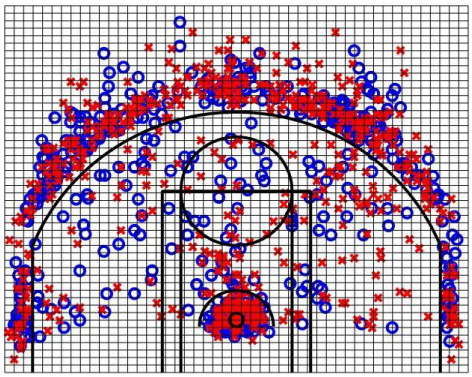


(b) James Harden (1296 shots)

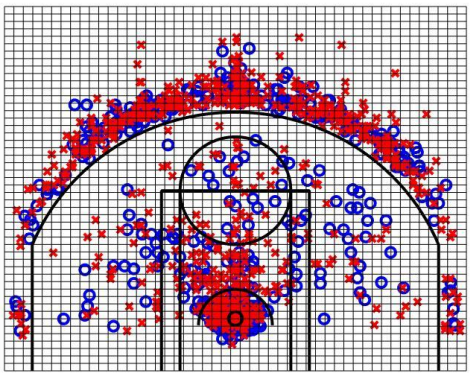
- Jicheng Lu



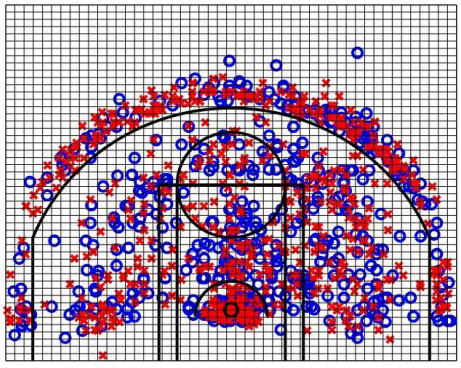
(c) Kawhi Leonard (1163 shots)
Figure 2 Original data points



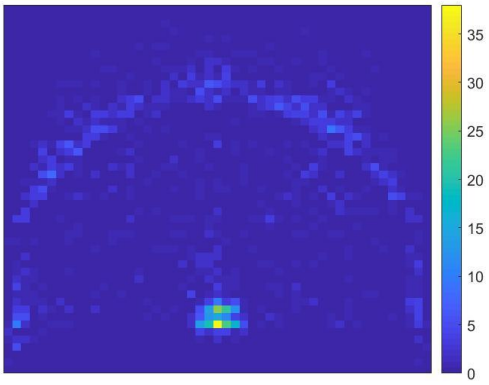
(a) Stephen Curry (1238 shots)



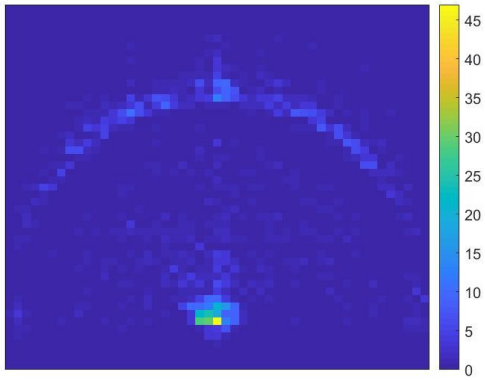
(b) James Harden (1296 shots)



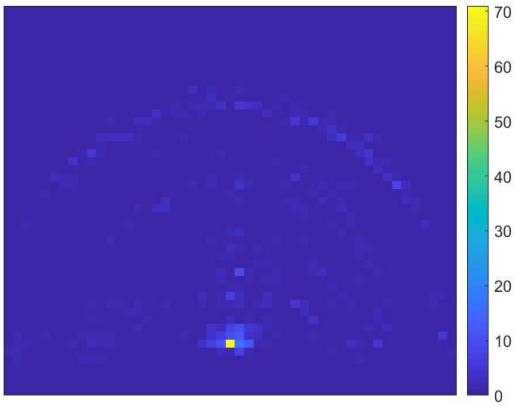
(c) Kawhi Leonard (1163 shots)
Figure 3 Discretized elements



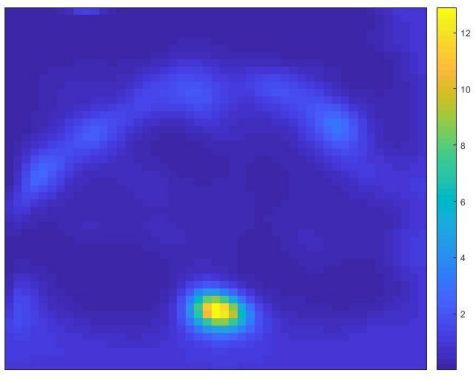
(a) Stephen Curry



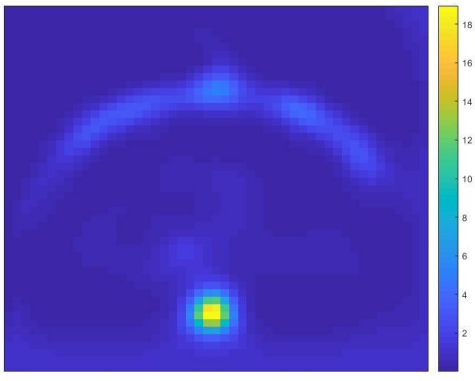
(b) James Harden



(c) Kawhi Leonard
Figure 4 Count map

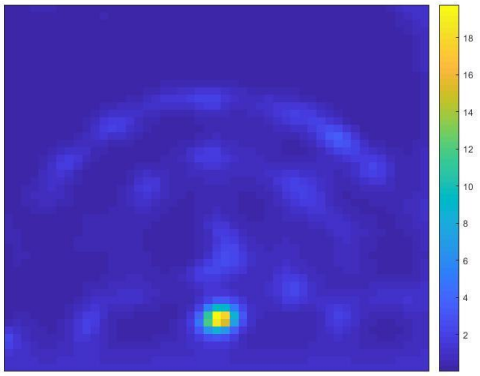


(a) Stephen Curry

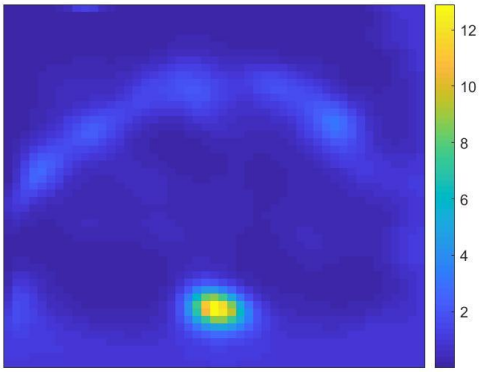


(b) James Harden

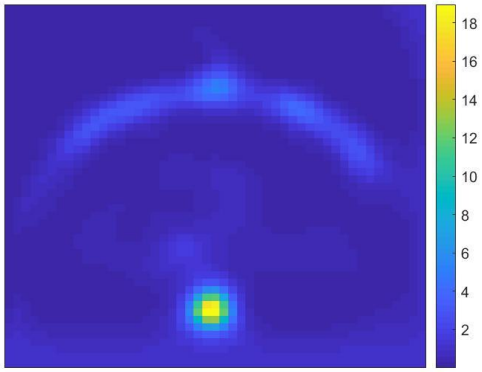
- Jicheng Lu



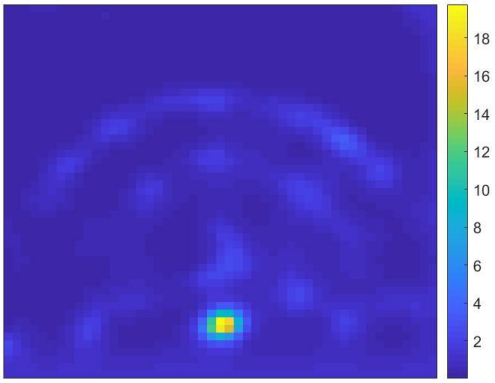
(c) Kawhi Leonard
Figure 5 LGCP surface



(a) Stephen Curry



(b) James Harden



(c) Kawhi Leonard
Figure 6 LGCP-NMF reconstructed surface

Table 1. Errors between the LGCP surfaces and NMF reconstructed surface

Player	Error
Stephen Curry	7.85×10^{-10}
James Harden	8.40×10^{-9}
Kawhi Leonard	7.00×10^{-9}

5 CONCLUSIONS

In this project, we apply an unsupervised protocol to evaluate the shooting habits of different players. We first model the spatial shot data as a point process, Log-Cox Gaussian Process, which is a doubly-stochastic Poisson process with a spatially varying intensity function modeled as an exponential Gaussian process. Using the elliptical slice sampling method and maximizing the posterior, we obtained the intensity surface based on the count matrix. We then use iterative NMF algorithm to minimize the KL loss function and decompose the intensity surface into two low-rank matrices: weight matrix and basis matrix. The results clearly summarize the shooting habits of the three excellent NBA players. The basis matrix needs to be further studied in order to quantify the shooting habits of different players.

REFERENCES

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

[2] Puranmalka, K., 2013. *Modelling the NBA to make better predictions* (Doctoral dissertation, Massachusetts Institute of Technology).

[3] Torres, R.A., Prediction of NBA games based on Machine Learning Methods. University of Wisconsin Madison.

[4] Miller, A., Bornn, L., Adams, R. and Goldsberry, K., 2014, January. Factorized point process intensities: A spatial analysis of professional basketball. In *International Conference on Machine Learning* (pp. 235-243).

[5] Chen, C.H., Liu, T.L., Wang, Y.S., Chu, H.K., Tang, N.C. and Liao, H.Y.M., 2015, October. Spatio-Temporal Learning of Basketball Offensive Strategies. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1123-1126). ACM.

[6] Block, A., Machine Learning Examination of NBA Defense.

[7] Shah, R. and Romijnders, R., 2016. Applying deep learning to basketball trajectories. *arXiv preprint arXiv:1608.03793*.

[8] Murray, Iain, Adams, Ryan P., and MacKay, David J.C. Elliptical slice sampling. Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS), 9:541–548, 2010.

[9] Lee, D.D. and Seung, H.S., 2001. Algorithms for non-negative matrix factorization. In *Advances in*

neural information processing systems (pp. 556-562).

[10] Brunet, J.P., Tamayo, P., Golub, T.R. and Mesirov, J.P., 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, *101*(12), pp.4164-4169.