Some Simple Linguistics of Ice Hockey

Denis Bashkirov

August 27, 2018

1 Introduction

We perform a simple statistical analysis of NHL Hockey games in seasons 2010-2017 and find statistically significant distinction in the game styles of different NHL teams based on their 'vocabulary' of attacks on goalposts in home games.

2 Representation of a game

We have information about approximately 10,000 games – all games played in NHL in seasons 2010-2017, both in regular seasons and playoffs. All the data was taken from the website *hockeystats.ca* in the form of couples of json files –lists of events and lists of shots coordinates and subsequently reworked in Python.

The data for each of the game is that of attacks on goalposts: time of the shot, its coordinates on the rink, its type – whether it was blocked, missed, saved or a goal was scored and the number of players of each team on the rink at the moment of the shot. The initial data was not perfect, however which resulted in discarding about 15 percent of the games, and in the remaining 85 percent about 10 shots (which is a very small fraction of all shots) per team could not be labeled by any type.

All this information about a game can be represented pictorially which we do in Figures 1,2,3 for the last NHL game to date – the final game of the playoff 2017 between Washington Capitals and Vegas Golden Knights, which ended with Capitals winning the Stanley Cup.

In this representation a game is a collection of lists of length five $[t, x, y, type, number]_A$ corresponding to attacks of the away team A and a similar collection of lists $[t, x, y, type, number]_H$ describing attacks of the opponent (home) team H. Here t is the time of the attack, (x, y) are its coordinates on the rink (with goalposts located at points $(\pm 90, 0)$), *type* is one of the four letters $('G', 'B', 'M', 'S')$[1] (goal, blocked, missed, saved) and *num* is the difference between the number of players of the away and home teams.

We can view a game as a text written by two authors – teams A and H. The text is the chronologically ordered set of attacks represented by the length-five lists which can be viewed as letters.

The questions we would like to address here is whether, given a collection of texts by the same author(s), we can, based on the used words, reliably guess the author. In fact, we will use a coarse-grained version of the text. First, we neglect all information about the positions of the attacks. Second, we will look at only (roughly), half of the attacks – all those by one of the team. Of course, a priori, the pattern of the attacks should depend on both the attacking and defending team, but for simplicity, we will neglect the contribution of the defending team. Next, we introduce the neutral letter Θ which corresponds to no attack.

Then the following part of the text, for example,

$$...'B'\Theta\Theta'M'\Theta'B'\Theta... \tag{1}$$

corresponds to a blocked shot which happened at a moment t_0 seconds, followed by a missed shot in three seconds (at $t_0 + 3$ seconds), and another blocked shot after one more second (at $t_0 + 5$ seconds) – all from one of the teams.

In fact, we will coarse-grain this description even further by ignoring the difference between types of shots, so the the text becomes just a sequence of ones and zeros with ones standing for attacks and zeroes –

[1] For rare shots its type is not contained in the data at our disposal, so we label them with $'N'$.

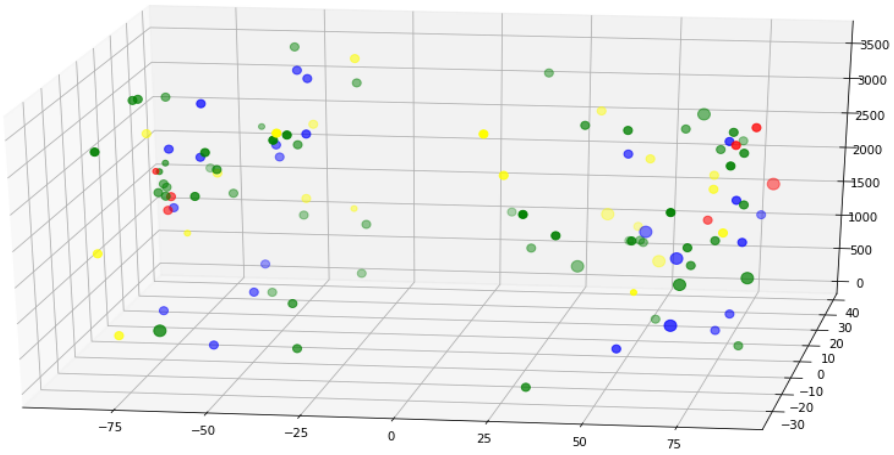

Figure 1: 3D representation of the game. Z-axis is time in seconds. Red–goal. Blue–blocked. Green–saved. Yellow–missed. Size–number of players difference.

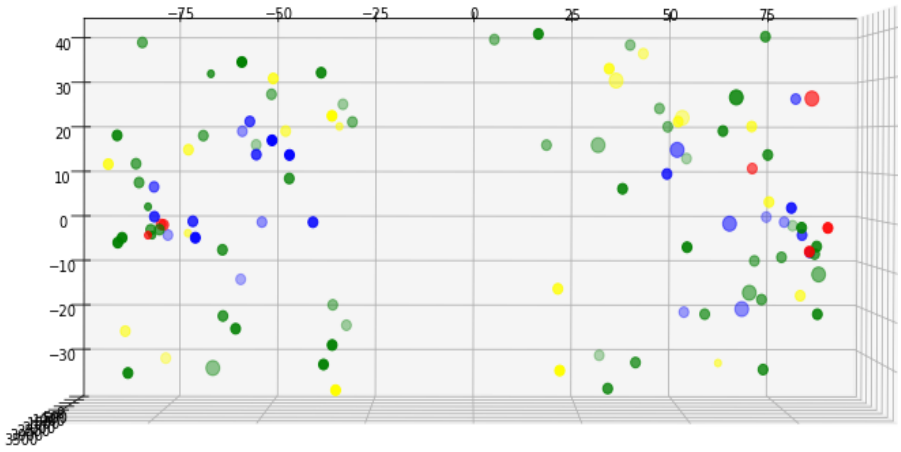

Figure 2: Projection to the rink.

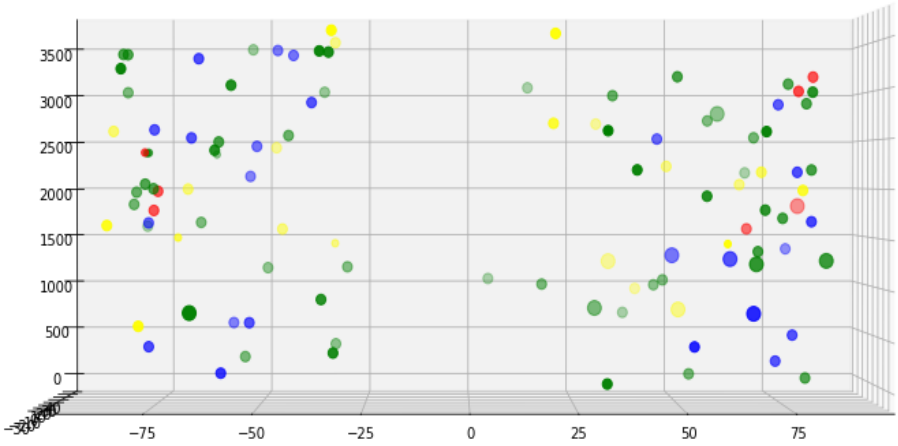

Figure 3: Time evolution of the game.

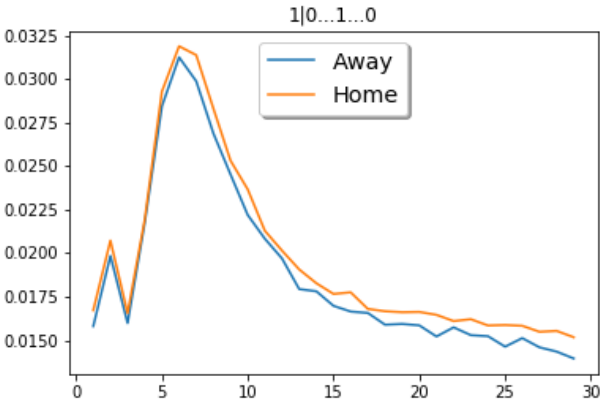

Figure 4: Conditional probabilities for $N = 30$.

for their absence. Then the previous example becomes

$$...1001010... \tag{2}$$

We do not know a priori if we neglected too much of the information to be able to tell apart styles of different teams. It will turn out that we did not. Given a collection of 20 texts (roughly, half of home games in a regular season) corresponding to home games, there are teams which can be reliably distinguished, and with a collection of 60 games, all of them can be by their 'vocabulary' when one looks at frequencies of only THREE words: $(2 * 1)$, (11) and (101). Here the first word corresponds to two attacks within a second, the second word – to two attacks separated by one second, and the third – two attacks separated by two seconds.

3 Independent words, conditional probabilities and the correlation length of the game.

Before discussing different authors/teams and difference in their style/voacbulary, one has to make sure that there is a style/vocabulary at all in our context.

The most trivial and uninteresting case would be when individual letters (in our case zeroes and ones) are independent random variables. In such a situation it does not make sense to speak of words – it is not a useful/meaningful concept.

To find out if it is our case, we look at sequences (0...0...1...01) of length N containing two ones and $N - 2$ zeroes and compute conditional probabilities according to its definition:

$$P(1|0...1...0) = \frac{P(0...1...01)}{P(0...1...0)}. \tag{3}$$

If letters are independent random variables the result should be a constant value $P = P(1) = q = 1 - p$ (the probability of one).

The results for $N = 30$ and $N = 60$ are shown in figures 4 and 5.

Obviously, we are not in the situation of statistically independent letters. Not only the conditional probabilities show dependence on positions of the previous one, the resulting profile is very interesting: it is not monotonous with a local minimum at $N = 3$ and a maximum at $N = 6$. Furthermore, in the second figure we see that close to $\Delta T = 60$ – when that two letters are separated by about a minute, the conditional probability becomes approximately constant. In other words, the two attacks separated by about a minute become statistically independent. Thus we conclude that the correlation length of Hockey games is around one minute:

$$T_{corr} \approx 60s \tag{4}$$

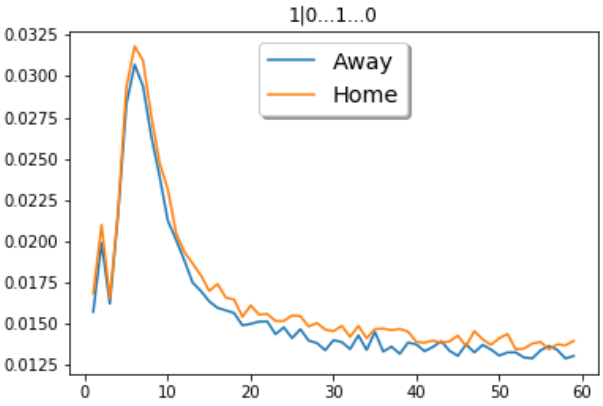

Figure 5: Conditional probabilities for $N = 60$.

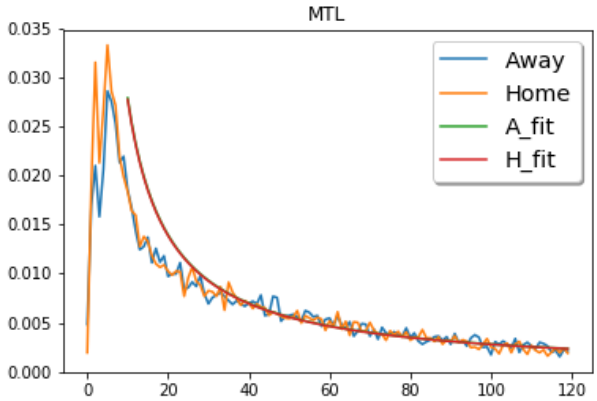

Figure 6: P(N) for Montreal Canadians.

In addition, these two figures show statistically significant difference between the away and home games – the separation between the two plots in each figure is obviously larger than fluctuations in probabilities within each of them. The conclusion is that home games favor more short words of length a minute and less than away games.

Next, let us take a look at probabilities of words of length N – sequences of the form (10...01) ($N-1$ zeroes between two ones.)

If the probability of zero is p and that of one is, correspondingly, $q = 1 - p$, the probability of a word of length in the assumption of statistical independence of letters would be

$$P(N) = p^{N-1}q = p^{N-1}(1-p) \tag{5}$$

This time, let us look at two teams: the Montreal Canadians and the current Stanley Cup champion Washington Capitals.

Again, we clearly see that the letters are not statistically independent, and the concept of words as correlated 'chunks' of letters should be useful in our context. In contrast to words in natural languages, here they are not 'sharp' as they are not, in the first approximation , statistically independent if they happen within one minute from each other. Nevertheless, we can find their frequencies and compare between various teams.

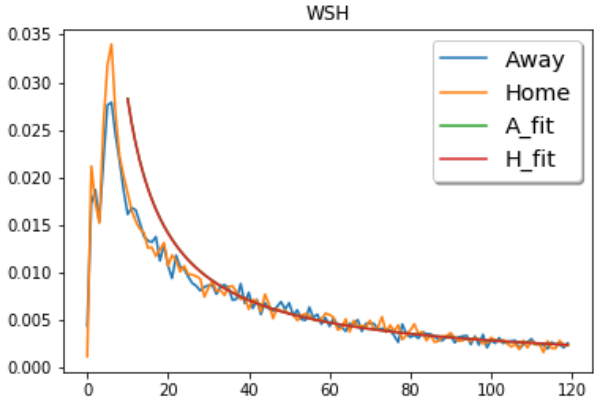

Figure 7: P(N) for Current champion Washington Capitals.

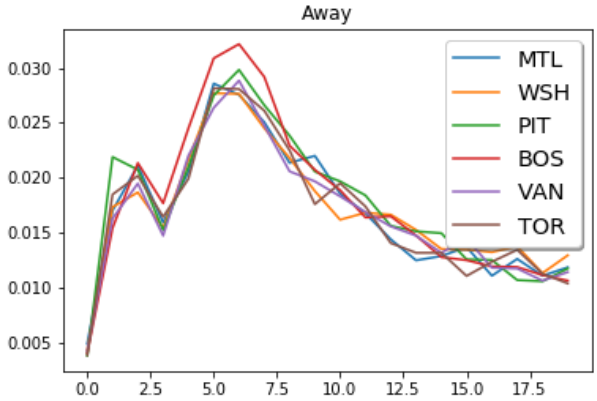

Figure 8: P(N). Away games for six teams.

Note that the values of probabilities are close to those of conditional probabilities considered above. This is due to a small value of probability of an attack – of order 10^{-2}.

4 Home vs Away games

As Figures 8,9 show, there is quite a significant difference between the games played at home and those played away. Namely, there is a much greater variance in the teams' vocabularies in the former case – they show more individuality. In later chapters we will see that more explicitly in 3d plots with probabilities of different length words taken as features.

One should mention here that the Figure 9 for the home games by itself does not warrant the conclusion that there is a statistically significant difference in style of the six teams – this could just be a statistical fluke due to a relatively small number of home games played by each team during the eight seasons considered (around 250 games each).

To see if there is a statistically significant difference, one can break down the sets of the home games for each team into a number of nonintersecting subsets (to guarantee their independence), and check if they tend to be closer to each other than those of other teams. In other words, to check if the variance in the feature vectors (whose coordinates are given by probabilities) of each team is noticeably less than the variance of

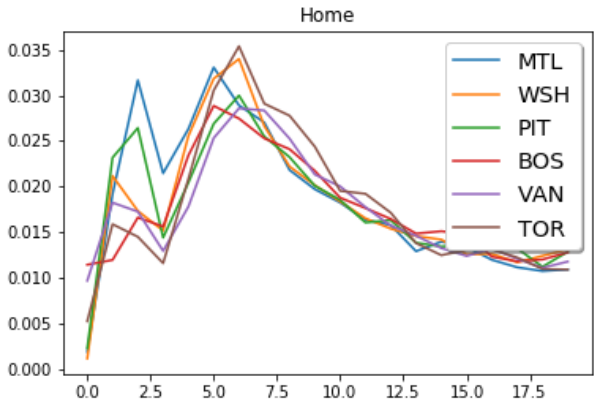

Figure 9: P(N). Home games for six teams.

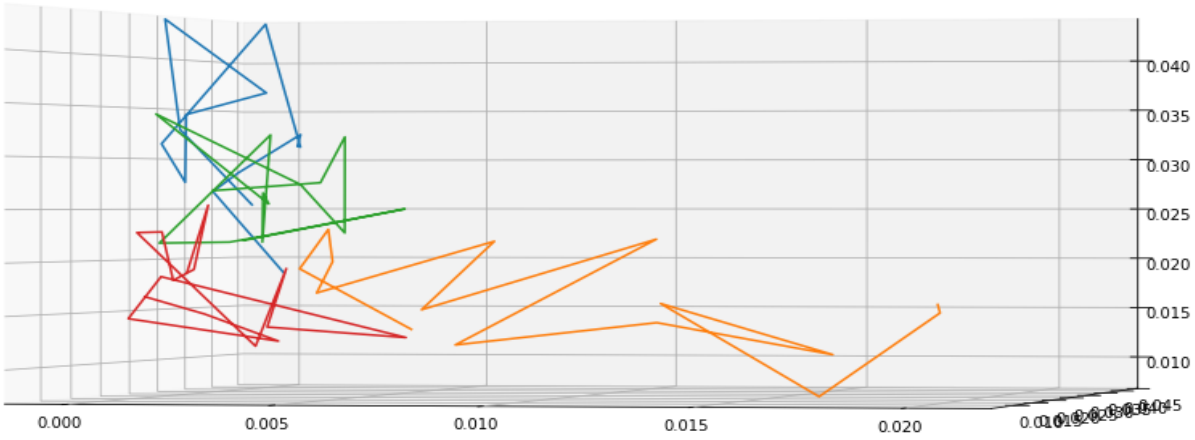

Figure 10: 'MTL', 'BOS', 'PIT' and 'WSH'. Projection in (x_0, x_1, x_3) .

the entire set of the feature vectors. We will find this to be the case (Figures 10, 11, 12, 15, 16).

5 Vocabularies of several teams

Consider a sub-vocabulary consisting of only words of three words: length zero $(2 * 1)$ (two attacks within a second), length one (11) (time interval between the two shots is between one and two seconds) and length two (101).

Each point of the Figures 10, 11,12 corresponds to 20 games, and none of the points contain a common game. Hence all points are independent. A priori, in the absence of distinct styles of different teams one would expect to find all points to be randomly distributed in some region of the three dimensional space. It is not what the pictures show!

The most important thing in these figures is that instead of being randomly distributed over the entire allowed region in the (x_0, x_1, x_2) space, the four teams tend to occupy small subregions hardly intersecting each other.

The next thing to notice is that the yellow team (it is 'BOS') shows a lot more variation along, essentially, a line in parallel to the (x_0, x_1) plane than the others. What is important here is that this is not a random variation – we see that this a drift with time, as the plots are time-ordered.

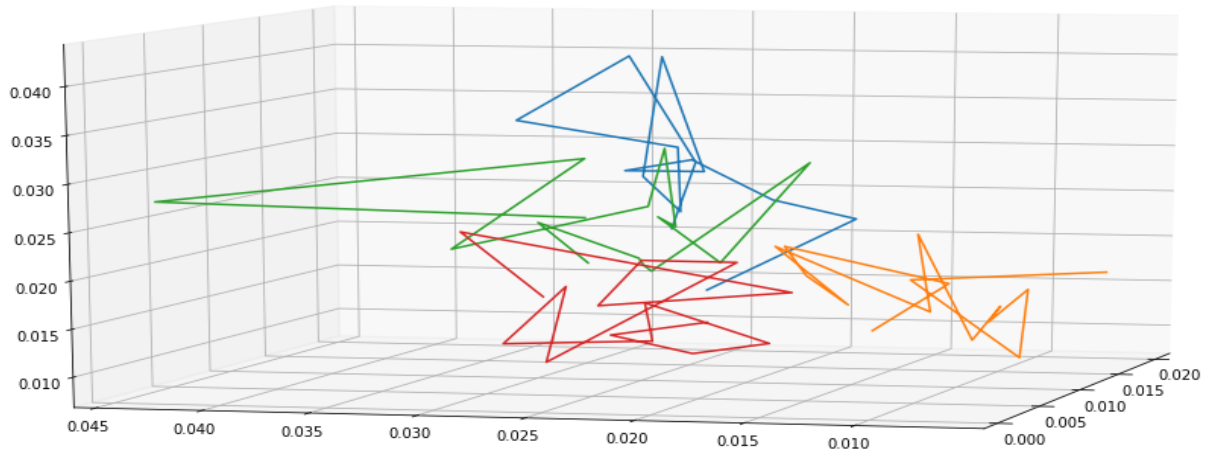

Figure 11: 'MTL', 'BOS', 'PIT' and 'WSH'. Another projection in (x_0, x_1, x_3).

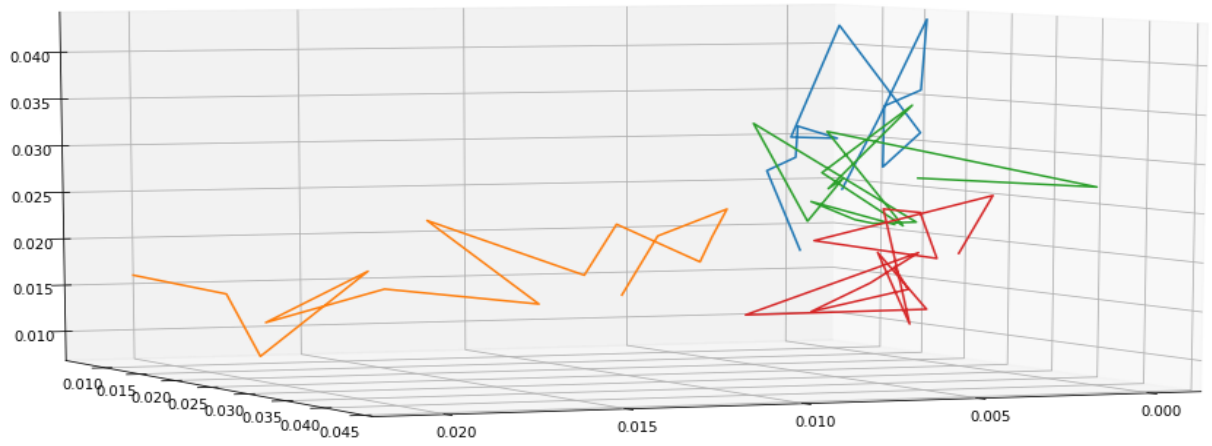

Figure 12: 'MTL', 'BOS', 'PIT' and 'WSH'. Third projection in (x_0, x_1, x_3).

This is particularly clear in figure 12.

Comparing to the rather 'stationary' situation of the other three teams, and seeing how the magnitudes of their variations are team-independent, we are led to the following interpretation:

- Each style corresponds to roughly a ball in these figures whose dimension is universal.
- Some teams (in this case 'BOS') show an evolution of style over time.
- Finally, the evolution of 'BOS' seem to be well-directed – happens along, essentially, a line.

As promised in the first section, let us look at the same features for away games to see what the situation could be, and to convince ourselves that team style is determined by home games, and not away games. Figures 13,14 show a couple of projections in the (x_0, x_1, x_2) space for the same four teams.

Of course, adding teams makes the available space more crowded. Here is how it looks (Figure 15,16) with Toronto Maple Leaves added to the previous four.

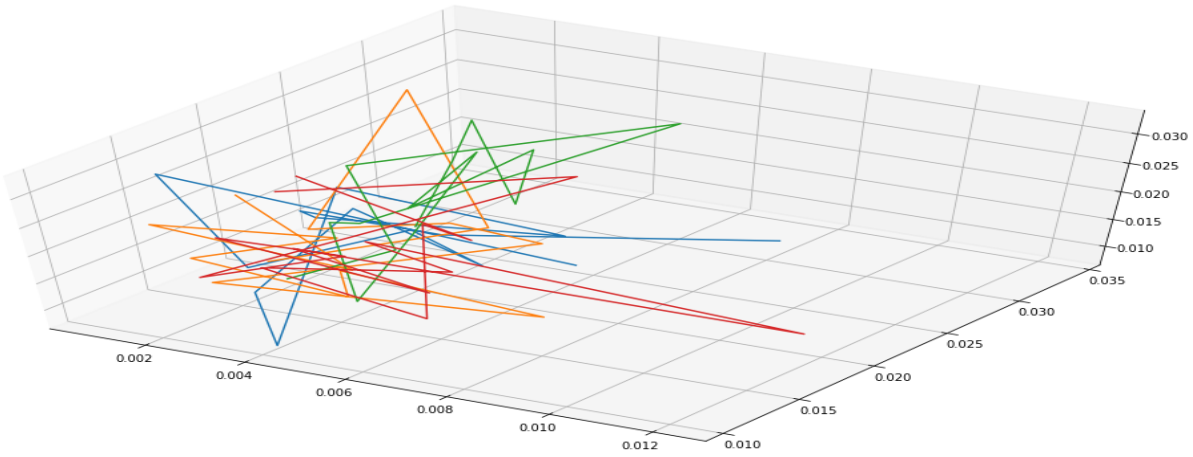

Figure 13: 'MTL', 'BOS', 'PIT' and 'WSH', away games. Projection in (x_0, x_1, x_3) .

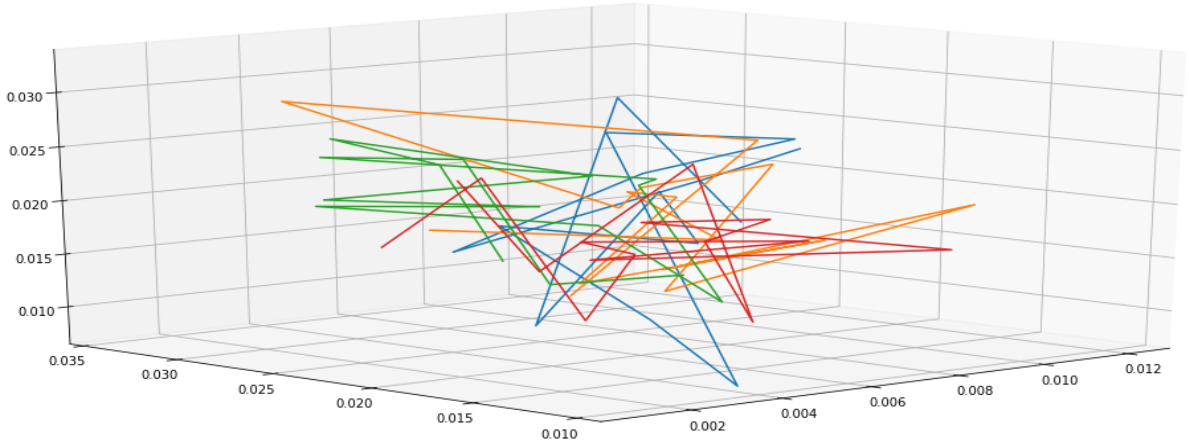

Figure 14: 'MTL', 'BOS', 'PIT' and 'WSH', away games. Another projection in (x_0, x_1, x_3).

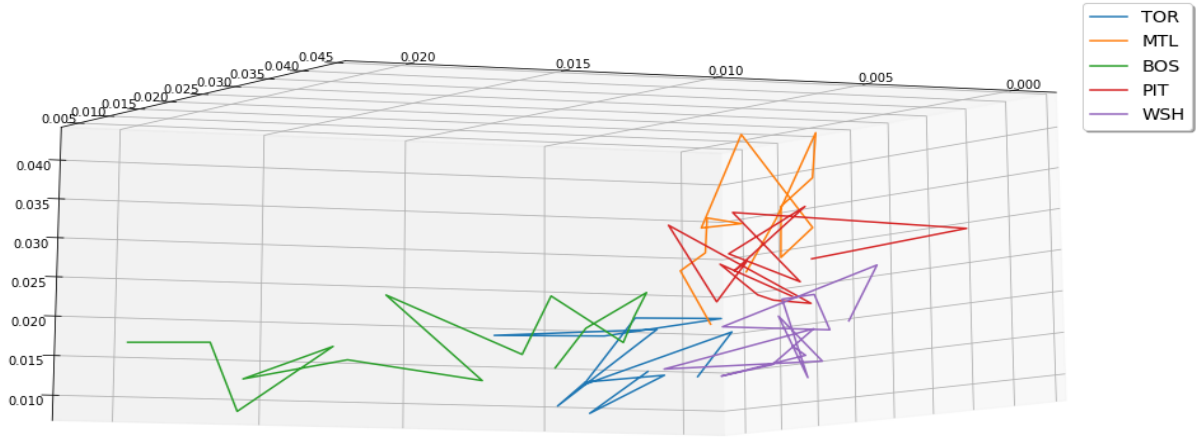

Figure 15: 'TOR','MTL', 'BOS', 'PIT','WSH', home games. Projection in (x_0, x_1, x_3) .

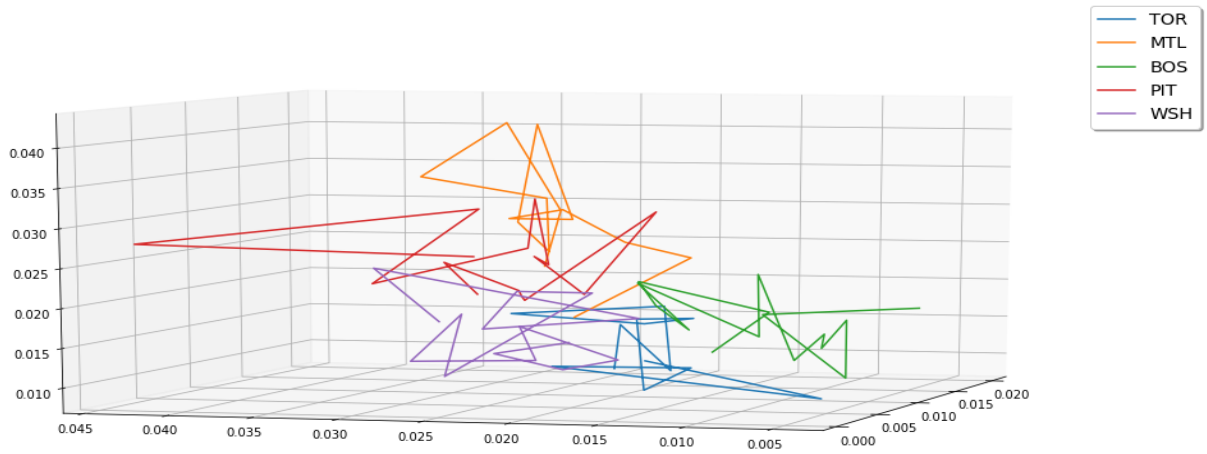

Figure 16: 'TOR','MTL', 'BOS', 'PIT' and 'WSH', home games. Another projection in (x_0, x_1, x_3).

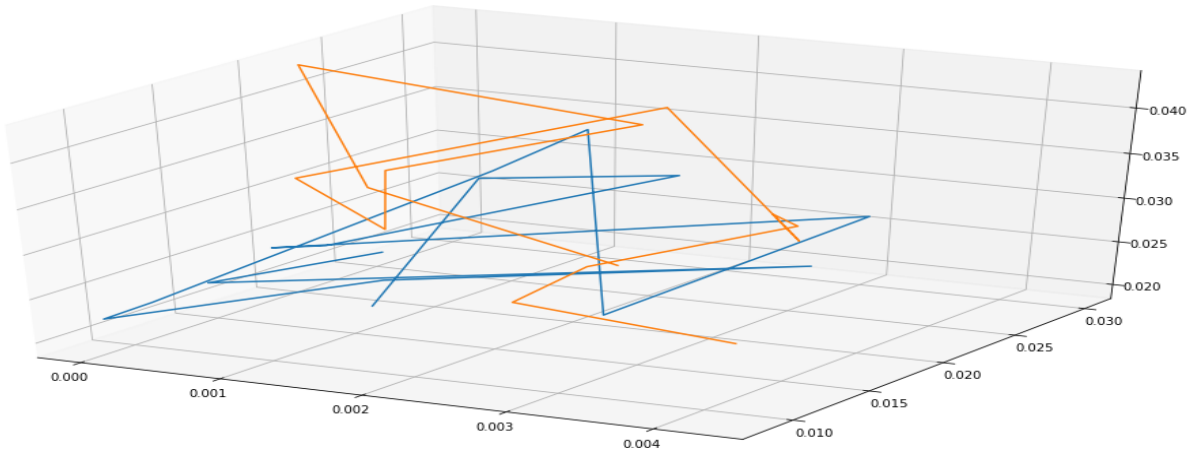

Figure 17: 'L.A' and 'MTL'. Absence of separation in (x_0, x_1, x_2).

6 Separability of 'MTL' and 'L.A'

Not all teams show a clear separation of feature vectors for $N = 20$ in the space (x_0, x_1, x_2). As an example, we show a couple of pictures, Figures 17,18, for the *Montreal Canadians* and *Los Angeles Kings* which demonstrate this.

The conclusion is that the statistics accumulated for sets of $N_{set} = 20$ games does not allow to tell apart the two teams when one looks at words $(2 * 1)$, (11) and (101) only. One can try to suppress the statistical fluctuations by increasing the number of games from which probabilities are calculated (at the expense of the number of feature vectors). This does help – for $N_{set} = 60$ there is a clear separation. But for these two teams there is another way to see the difference – to consider a different set of words, name words (10001), (100001) and (1000001). The corresponding probabilities we denote as (x_4, x_5, x_6) and show the result in Figure 19. The separation is also obvious in the space (x_4, x_5, x_7) (Figure 20).

7 Sets of 60 home games.

Finally, we show results when the probabilities are computed for sets of $N_{set} = 60$ different games – this corresponds to 4-5 feature vectors for each team. Again, and this becomes even clearer, the difference in

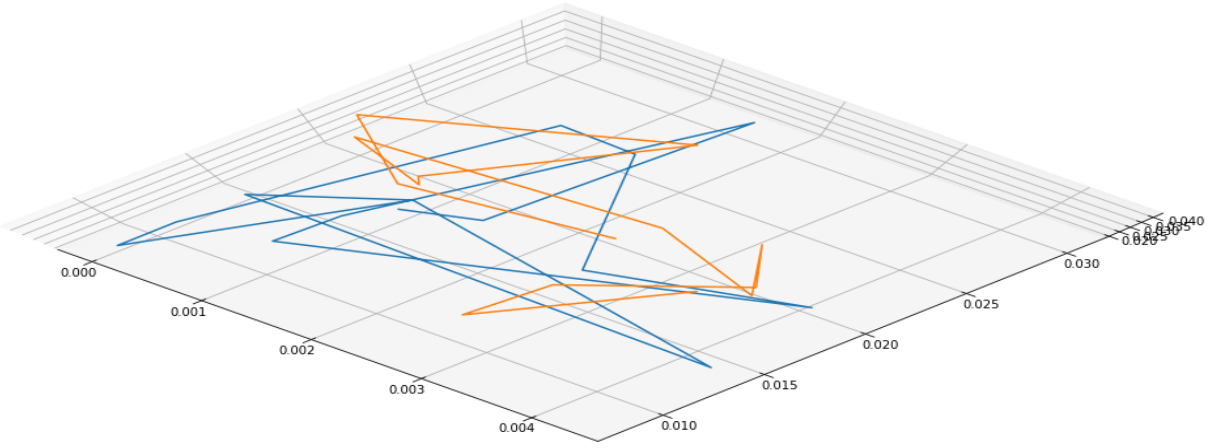

Figure 18: 'L.A' and 'MTL'. Another projection in (x_0, x_1, x_2) .

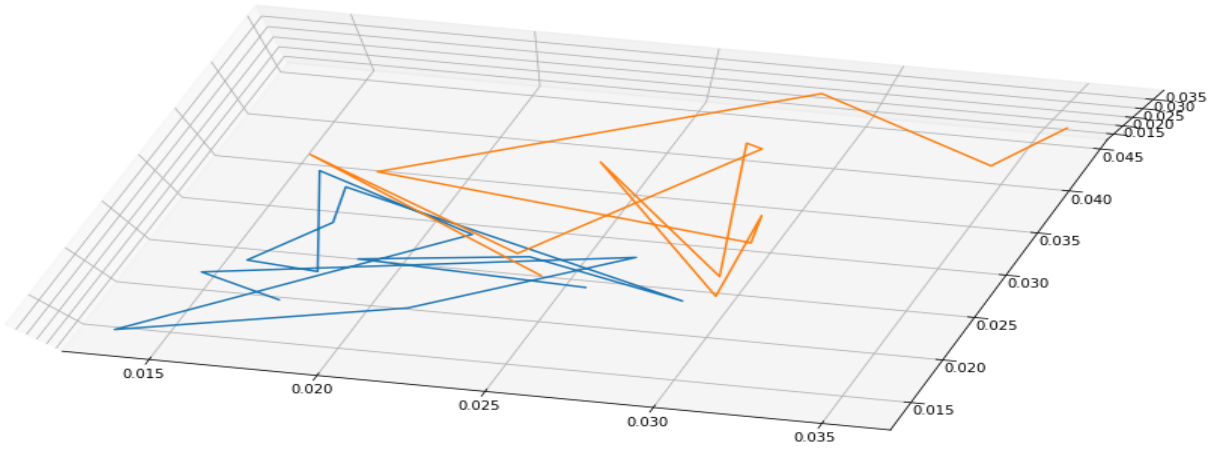

Figure 19: 'L.A' and 'MTL'. Separation in (x_4, x_5, x_6).

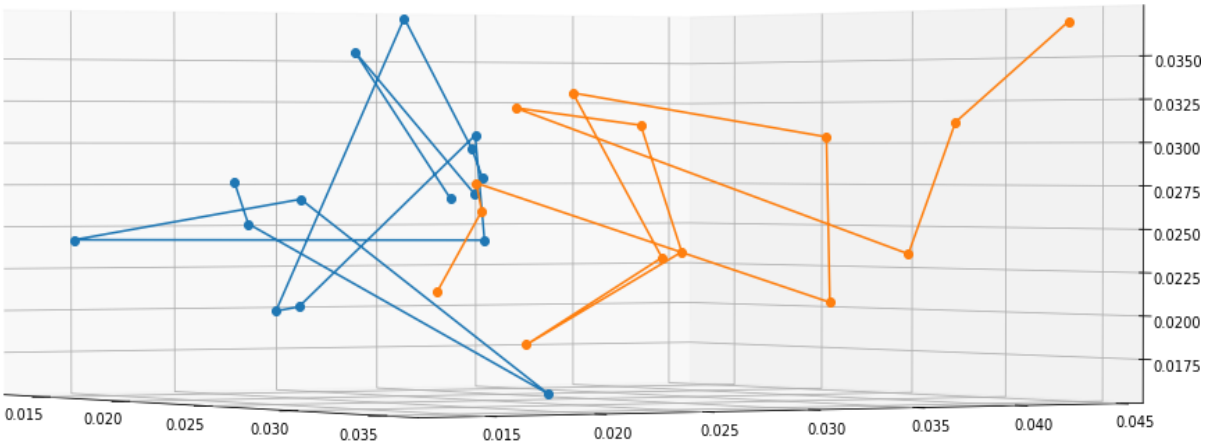

Figure 20: 'L.A' and 'MTL'. Projection in (x_4, x_5, x_7) .

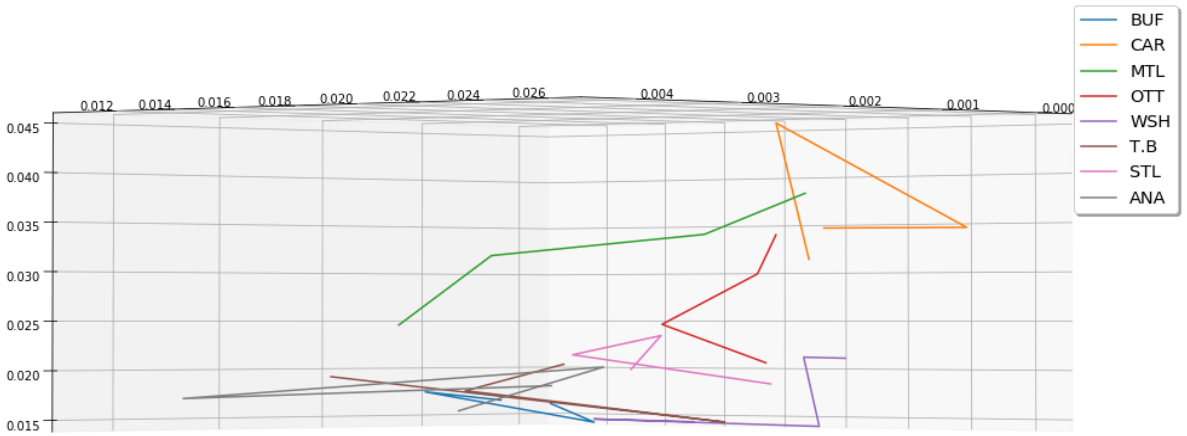

Figure 21: Eight teams inseparable for $N_{set} = 20$. Here for $N_{set} = 60$.

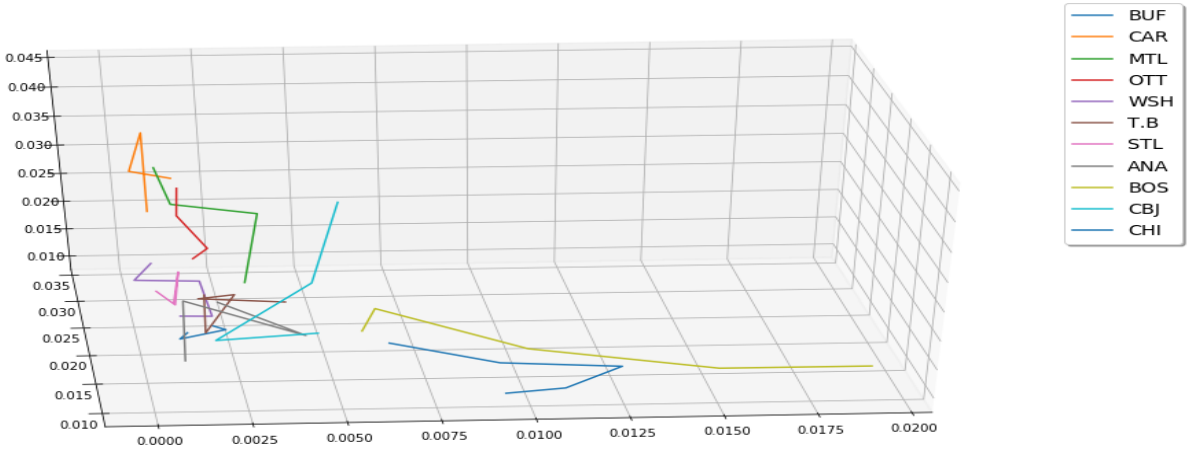

Figure 22: Some of the worst for $N_{set} = 20$. Here for $N_{set} = 60$. Each team has 4-5 independent points. The hierarchy in variances for teams vs total is obvious!

style is obvious – there is a significant hierarchy between variance of feature vectors within a team and the variance in all feature vectors. The former occupy a tiny subregion in the total allowed region.

8 Principal Component Analysis for $N_{set} = 60$

Having noticed an unusual flatness of the Boston Bruins' 5-point plot in Figure 24 we performed a Principal Component Analysis on each of the teams to see if, similarly, their feature points tend to lie on a two-dimensional plane. We used a StandardScaler from Python's scipy module to shift and scale the data. The fit of the scaler was done on Montreal Canadians (so that the mean is zero vector and all three feature coordinates had unit variance on the 'MTL' dataset) which has one of the most 'symmetrical' feature coordinates (as can be seen from the figures). We did not performed rescalings on other teams to learn the relative variances. To find the principal components and their variances we used scipy's standard decomposition.PCA on the rescaled data. The results are summarized in Table 1.

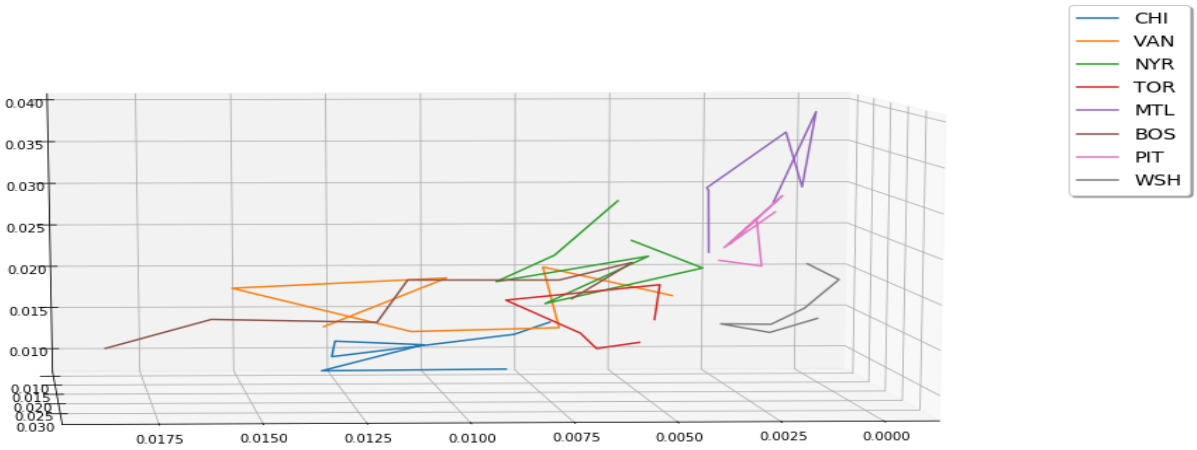

Figure 23: Here for $N_{set} = 40$.

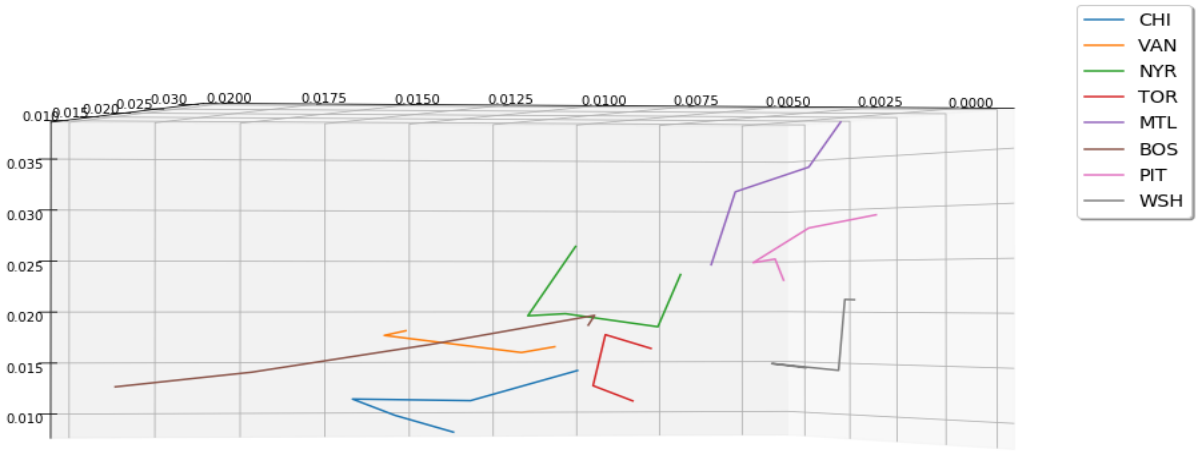

Figure 24: $N_{set} = 60$. Each team has 4-5 independent points. The hierarchy is obvious! Note that in this projection the 5 points of 'BOS' lie essentially on a line!

Team	number of pts	1st PC variance	2nd PC variance	3rd PC variance
ANA	4	1.9	1.7	0.05
ARI	2	1.1	0	0
PHX	2	2.8	0	0
BOS	5	28.8	0.4	0.001
BUF	4	0.4	0.05	0.01
CAR	4	2	0.5	0.2
CBJ	4	15.4	1.2	0.02
CGY	4	0.9	0.5	0.06
CHI	5	4.5	0.4	0.04
COL	4	12.4	3.5	0.05
DAL	4	2.4	0.4	0.001
DET	4	5.4	0.4	0.08
EDM	4	0.8	0.6	0.05
FLA	4	1.2	0.6	0.2
L.A	5	3.3	0.5	0.02
MIN	4	3.6	0.8	0.09
MTL	4	3.3	0.7	0.002
NSH	5	4.4	0.2	0.1
N.J	4	3.4	0.2	0.0005
NYI	4	10.3	3.7	0.4
NYR	5	2.4	1.3	0.03
OTT	4	1.4	0.2	0.02
PHI	4	1.1	0.5	0.04
PIT	5	3.4	0.3	0.04
S.J	5	10.6	0.7	0.12
STL	4	0.9	0.2	0.006
T.B	5	2.2	0.8	0.1
TOR	4	1.6	0.4	0.01
VAN	4	7.7	1.2	0.002
WPG	4	2.4	0.6	0.009
WSH	5	1.5	0.9	0.06

Table 1: Principal Components' Variances.

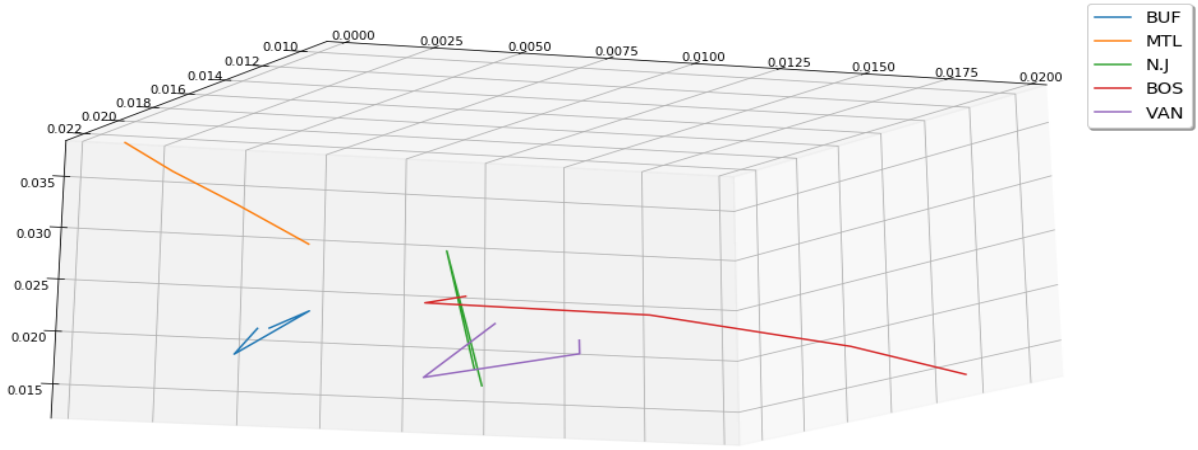

Figure 25: MTL plane.

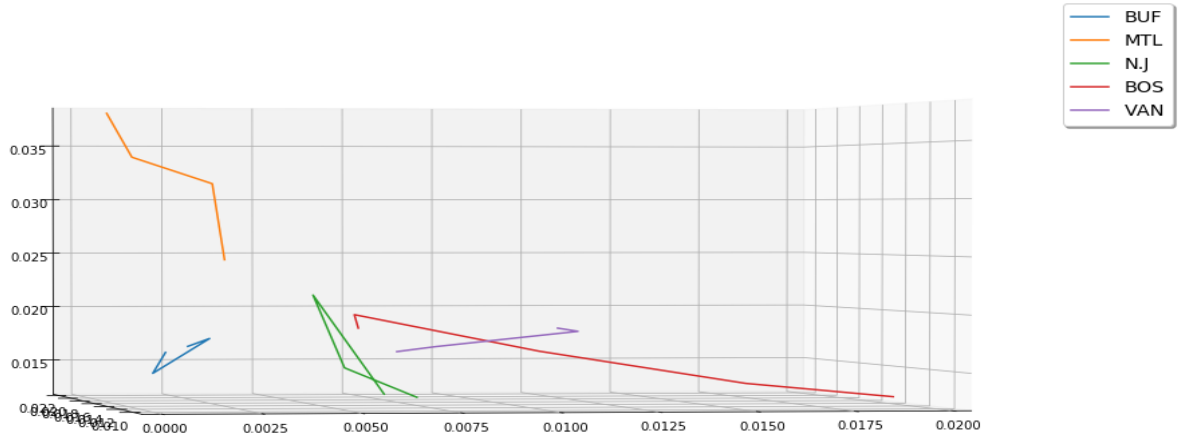

Figure 26: VAN plane.

Obviously, Boston Bruins is not an exception (although it is the 'flattest' team) – with few exceptions, the variance of the least-variant Principal Component is one-two orders of magnitude less than the next-to-least-variant. This cannot be a coincidence, and should be a reflection of the game's structure. The planes are, typically, not parallel to each other, but there seems to be a decomposition into classes – many of them have parallel orthogonal planes, as seen in Figure 28 for 'N.J', 'BUF', 'BOS' and 'MTL'.

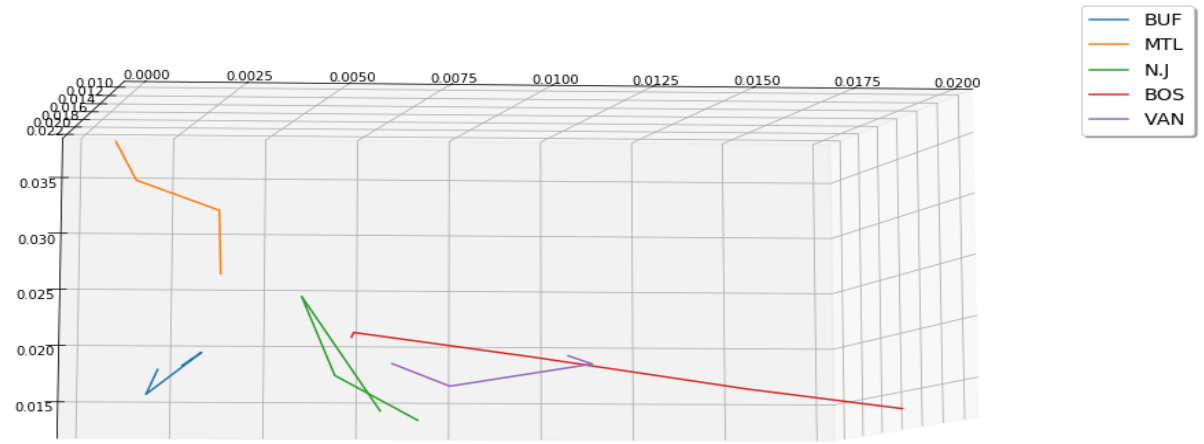

Figure 27: BOS plane.

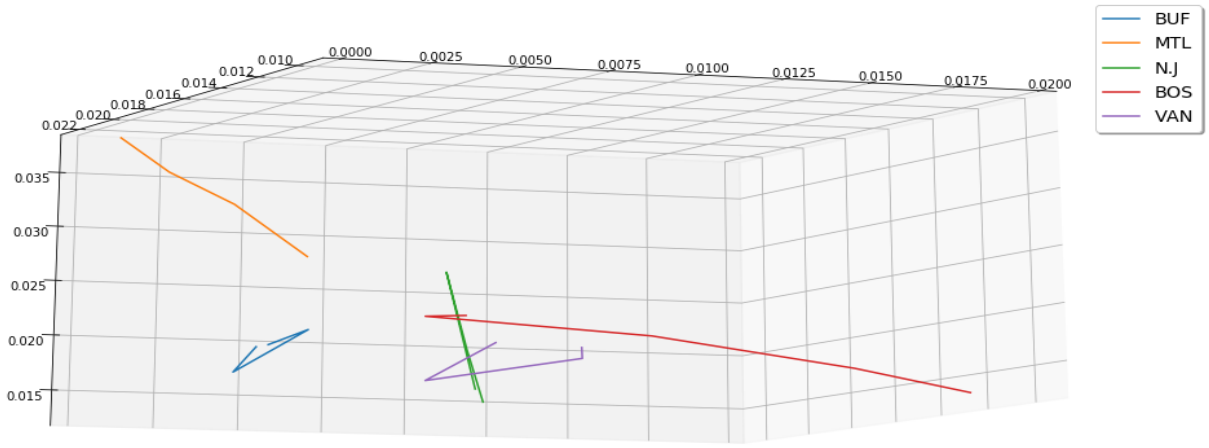

Figure 28: N.J plane. Note that in this projection the other three except VAN reduce to almost line intervals as well!

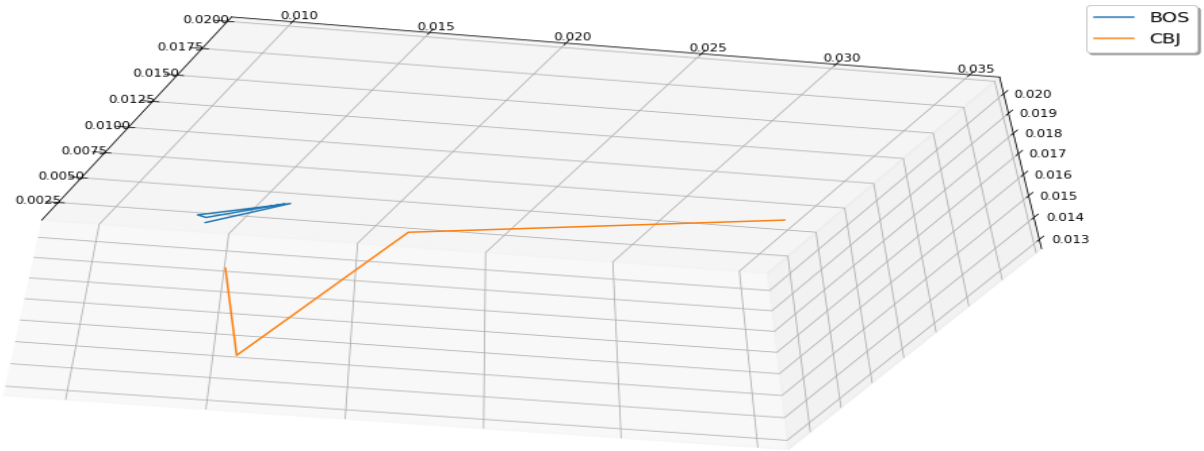

Figure 29: BOS and CBJ. Note hierarchy in variances.

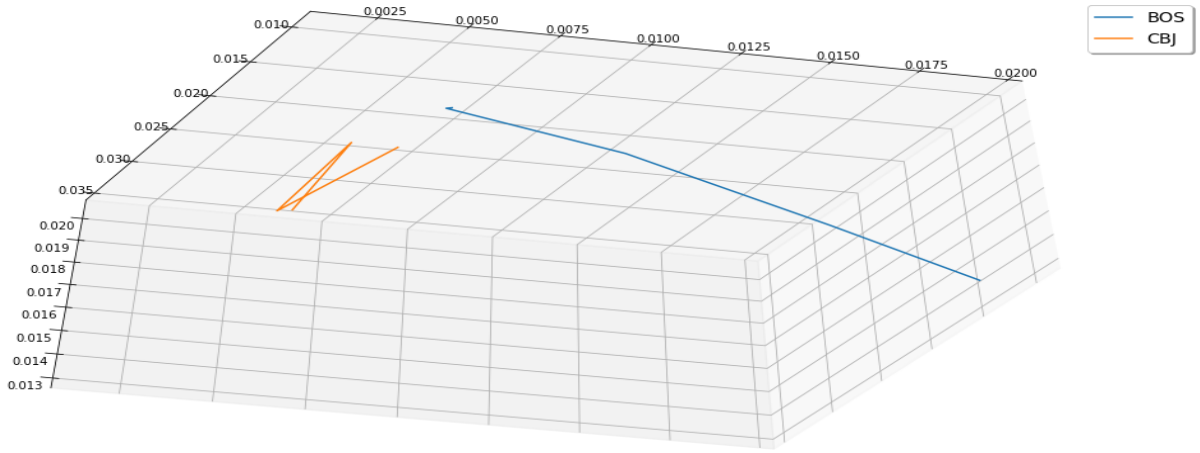

Figure 30: CBJ and BOS. Note hierarchy in variances.