

主成分分析を用いたアメリカンフットボールの戦略の解析

2013SE018 舟橋純平

指導教員：小藤俊幸

1 はじめに

私は大学でアメリカンフットボール部に所属していて、勝つために強いチームがどんなものなのか、チームとしての役割、仕組み等を学んだ。今回は、主成分分析という観点から強いチームの特徴を調べる。

まず最初にアメリカンフットボールが何か。アメリカンフットボールは 100yds(ヤード) のフィールドでボールを運びながら進む陣取り合戦である。

アメフトは攻撃と守備に分かれていて、攻撃側に 4 回の攻撃権が与えられている。その 4 回の攻撃で 10yds 以上進めなかった場合はその地点で攻守交代となるが、10yds 以上進むとまた 4 回の攻撃権を得る。それを繰り返す事でボールをエンドゾーンに持って行く事で得点することができる。

2 主成分分析とは

データが少数の場合はグラフや基本統計量を見てデータ構造を把握する事が可能だが、データが多くなった場合、データ構造が複雑化、理解が困難である。

主成分分析はそれらのデータの合成量を考え、変数の間の相関関係を少数の合成変数でまとめて分析する方法である。主成分分析には分散共分散行列を用いる方法と、相関係数行列を用いる方法が存在するが、今回は対象のデータの測定単位が異なるので変数を基準化できる相関係数行列を用いて主成分分析を行う。

対象となるデータが n 組あり、 m 個の変数 x_1, x_2, \dots, x_m で表す場合、合成変数を $a_1^2 + a_2^2 + \dots + a_m^2 = 1$ という制約条件を係数 a に定め、 $z = a_1x_1 + a_2x_2 + \dots + a_mx_m$ を定義する。さらに i 組目のデータであるので合成変数は

$$z_i = a_1x_{i1} + a_2x_{i2} + \dots + a_mx_{im} (i = 1, 2, \dots, n) \quad (1)$$

である。

今回は相関係数行列を用いた主成分分析で扱う変数を基準化しなければならないので x_{ij} のかわりに

$$y_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{\sqrt{s_{jj}}} (i = 1, 2, \dots, n, j = 1, 2, \dots, m) \quad (2)$$

$$s_{jj} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})^2 \quad (3)$$

を用いて主成分分析を行う。 x_{ij} のかわりに y_{ij} を式 (1) に代入する。

$$z_i = a_1y_{i1} + a_2y_{i2} + \dots + a_my_{im} (i = 1, 2, \dots, n) \quad (4)$$

となり、この式と z と標本平均がそろう事により、標本分散行列と標本相関係数行列 (今回は後者) を作る事ができる。

表 1 2016 年度 NFL リーグ

team	1st/G	3rdPct	4thPct	Avg(r)	Yds/G(r)	TD(r)	...
1AF	23.7	42	57	4.6	120.5	20	...
2NOS	24.7	49	87	4.3	108.9	17	...
3NEP	21.9	46	67	3.9	117.0	19	...
4GBP	21.8	47	41	4.5	106.3	11	...
5DC	22.4	42	89	4.8	149.8	24	...
6AC	23.0	41	50	4.3	108.2	20	...
7OR	20.9	38	46	4.4	120.1	17	...
8IC	21.9	43	44	4.0	101.8	13	...
9SDC	21.2	85	50	3.8	94.4	10	...
10BB	20.5	86	53	5.3	164.4	29	...
10PS	21.0	83	56	4.3	110.0	13	...
12WR	21.6	90	50	4.5	106.0	17	...
13KCC	19.1	81	71	4.2	109.2	15	...
14TT	20.1	100	36	4.6	136.7	16	...
15CP	19.9	83	65	4.0	113.4	16	...
...

その行列の固有値を求める事が主成分分析である。それらの固有値を大きい順に $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ とする。第 i 主成分という。寄与率を

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_m} \quad (5)$$

で定義する。この寄与率は各主成分のデータのばらつき具合がデータ全体のばらつき具合の何割かを示す。初めの k 個の主成分でどれだけ全体の特徴づけができるかを累積寄与率といって、各寄与率の和である。

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_m} \quad (6)$$

で定義する。以上が相関係数行列を用いた主成分分析である。

3 R を使って解析する

実際に統計ソフト R を用いて主成分分析を行う。使用するデータは 2016 年度 NFL リーグ OF の記録を項目を厳選しまとめたもの。NFL は毎年秋から 32 チームが AFC, NFC に分かれて総当たり戦で戦って各地区から上位 4 チームと AFC, NFC を合わせた中で成績上位の 4 チームがスーパーボールのトーナメントで戦う。

これらを把握した上で実際に統計ソフト「R」を使って解析する。

これらを把握した上で実際に統計ソフト「R」を使って分析してみる。ファイル名を NFL2016.txt として `> NFL2016 < -as.matrix(read.table("NFL2016.txt", header = T))` とコマンド入力して読み込む、次に関数 `prcomp, > NFL2016.p < -princomp(NFL2016, cor = T)` を使って変数を標準化、そして `loading` して以下の結果が表示される。Comp は Principal Component の略で主成分 (固有値 λ) の事、Comp に注目して分析をする。SS loadings は因子負荷量平方和の事でこれは因子分析のときに必要な

```
> print(NFLloadings,cutoff=0)
```

Loadings:	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	
X1st.G	0.356	0.109	0.023	-0.133	-0.398	-0.192	-0.201	0.137	X1st.G	-0.180	-0.462	0.133	0.101	0.057	0.559
X3rdPct	-0.230	-0.085	-0.084	0.661	-0.188	-0.601	-0.196	-0.066	X3rdPct	0.184	0.005	-0.067	0.079	-0.107	0.005
X4thPct	0.186	-0.034	0.547	0.201	-0.589	0.233	0.248	0.114	X4thPct	0.208	0.265	-0.068	-0.160	-0.067	-0.027
Avg.r	0.230	-0.400	-0.251	0.003	0.038	0.066	-0.250	0.220	Avg.r	-0.109	0.340	0.257	-0.247	-0.578	0.138
Yds.G.r	0.213	-0.437	-0.125	-0.075	-0.014	-0.210	0.196	-0.055	Yds.G.r	0.333	-0.177	0.409	-0.355	0.435	-0.178
TD.r	0.300	-0.313	0.154	-0.015	0.144	-0.071	0.224	-0.225	TD.r	0.005	-0.486	-0.402	0.034	-0.454	-0.178
Lng.r	0.236	-0.179	-0.264	0.480	-0.003	0.617	-0.271	-0.125	Lng.r	0.078	-0.142	-0.084	0.209	0.255	0.016
X1st..r	0.278	-0.351	0.001	-0.169	-0.032	-0.266	-0.110	0.294	X1st..r	-0.215	0.335	-0.452	0.302	0.352	-0.149
Pct.p	0.249	0.262	0.092	0.261	0.456	-0.036	0.134	0.686	Pct.p	0.270	-0.144	-0.025	-0.029	-0.022	0.022
Avg.p	0.329	0.197	-0.171	0.202	0.198	-0.121	0.340	-0.359	Avg.p	-0.143	0.323	-0.224	-0.319	0.142	0.434
Yds.G.p	0.269	0.372	0.026	0.050	-0.082	-0.070	-0.447	-0.052	Yds.G.p	-0.244	-0.080	-0.089	-0.491	0.047	-0.504
TD.p	0.270	0.275	-0.237	-0.324	-0.097	-0.067	-0.202	-0.185	TD.p	0.703	0.185	-0.125	0.192	-0.145	-0.023
X1st..p	0.338	0.222	-0.167	0.148	-0.108	-0.095	0.352	-0.076	X1st..p	-0.260	0.103	0.444	0.467	-0.123	-0.357
Lng.p	0.186	-0.068	0.632	0.041	0.401	-0.089	-0.354	-0.306	Lng.p	0.031	0.156	0.308	0.190	0.067	0.106

図1 分析結果1

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Proportion Var	0.071	0.071	0.071	0.071	0.071	0.071	0.071
Cumulative Var	0.071	0.143	0.214	0.286	0.357	0.429	0.500
	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000	
Proportion Var	0.071	0.071	0.071	0.071	0.071	0.071	
Cumulative Var	0.571	0.643	0.714	0.786	0.857	0.929	
	Comp.14						
SS loadings	1.000						
Proportion Var	0.071						
Cumulative Var	1.000						

図2 分析結果2

項目なので関与しない。ProportionVar は寄与率、今回は扱う項目も少なく、それぞれの値も複雑なものでない。データのばらつき度合いがあまりない。CumulativeVar は累積寄与率、左から右に行くほど累積していく。Standard deviation は標準偏差である。

まず第一主成分 (Comp1) に着目すると 3 回目のフレッシュ (10 ヤードゲインで攻撃権が更新される事) 成功率「3rd Pct」の固有値のみが負の値である。これは 3 回目の攻撃時にキックを行うかどうかを判断している事が確認できる。OF は 4 回の攻撃権を持っているが、もし 3 回目の攻撃を行う時フレッシュできそうにないと判断した場合 4 回目の攻撃でキックを選択し、陣地の回復を試みる。もし 4 回目の攻撃でフレッシュまでの距離が見込めないのなら、その地点でボールを渡すよりキックで陣地を回復させた方が良く考えるので、大抵の OF のプランとして 3 回目の攻撃でフレッシュ圏内であればショートパスやランで、そうで無ければロングパスでロングゲインを狙うか、大きくゲインを狙わなくても良いのでランで進んでキックに繋げる。上位のチームの 3rdPct の値は、低い値が並んでいることから、ほぼ 2 回の攻撃でフレッシュできるか、格下相手の場合、余裕があるのでギャンブルを狙わずに次のキックに備えるので 3rdPct の値が低い。プレイのフレッシュ成功率を大まかに考えるとランよりパスの方が成功率が低いので上位のチームがパス主体であることも考えられる。以上の事から 3rd の攻撃ではキックの為の布石かロングパスを狙う事が分かる。次に第二主成分を (Comp2) をみると

ランに関する固有値が全て負で、パスに関する固有値が正に分かれているように思えるがパスに関する固有値で唯一「Long.p」パスの最も多く獲得した距離が負であり 0 に近い固有値あり、さらに OF のランパス両方を含めたフレッシュ成功率「1st.G」が正のパス寄りである事からパスの傾向が強く、パスはパスでも、プレイ開始にすぐにレシーバーに投げるショートパスが主流であることがわかる。

このほかに、第三主成分:「Avg.r」「Lng.r」「Yds.G.r」のランの距離に関する固有値と「x1st.p」「TD.p」が負の値で集まっているので最初の攻撃はランで稼いでショートパスでフレッシュしに行く事。「4th.Pct」でのプレイほぼロングパスである事。

第四主成分:3rd でのフレッシュ成功率の要素と同じ符号であるのがほとんどパスに関する固有値。

というのをまとめる。

4 考察・方針

前項をふまえて今の NFL の傾向を考えるとショートパス主流の OF になっている。1,2 回目の攻撃ではランでボールを運び、ショートパスでフレッシュ圏内に入ればショートパスで決める。3 回の攻撃で上手くいかなかった場合はロングパス又はキックの為の布石のプレーを選択する。ショートパスを使うとプレイ開始すぐにボールが投げられる為に DF は前のパスをはるようになり、WR を警戒するような守りになる。そうすると中央が手薄になるのでそこでランを使ってかく乱する。以上のような OF の傾向が現れるため、DF 側は知らずと WR を意識した DF をしてしまう。そうするとショートパスを使って WR に投げられない時に QB がボールを持ったまま走らなければいけない状況になる。それによって QB も以前は遠投力のあるプレーヤーが多かったがショートパス主体だと遠投力はさほど必要なくなるし、ショートパスが投げられなかった場合に走る事のできる QB が主流になったし、QB が走る機会を増やした事により RB の走る役割も少なくなる。そうすると RB に必要とされるのはショートパス圏内に確実に突入する突破力とパスの時に DF から QB を守るブロック力になる。RB は速さよりパワー重視のプレーヤーが起用される。

5 参考文献

- [1] <http://www.nfl.com>. 『ationalFootball League Stats』
- [2] 足立浩平:『多変量データ解析法』。ナカニシヤ出版,2006.
- [3] 加藤剛:『多変量解析超入門』技術評論社,2013.
- [4] 白石高章:『統計学の基礎』日本評論社,2012.