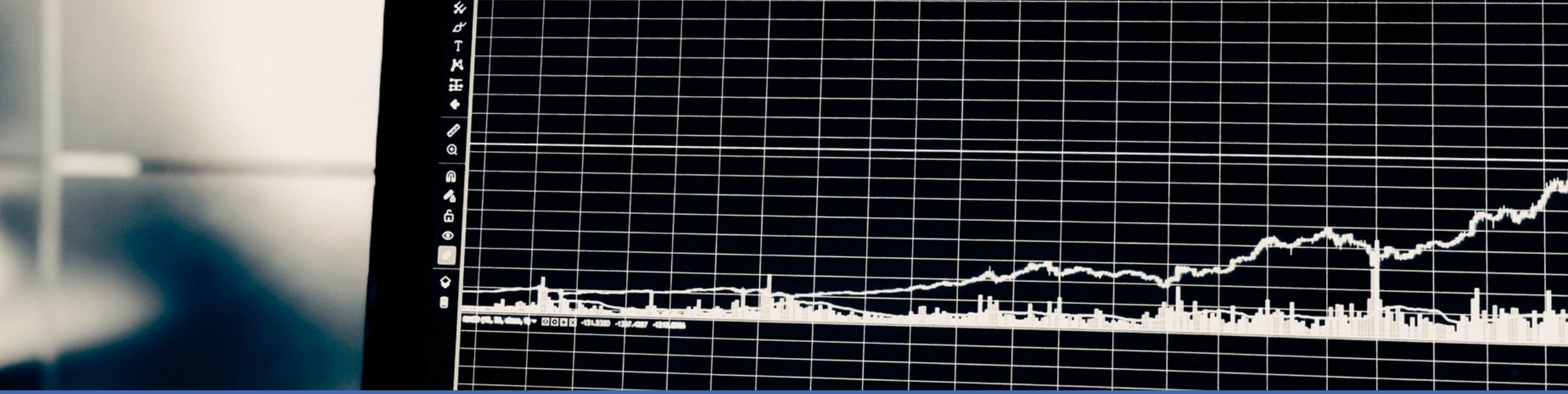


PRINCIPAL COMPONENT ANALYSIS

Francesc Contreras & Albert Pérez

PRINCIPAL COMPONENT ANALYSIS

“Simplify the complexity of the data reducing the number of dimensions and maintain the information.”

1 Large data-set

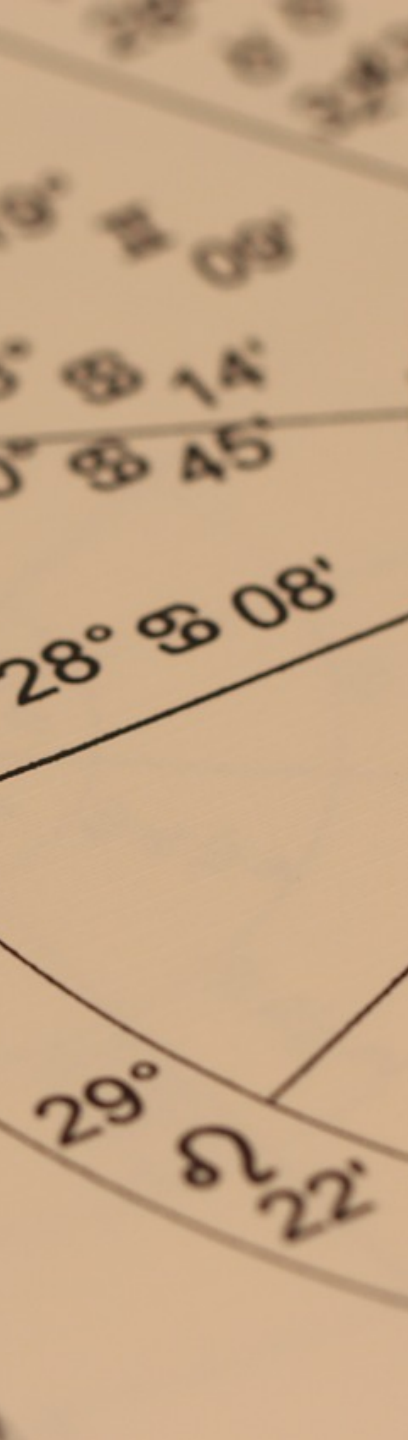
In order to **interpret such data-sets**, methods are required to **drastically reduce** their dimensionality in an interpretable way.

2 Technique

Reduce the **dimensionality** of a **data-set**, while preserving as much ‘variability’ as possible.

3 Benefits

Increasing interpretability but at the same time **minimizing information loss**.



ANALYSIS STEPS

Get Data

Obtain **X matrix.** It is a set of all data vectors, one vector per row.

Check correlation

Checking if the dataset has **correlation**.

Centre & Reduce Data

eigen(R)

Calculate Components

DATA-SET

BODYFAT.CSV

```
> names(dataframe) # column names
[1] "BODYFAT"   "DENSITY"   "AGE"       "WEIGHT"    "HEIGHT"    "ADIPOSITY" "NECK"      "CHEST"     "ABDOMEN"
"HIP"       "THIGH"     "KNEE"      "ANKLE"     "BICEPS"    "FOREARM"   "WRIST"
```

```
> dim(dataframe) # dimension of dataframe
[1] 252  16
```

```
> sapply(dataframe, typeof)
BODYFAT   DENSITY       AGE    WEIGHT    HEIGHT ADIPOSITY      NECK     CHEST   ABDOMEN       HIP     THIGH
KNEE     ANKLE    BICEPS   FOREARM     WRIST
"double"  "double" "integer"  "double"  "double"  "double"  "double"  "double"  "double"  "double"  "double"
"double"  "double"  "double"  "double"  "double"
```

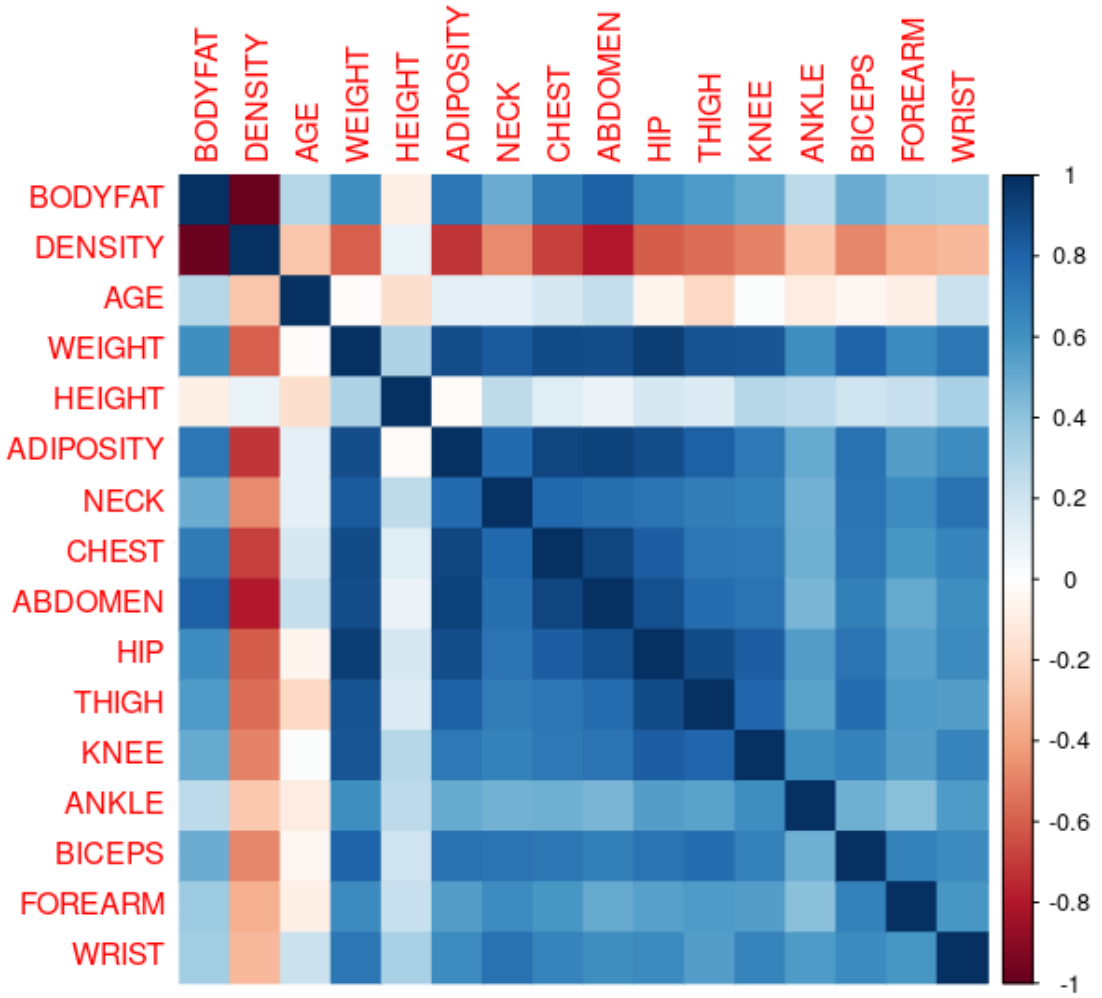
```
> summary(dataframe)
```

```
    BODYFAT         DENSITY           AGE            WEIGHT          HEIGHT         ADIPOSITY          NECK            CHEST           ABDOMEN            HIP            THIGH            KNEE           ANKLE           BICEPS          FOREARM     
 Min.   : 0.00   Min.   :0.995   Min.   :22.00   Min.   :118.5   Min.   :29.50   Min.   :18.10   Min.   :31.10   Min.   : 79.30   Min.   : 69.40   Min.   : 85.0   Min.   :47.20   Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00  
 1st Qu.:12.80   1st Qu.:1.041   1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25   1st Qu.:23.10   1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30  
 Median :19.00   Median :1.055   Median :43.00   Median :176.5   Median :70.00   Median :25.05   Median :38.00   Median : 99.65   Median : 90.95   Median : 99.3   Median :59.00   Median :38.50   Median :22.8   Median :32.05   Median :28.70  
 Mean   :18.94   Mean   :1.056   Mean   :44.88   Mean   :178.9   Mean   :70.15   Mean   :25.44   Mean   :37.99   Mean   :100.82   Mean   : 92.56   Mean   : 99.9   Mean   :59.41   Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66  
 3rd Qu.:24.60   3rd Qu.:1.070   3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:27.32   3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00  
 Max.   :45.10   Max.   :1.109   Max.   :81.00   Max.   :363.1   Max.   :77.75   Max.   :48.90   Max.   :51.20   Max.   :136.20   Max.   :148.10   Max.   :147.7   Max.   :87.30   Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90  
     WRIST      
 Min.   :15.80  
 1st Qu.:17.60  
 Median :18.30  
 Mean   :18.23  
 3rd Qu.:18.80  
 Max.   :21.40  
```

```
> View(dataframe)
```

	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
1	12.6	1.0708	23	154.25	67.75	23.7	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
2	6.9	1.0853	22	173.25	72.25	23.4	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
3	24.6	1.0414	22	154.00	66.25	24.7	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
4	10.9	1.0751	26	184.75	72.25	24.9	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
5	27.8	1.0340	24	184.25	71.25	25.6	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7
6	20.6	1.0502	24	210.25	74.75	26.5	39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8
7	19.0	1.0549	26	181.00	69.75	26.2	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
8	12.8	1.0704	25	176.00	72.50	23.6	37.8	99.6	88.5	97.1	60.0	39.4	23.2	30.5	29.0	18.8
9	5.1	1.0900	25	191.00	74.00	24.6	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2
10	12.0	1.0722	23	198.25	73.50	25.8	42.1	99.6	88.6	104.1	63.1	41.7	25.0	35.6	30.0	19.2

```
> R <- cor(centrate_data, method = "pearson")
> corrplot(R, method="color")
```



X MATRIX

	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
1	12.6	1.0708	23	154.25	67.75	23.7	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
2	6.9	1.0853	22	173.25	72.25	23.4	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
3	24.6	1.0414	22	154.00	66.25	24.7	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
4	10.9	1.0751	26	184.75	72.25	24.9	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
5	27.8	1.0340	24	184.25	71.25	25.6	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7
6	20.6	1.0502	24	210.25	74.75	26.5	39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8
7	19.0	1.0549	26	181.00	69.75	26.2	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
8	12.8	1.0704	25	176.00	72.50	23.6	37.8	99.6	88.5	97.1	60.0	39.4	23.2	30.5	29.0	18.8
9	5.1	1.0900	25	191.00	74.00	24.6	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2
10	12.0	1.0722	23	198.25	73.50	25.8	42.1	99.6	88.6	104.1	63.1	41.7	25.0	35.6	30.0	19.2
11	7.5	1.0830	26	186.25	74.50	23.6	38.5	101.5	83.6	98.2	59.7	39.7	25.2	32.8	29.4	18.5
12	8.5	1.0812	27	216.00	76.00	26.3	39.4	103.6	90.9	107.7	66.2	39.2	25.9	37.2	30.2	19.0
13	20.5	1.0513	32	180.50	69.50	26.3	38.4	102.0	91.6	103.9	63.4	38.3	21.5	32.5	28.6	17.7
14	20.8	1.0505	30	205.25	71.25	28.5	39.4	104.1	101.8	108.6	66.0	41.5	23.7	36.9	31.6	18.8
15	21.7	1.0484	35	187.75	69.50	27.4	40.5	101.3	96.4	100.1	69.0	39.0	23.1	36.1	30.5	18.2
16	20.5	1.0512	35	162.75	66.00	26.3	36.4	99.1	92.8	99.2	63.1	38.7	21.7	31.1	26.4	16.9
17	28.1	1.0333	34	195.75	71.00	27.3	38.9	101.9	96.4	105.2	64.8	40.8	23.1	36.2	30.8	17.3
18	22.4	1.0468	32	209.25	71.00	29.2	42.1	107.6	97.5	107.0	66.9	40.0	24.4	38.2	31.6	19.3
19	16.1	1.0622	28	183.75	67.75	28.2	38.0	106.8	89.6	102.4	64.2	38.7	22.9	37.2	30.5	18.5
20	16.5	1.0610	33	211.75	73.50	27.6	40.0	106.2	100.5	109.0	65.8	40.6	24.0	37.1	30.1	18.2
21	19.0	1.0551	28	179.00	68.00	27.3	39.1	103.3	95.9	104.9	63.5	38.0	22.1	32.5	30.3	18.4
22	15.3	1.0640	28	200.50	69.75	29.1	41.3	111.4	98.8	104.8	63.4	40.6	24.6	33.0	32.8	19.9
23	15.7	1.0631	31	140.25	68.25	21.2	33.9	86.0	76.4	94.6	57.4	35.3	22.2	27.9	25.9	16.7
24	17.6	1.0584	32	148.75	70.00	21.4	35.5	86.7	80.0	93.4	54.9	36.2	22.1	29.8	26.7	17.1
25	14.2	1.0668	28	151.25	67.75	23.2	34.5	90.2	76.3	95.8	58.4	35.5	22.9	31.1	28.0	17.6
26	4.6	1.0911	27	159.25	71.50	21.9	35.7	89.6	79.7	96.5	55.0	36.7	22.5	29.9	28.2	17.7
27	8.5	1.0811	34	131.50	67.50	20.3	36.2	88.6	74.6	85.3	51.7	34.7	21.4	28.7	27.0	16.5
28	22.4	1.0468	31	148.00	67.50	22.9	38.8	97.4	88.7	94.7	57.5	36.0	21.0	29.2	26.6	17.0
29	4.7	1.0910	27	133.25	64.75	22.4	36.4	93.5	73.9	88.5	50.1	34.5	21.3	30.5	27.9	17.2
30	9.4	1.0790	29	160.75	69.00	23.8	36.7	97.4	83.5	98.7	58.9	35.3	22.6	30.1	26.7	17.6
31	12.3	1.0716	32	182.00	73.75	23.6	38.7	100.5	88.7	99.8	57.5	38.7	33.9	32.5	27.7	18.4
32	6.5	1.0862	29	160.25	71.25	22.2	37.3	93.5	84.5	100.6	58.5	38.8	21.5	30.1	26.4	17.9
33	13.4	1.0719	27	168.00	71.25	23.3	38.1	93.0	79.1	94.5	57.3	36.2	24.5	29.0	30.0	18.8
34	20.9	1.0502	41	218.50	71.00	30.5	39.8	111.7	100.5	108.3	67.1	44.2	25.2	37.5	31.5	18.7
35	31.1	1.0263	41	247.25	73.50	32.2	42.1	117.0	115.6	116.1	71.2	43.3	26.3	37.3	31.7	19.7
36	38.2	1.0101	49	191.75	65.00	32.0	38.4	118.5	113.1	113.8	61.9	38.3	21.9	32.0	29.8	17.0
37	23.6	1.0438	40	202.25	70.00	29.1	38.5	106.5	100.9	106.2	63.5	39.9	22.6	35.1	30.6	19.0

Showing 1 to 37 of 252 entries, 16 total columns

Y MATRIX

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	-0.819407078	0.80164696	-1.740073293	-0.841245827	-0.65620473	-0.47705824	-0.738664699	-0.918047611	-0.68353321	-0.7559288203	-0.077478850	-0.53613146	-0.710825680	-0.09067593	-0.62671828	-1.21254120
2	-1.556273245	1.56506057	-1.819583442	-0.193462390	0.57478975	-0.55945623	0.209364662	-0.858620948	-0.88796263	-0.1685021952	-0.134735948	-0.53613146	0.175946951	-0.58814329	0.11707924	-0.03194260
3	0.731890114	-0.74623993	-1.819583442	-0.849769293	-1.06653622	-0.20239825	-1.645475391	-0.597143632	-0.43264256	-0.0985704541	0.037035345	0.12859242	0.530656003	-1.15193963	-1.71762131	-1.74917692
4	-1.039174181	1.02803858	-1.501542847	0.198617058	0.57478975	-0.14746625	-0.244040684	0.115976320	-0.57202626	0.1811565102	0.132463841	-0.53613146	-0.178762102	0.04198203	0.36501175	-0.03194260
5	1.145569365	-1.13584412	-1.660563144	0.181570125	0.30123542	0.04479574	-1.480600720	-0.418863644	0.69171924	0.2790609477	0.724120516	1.49958540	0.530656003	-0.02434695	-0.47795878	-0.56857833
6	0.214791050	-0.28292685	-1.660563144	1.068010617	1.25867557	0.29198972	0.415458001	0.436880299	0.17135345	1.1042554925	1.258520093	1.41649492	1.476546809	1.13641023	0.96004976	0.61202027
7	0.007951424	-0.03547554	-1.501542847	0.070765064	-0.10909607	0.20959173	-0.656227363	0.508192294	-0.17245967	0.0552793763	-0.191993045	-0.12067904	-0.119643926	-0.12384042	-0.42837228	-0.56857833
8	-0.793552125	0.78058728	-1.581052996	-0.099704261	0.64317833	-0.50452423	-0.079166013	-0.145500996	-0.37688908	-0.3922837666	0.113378142	0.33631863	0.057710600	-0.58814329	0.16666574	0.61202027
9	-1.788967823	1.81251187	-1.581052996	0.411703715	1.05350982	-0.22986425	0.044489991	0.009008327	-0.93442386	-0.0006660166	0.666863418	-0.12067904	0.412419652	1.20273921	1.20798227	-0.03194260
10	-0.896971938	0.87535586	-1.740073293	0.658884236	0.91673266	0.09972773	1.693236705	-0.145500996	-0.36759684	0.5867606085	0.705034816	1.29185919	1.121837757	1.10324574	0.66253075	1.04132885
11	-1.478708385	1.44396737	-1.501542847	0.249757856	1.19028698	-0.50452423	0.209364662	0.080320323	-0.83220915	-0.2384339363	0.056121044	0.46095435	1.240074107	0.17464000	0.36501175	0.29003883
12	-1.349433619	1.34919879	-1.422032698	1.264050342	1.60061848	0.23705772	0.580332673	0.329912306	-0.15387517	1.0902691443	1.296691491	0.25322814	1.653901335	1.63387759	0.76170376	0.82667456
13	0.201863573	-0.22501271	-1.024481955	0.053718131	-0.17748465	0.23705772	0.168145994	0.139746985	-0.08882945	0.5587879121	0.762291914	-0.12067904	-0.947298382	0.07514652	-0.03168026	-0.56857833
14	0.240646003	-0.26713208	-1.183502252	0.897541292	0.30123542	0.84130969	0.580332673	0.389338969	0.85897967	1.2161462783	1.258520093	1.20876871	0.353301477	1.53438411	1.45591478	0.61202027
15	0.356993292	-0.37769543	-0.785951509	0.300898653	-0.17748465	0.53918371	1.033738019	0.056549658	0.35719837	0.0273066799	1.831091069	0.17013766	-0.001407576	1.26906819	0.91046326	-0.03194260
16	0.201863573	-0.23027763	-0.785951509	-0.551447974	-1.13492480	0.23705772	-0.656227363	-0.204927658	0.02267751	-0.0985704541	0.705034816	0.04550193	-0.829062031	-0.38915635	-1.12258329	-1.42719549
17	1.184351795	-1.17269857	-0.865461658	0.573649574	0.23284684	0.51171771	0.374239334	0.127861653	0.35719837	0.7406104389	1.029491703	0.91795201	-0.001407576	1.30223268	1.05922276	-0.99788691
18	0.447485629	-0.46193418	-1.024481955	1.033916752	0.23284684	1.03357168	1.693236705	0.805325608	0.45941308	0.9923647068	1.430291386	0.58559008	0.767128704	1.96552249	1.45591478	1.14865599
19	-0.366945397	0.34886372	-1.342522550	0.164523193	-0.65620473	0.75891170	0.003271323	0.710242947	-0.27467437	0.3489926888	0.914977508	0.04550193	-0.119643926	1.63387759	0.91046326	0.29003883
20	-0.315235491	0.28568466	-0.944971806	1.119151415	0.91673266	0.59411570	0.827644680	0.638930952	0.73818047	1.2720916711	1.220348695	0.83486153	0.530656003	1.60071310	0.71211726	-0.03194260
21	0.007951424	-0.02494570	-1.342522550	0.002577334	-0.58781614	0.51171771	0.456676669	0.294256308	0.31073714	0.6986513942	0.781377613	-0.24531476	-0.592589329	0.07514652	0.81129026	0.18271169
22	-0.470365210	0.44363231	-1.342522550	0.735595433	-0.10909607	1.00610568	1.363487362	1.256968244	0.58021228	0.6846650460	0.762291914	0.83486153	0.885365055	0.24096898	2.05095279	1.79261886
23	-0.418655304	0.39624801	-1.103992104	-1.318559938	-0.51942756	-1.16370820	-1.686694059	-1.761906221	-1.50125088	-0.7419424721	-0.382850037	-1.36703630	-0.533471154	-1.45042005	-1.37051580	-1.64184978
24	-0.173033248	0.14879671	-1.024481955	-1.028762085	-0.04070749	-1.10877620	-1.027195373	-1.678708893	-1.16673002	-0.9097786507	-0.859992517	-0.99312912	-0.592589329	-0.82029472	-0.97382379	-1.21254120
25	-0.612567453	0.59105011	-1.342522550	-0.943527422	-0.65620473	-0.61438823	-1.439382052	-1.262722254	-1.51054313	-0.5741062935	-0.191993045	-1.28394581	-0.119643926	-0.38915635	-0.32919927	-0.67590547
26	-1.853605206	1.87042601	-1.422032698	-0.670776501	0.36962400	-0.97144621	-0.944758038	-1.334034250	-1.19460676	-0.4762018559	-0.840906818	-0.78540291	-0.356116628	-0.78713023	-0.23002627	-0.56857833
27	-1.349433619	1.34393387	-0.865461658	-1.616881257	-0.72459331	-1.41090218	-0.738664699	-1.452887575	-1.66851132	-2.0426728562	-1.470734891	-1.61630775	-1.006416557	-1.18510412	-0.82506429	-1.85650407
28	0.447485629	-0.46193418	-1.103992104	-1.054332483	-0.72459331	-0.69678622	0.333020666	-0.406978311	-0.35830459	-0.7279561238	-0.363764338	-1.07621960	-1.242889259	-1.01928167	-1.02341029	-1.31986834
29	-1.840677730	1.86516109	-1.422032698	-1.557216993	-1.47686771	-0.83411622	-0.656227363	-0.870506281	-1.73355704	-1.5951097133	-1.776106078	-1.69939823	-1.065534733	-0.58814329	-0.37878577	-1.10521405
30	-1.233086330	1.23337052	-1.263012401	-0.619635704	-0.31426182	-0.44959224	-0.532571359	-0.406978311	-0.84150140	-0.1685021952	-0.096564549	-1.36703630	-0.296998453	-0.72080125	-0.97382379	-0.67590547
31	-0.858189508	0.84376633	-1.024481955	0.104858929	0.98512124	-0.50452423	0.291801998	-0.038533003	-0.35830459	-0.0146523648	-0.363764338	0.04550193	6.383355366	0.07514652	-0.47795878	0.18271169
32	-1.607983151	1.61244486	-1.263012401	-0.636682636	0.30123542	-0.88904821	-0.285259352	-0.870506281	-0.74857893	0.0972384209	-0.172907346	0.08704717	-0.947298382	-0.72080125	-1.12258329	-0.35392404
33	-0.715987266	0.85956110	-1.422032698	-0.372455182	0.30123542	-0.58692223	0.044489991	-0.929932943	-1.25036023	-0.7559288203	-0.401935736	-0.99312912	0.826246880	-1.08561065	0.66253075	0.61202027
34	0.253573480	-0.28292685	-0.308890617	1.349285004	0.23284684	1.39062966	0.745207344	1.292624242	0.73818047	1.1741872336	1.468462784	2.33049024	1.240074107	1.73337106	1.40632827	0.50469312
35	1.572176093	-1.54124306	-0.308890617	2.329483625	0.91673266	1.85755163	1.693236705	1.922546866	2.14130966	2.2651223945	2.250976451	1.95658306	1.890374037	1.66704208	1.50550128	1.57796457
36	2.490026931	-2.39416033	0.327190572	0.437274114	-1.40847913	1.80261964	0.168145994	2.100826855	1.90900350	1.9434363855	0.476006426	-0.12067904	-0.710825680	-0.09067593	0.56335775	-1.31986834
37	0.602615348	-0.61988182	-0.388400766	0.795259697	-0.04070749	1.00610568	0.209364662	0.674586950	0.77534945	0.8804739211	0.781377613	0.54404484	-0.296998453	0.93742328	0.96004976	0.82667456

Showing 1 to 37 of 252 entries, 16 total columns

Z MATRIX (PCA$X)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
1	-2.519266098	0.639629420	1.828389973	0.354805116	0.198342471	-0.326638223	-0.044127097	0.310455301	-0.116000284	0.0675203968	-0.310343003	-0.1206582368	0.311936898	-0.2475614513	0.0567811644	-0.0468047920
2	-1.752463194	2.568320311	0.618031612	0.394893960	-0.015958653	-0.335988192	-0.580322768	-0.522361428	-0.255324256	0.2038389897	0.184113528	0.2000272526	0.064993363	-0.0027267057	-0.0309083473	0.0038939501
3	-1.829808385	-0.959771167	2.856036137	-1.404382115	-1.409199173	0.532257754	0.039529592	-0.137195196	-0.210966259	-0.0403842191	-0.516227523	-0.0694386693	0.149203077	0.1787506958	0.2605752210	-0.0736956316
4	-0.670371032	1.862274182	0.759338970	0.288553051	0.474194544	-0.409963554	-0.438943010	-0.248990116	0.223292841	-0.5453519693	0.122481606	-0.0516395488	-0.187442678	0.1889607439	-0.0603669909	0.0450969331
5	0.800083830	-0.041233901	2.119131968	-1.900253214	-0.537629253	0.730004277	1.094389287	-0.050904565	-0.398726353	-0.4847874299	-0.593915522	0.0215987559	0.494173494	-0.4496568938	-0.0378667211	-0.0904774326
6	2.870458047	2.294945652	1.002185083	-0.671775689	-0.117624555	0.597410534	0.345498498	0.188746228	-0.162876560	-0.0339075006	-0.170320621	0.0636743807	-0.166012538	0.1534911838	0.0306650510	0.0399220492
7	-0.388704522	0.245372888	1.432285148	-0.410069208	0.016346963	-0.157702298	-0.476048995	-0.179146966	0.241711832	-0.6016061612	-0.649045589	-0.0441002544	-0.021768058	0.2296341090	0.0015412208	0.0208551060
8	-0.675002987	1.891695864	0.367689502	-0.111037501	0.027177477	-0.088208925	-0.057045832	-0.785071028	-0.590896599	-0.2192480375	-0.485118110	-0.3392772724	0.027079217	-0.1914003545	0.0228111709	0.0149204661
9	-0.198321705	3.440603622	0.568181881	0.982698505	0.652775924	-0.150471445	0.005073190	0.536877642	0.708291375	-0.2362476607	0.228105391	-0.3049185835	-0.131059235	0.1377191011	-0.1707873285	0.0075231553
10	1.773765851	3.180682988	0.275894145	0.543462951	-0.188270665	-0.220677404	-0.099062423	0.225113842	-0.688716756	0.8189564519	-0.547683617	0.3023956885	0.064582739	0.1358080192	0.1281639701	0.0008889770
11	-0.366124972	3.184771144	0.057520790	0.073384649	-0.468075248	0.148951911	-0.273942506	-0.049682428	0.312582364	0.0870286671	-0.462695266	-0.2370529138	-0.155965968	0.0944997109	-0.0474262485	0.0408710957
12	1.990850456	3.727036010	0.301923191	0.288466862	-0.323494303	-0.306277375	-0.357454035	0.839567846	0.401415945	-0.3639777384	0.637976772	0.0124031180	-0.282803996	0.0497645639	-0.1678730470	0.0345297642
13	0.243000665	-0.069730245	1.424298337	-0.057107066	0.761853646	-0.705296567	0.009628479	-0.132630068	-0.210818444	0.1158657604	0.177151712	-0.1034462197	-0.142985867	0.2685724561	0.1569752853	0.0179823364
14	3.138282974	1.272409740	1.065586493	0.448309147	0.607602268	0.070738740	0.587103467	0.142212906	-0.075577966	-0.1802769538	0.011685990	0.2907723479	0.237904920	-0.2337081607	0.1585759598	-0.0074241090
15	1.884486299	0.470139950	1.152902499	0.697170974	0.829583129	0.061066131	0.053345926	0.646958783	-0.439030800	0.7188869553	0.451889427	-0.7467367597	0.170923704	-0.2916383598	-0.0478247673	-0.0350340139
16	-0.998961985	-1.118370071	1.999911766	-0.206912346	-0.265145306	-0.765845135	0.335526580	0.101505490	0.021850712	0.2374203162	-0.096506704	-0.4623673239	0.252601550	-0.0545085086	0.1714611084	-0.0328723720
17	2.201250113	-0.092222297	1.646946572	-0.477272825	1.091376562	0.620221684	0.651655242	0.569122925	0.270862033	0.6383107199	-0.034747782	0.2075320206	0.081825691	0.1570581105	-0.0034477271	-0.0552157766
18	3.771834093	1.298673401	0.607169641	0.940850730	0.684749556	0.210347560	-0.417848251	0.584310706	-0.522391415	0.2460929949	0.050781752	-0.0471472228	-0.065326226	0.1636598658	0.0777108645	0.0005823208
19	1.209175029	0.989085215	1.390198141	1.483595411	0.595192265	-0.109409052	-0.014897606	0.670582250	0.078677533	-0.6778516947	-0.240496583	-0.3965717697	0.026223589	0.2954224318	0.1964071435	-0.0049490831
20	2.671006440	1.732054175	0.794995749	0.001538716	0.376318696	-0.644145928	0.126009713	0.625177794	0.611576097	0.1343961713	0.189889634	0.2264197611	-0.090867590	-0.2061394067	0.0728930082	0.0229770165
21	0.870030330	0.309719921	1.366409648	0.834314631	0.728517194	-0.356315600	-0.310424213	-0.662160905	-0.390488309	-0.1539317348	0.260482768	-0.0219859779	-0.019823312	-0.0502178319	0.3653262316	0.0068773777
22	2.938372961	1.912791896	0.134085563	1.571813645	0.107223323	0.220368358	-0.531586696	-1.398753248	-0.244323603	-0.2320068705	-0.388104288	-0.2620726939	0.074820232	-0.0675614645	0.2726062085	0.0499320724
23	-4.305217072	-0.187974001	1.637779062	-0.704934684	-0.386286033	0.255836050	0.014685874	0.022135657	-0.420844703	0.0278951965	0.576437755	-0.0127795842	-0.039176656	0.2777651599	-0.0245372762	-0.0224920382
24	-3.486067454	0.002398607	1.005433559	-0.743210378	0.204077516	0.434368500	-0.100534350	0.245246722	-0.574206065	0.2141059144	0.025861998	0.3561278686	0.140860270	0.0896104149	-0.1239844699	-0.0298511096
25	-2.974025311	0.588657197	1.462073452	0.391391462	-0.134757962	0.635889450	-0.123145972	0.287479447	-0.431627468	-0.5411959084	0.450799643	-0.0569891842	0.035270113	0.4180794842	-0.0596323606	-0.0146637402
26	-3.424722990	2.124339077	0.528004975	0.311620907	-0.018989984	-0.316416810	-0.049429074	-0.333756862	0.133701725	-0.3021717159	0.162370910	0.4410627086	0.161670819	-0.0325598250	-0.0429294797	-0.0075950369
27	-5.323756548	0.257642712	0.662728754	0.554410607	0.365022625	0.099163680	-0.572174611	0.053503635	0.199171238	0.6722841970	-0.366990699	-0.1463279902	0.119409755	-0.0847827813	-0.1406021090	-0.0297311601
28	-2.199068557	-1.168367158	1.376238291	-0.415163141	0.714203305	-0.326148918	-0.903889630	-0.296409722	-0.680262404	0.8587909538	-0.349192654	-0.1525013582	-0.247606286	-0.1260085200	0.3096264533	-0.0289503188
29	-4.808350441	0.711623398	0.880954231	1.889181721	0.337964977	-0.176926588	-0.915246716	-0.034015784	0.225942594	-0.1042412828	-0.776335838	0.1511102661	0.133963299	0.1528033373	0.1152060879	-0.0354290203
30	-2.532592272	0.915310276	1.065413483	0.316216733	-0.360960778	-0.806144647	-0.861774410	-0.077501497	0.032801149	-0.2835016299	0.387656592	-0.1583486936	-0.222337944	0.1094023252	0.1525787740	0.0098631391
31	0.606479687	3.351033578	-0.281869152	-0.839506946	-4.223983439	2.859115489	-2.025515761	0.873174707	0.917633411	0.6911574236	0.022006599	0.1730343729	-0.353093613	-0.3053711243	0.2898979216	-0.0634716908
32	-2.472958301	1.659432619	0.604065258	-0.057890996	-0.215429443	-1.553954513	0.351187683	-0.298360648	-0.337410450	0.0802354038	-0.170568066	0.2490419891	-0.058200158	-0.0751090632	0.3685250215	-0.0026904936
33	-1.745832331	1.890413487	0.108987216	0.448395348	-0.152487846	1.101971520	-0.935374169	-0.880188779	-0.813802164	0.0539487164	0.295370449	-0.0845162423	0.024392461	0.2764570168	-0.2518240699	-0.0874960765
34	4.282693285	1.163194274	0.530185776	0.474052084	-0.307816541	0.127432538	1.006680375	0.290350970	0.774425725	0.1435550628	-0.525749056	-0.2062629826	0.253310854	0.2415920562	0.0081402727	0.0144218625
35	7.049967249	0.639727042	0.351716252	-0.534333299	-0.264552623	0.207760934	-0.021148942	-0.121340292	0.019079997	-0.0254876164	0.233385855	-0.0127172584	-0.116047381	-0.1468817722	-0.0410201879	0.0020549205
36	3.260035676	-3.732820582	1.836143832	-0.450420468	0.767023517	-0.094550171	-0.554553374	-0.874328956	1.139121780	-0.2727492523	0.210003566	0.3729962865	-0.281127453	0.2706514081	0.7043695697	-0.0654942649
37	2.522257462	-0.052283642	0.493211217	0.364261869	0.688349449	-0.132738948	0.259976594	-0.108516257	-0.210389832	-0.6761347063	0.037752464	0.0889926674	0.203513978	0.0436175455	-0.0364363013	0.0215806321

Showing 1 to 37 of 252 entries, 16 total columns

EIGEN$VECTORS & EIGEN$VALUES

```
> pca <- prcomp(data, center = TRUE, scale. = TRUE)
```

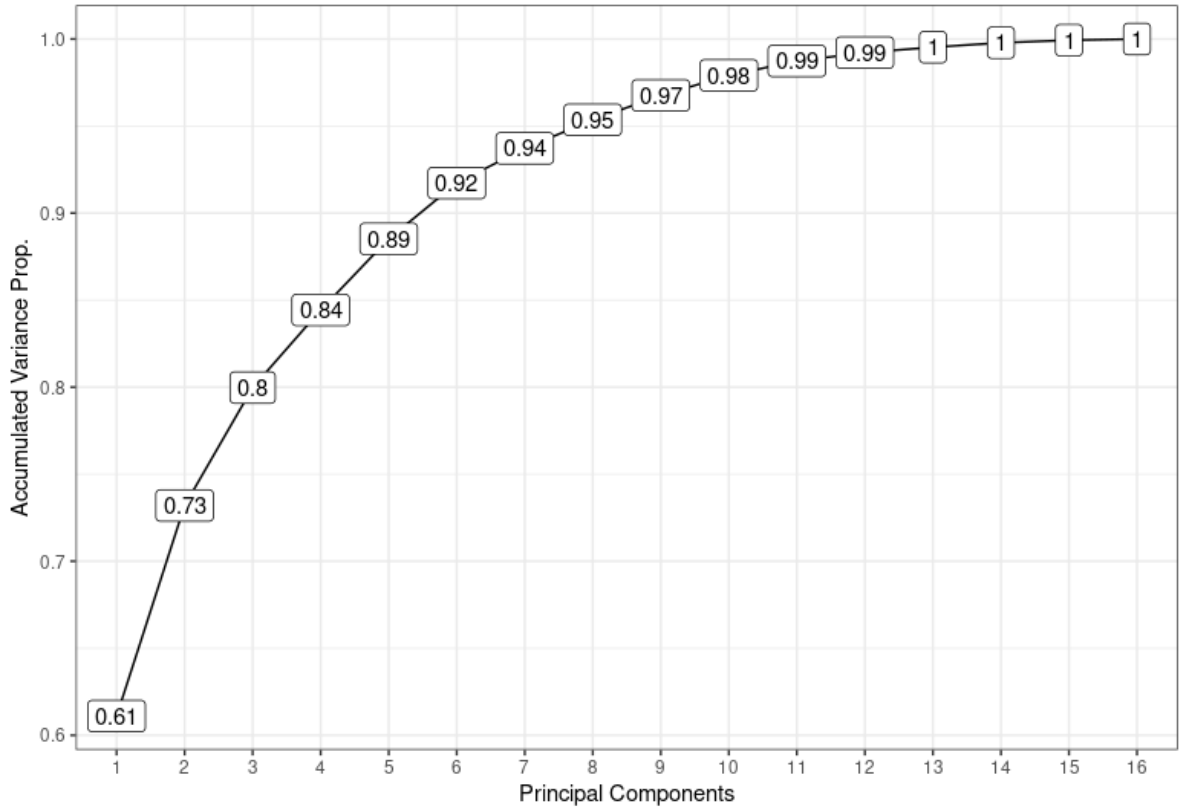
```
> pca$sdev ^ 2 # Eigen Values
[1] 9.77673901 1.93527587 1.08406418 0.71308511 0.65387211 0.51301292 0.31923652 0.26233662 0.22416034
0.18430129 0.13208077 0.07709517 0.04965932 0.04058785 0.02345977 0.01103317
```

```
> pca$rotation # Eigen Vectors
```

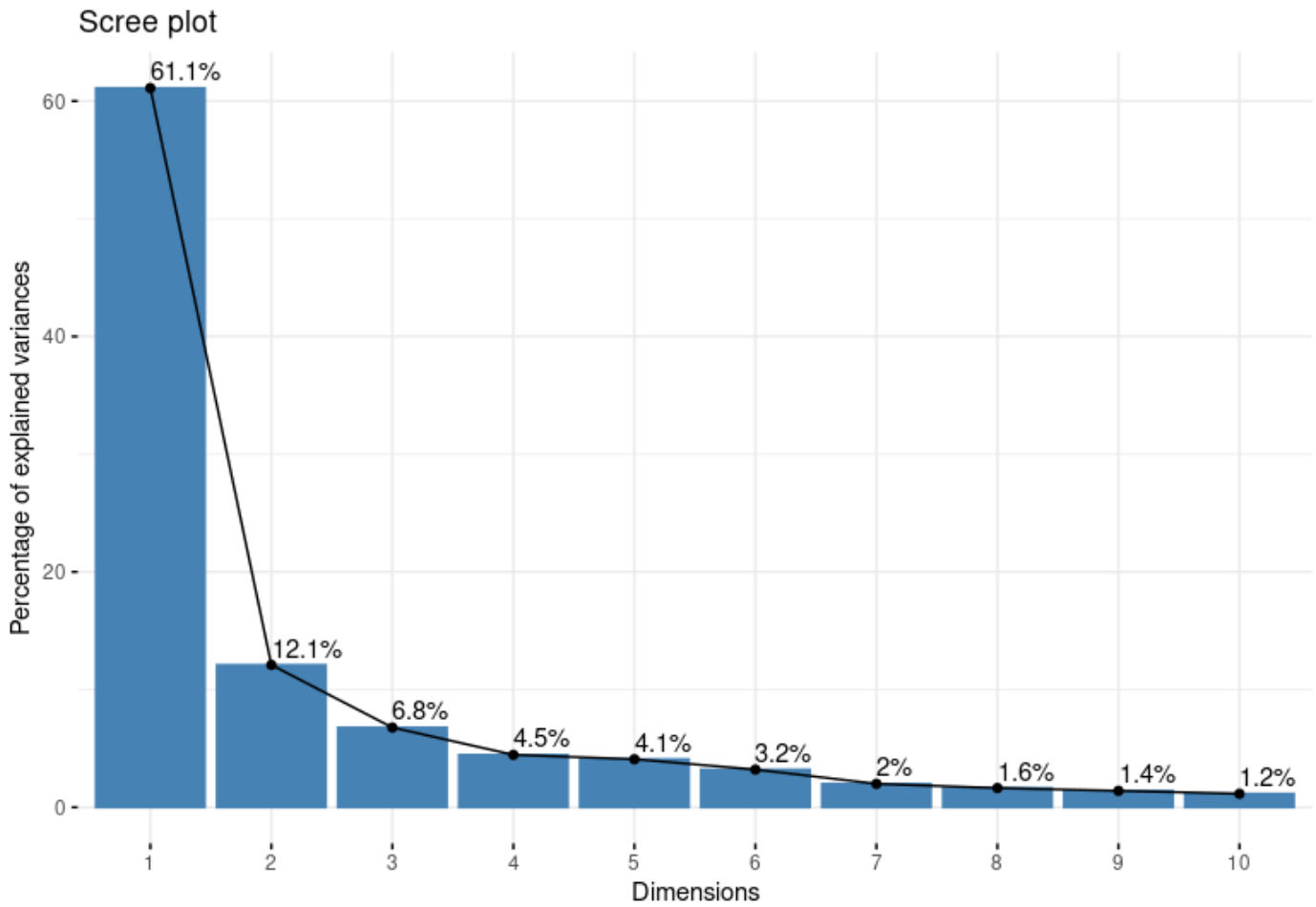
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
BODYFAT	0.23288808	-0.41096554	0.099700851	-0.287154779	0.134970132	0.289372075	-0.02773416	0.01223896	-0.22289342	0.02857962	-0.046314945	0.02306917	-0.086559082	0.114138580	-0.057333020	-0.710364368
DENSITY	-0.22812621	0.41418638	-0.117191684	0.294814120	-0.132425382	-0.315524493	0.02537825	-0.04904323	0.23634008	-0.04689151	0.050867097	-0.01382871	0.006673985	-0.084729134	-0.024343107	-0.697830237
AGE	0.02704686	-0.44116593	-0.706529347	0.073247997	-0.125422654	-0.035582211	0.27388965	0.15718468	0.25581970	0.11402451	0.302982189	-0.06227602	-0.046913023	0.072407881	-0.040723986	0.012190599
WEIGHT	0.30952265	0.09697852	0.018709169	-0.068772995	-0.025280970	-0.163176560	-0.02916258	-0.06070495	0.11428681	-0.06116268	0.009168349	0.18488738	-0.188299574	0.032780818	-0.874834758	0.070519208
HEIGHT	0.06601034	0.45935163	-0.337705028	-0.694947252	0.331422522	-0.023700277	-0.06383230	0.06212410	0.09918199	-0.03016399	0.132309319	-0.07106318	0.148944439	0.079649579	0.085958564	-0.009260225
ADIPOSITY	0.29808427	-0.13048434	0.098676457	0.139310201	-0.065461322	-0.147785363	-0.20997460	-0.05108397	0.16287069	-0.12466851	0.073835628	-0.05842572	0.766535354	0.390852682	-0.011944591	-0.027042981
NECK	0.27178834	0.08400583	-0.190438254	0.191794964	0.138325238	-0.211589669	-0.38765520	-0.03318384	-0.20684371	0.74882612	-0.083738156	0.08672212	-0.065926222	0.016753097	0.089723646	-0.002313431
CHEST	0.29473701	-0.09379545	-0.054416224	0.020593125	0.045151534	-0.140094968	-0.26372035	-0.13509424	0.39998175	-0.25353173	-0.384484170	-0.42760128	-0.418189952	0.134989492	0.206823164	0.025839594
ABDOMEN	0.29770225	-0.18600806	-0.007112912	-0.112277179	-0.007458984	-0.148099729	-0.09756727	-0.12694298	0.13505893	-0.10483096	0.046375404	0.12253615	0.201286784	-0.846978131	0.107525675	-0.025460305
HIP	0.29675137	0.03672213	0.166815537	-0.058977628	-0.117432665	-0.243465841	0.11653636	-0.10534511	0.05856668	-0.13310493	0.364084444	0.56915685	-0.300723865	0.253848379	0.385823340	0.012747376
THIGH	0.27964521	0.10274086	0.307077752	0.007212733	-0.055714371	-0.162642186	0.25146248	0.08214210	-0.17905379	0.13604748	0.498017240	-0.63606337	-0.092891047	-0.073653596	-0.010996731	-0.003915589
KNEE	0.27235885	0.13198699	-0.011117645	-0.117977609	-0.208971999	-0.059049164	0.67967026	-0.13548109	0.01241168	0.22754520	-0.531644744	0.01979244	0.158914160	0.031791370	0.074071573	-0.013234319
ANKLE	0.19627520	0.26083183	-0.034257003	-0.071443542	-0.675460435	0.549103321	-0.25270336	0.12405745	0.16450137	0.11256290	0.077547675	-0.00670446	-0.029225965	-0.044378152	0.036669753	-0.014022704
BICEPS	0.26629974	0.12082127	0.047176375	0.231447174	0.231096334	0.032722149	0.07675122	0.85595470	0.04910907	-0.13535566	-0.149302669	0.12187173	-0.001166518	-0.067896395	0.051643665	-0.022640811
FOREARM	0.21701211	0.19316560	-0.038348514	0.388841877	0.479152395	0.536887269	0.18798688	-0.35348278	0.22978253	0.01442646	0.171198893	0.03622766	0.011173372	-0.031353586	0.006340546	0.004727290
WRIST	0.24006835	0.16665045	-0.431988876	0.202578698	-0.112779515	0.005902735	-0.03685044	-0.13386162	-0.66382629	-0.45670456	-0.050971864	-0.04748396	-0.001179995	-0.007367905	0.026219304	0.008423323

ANALYSIS

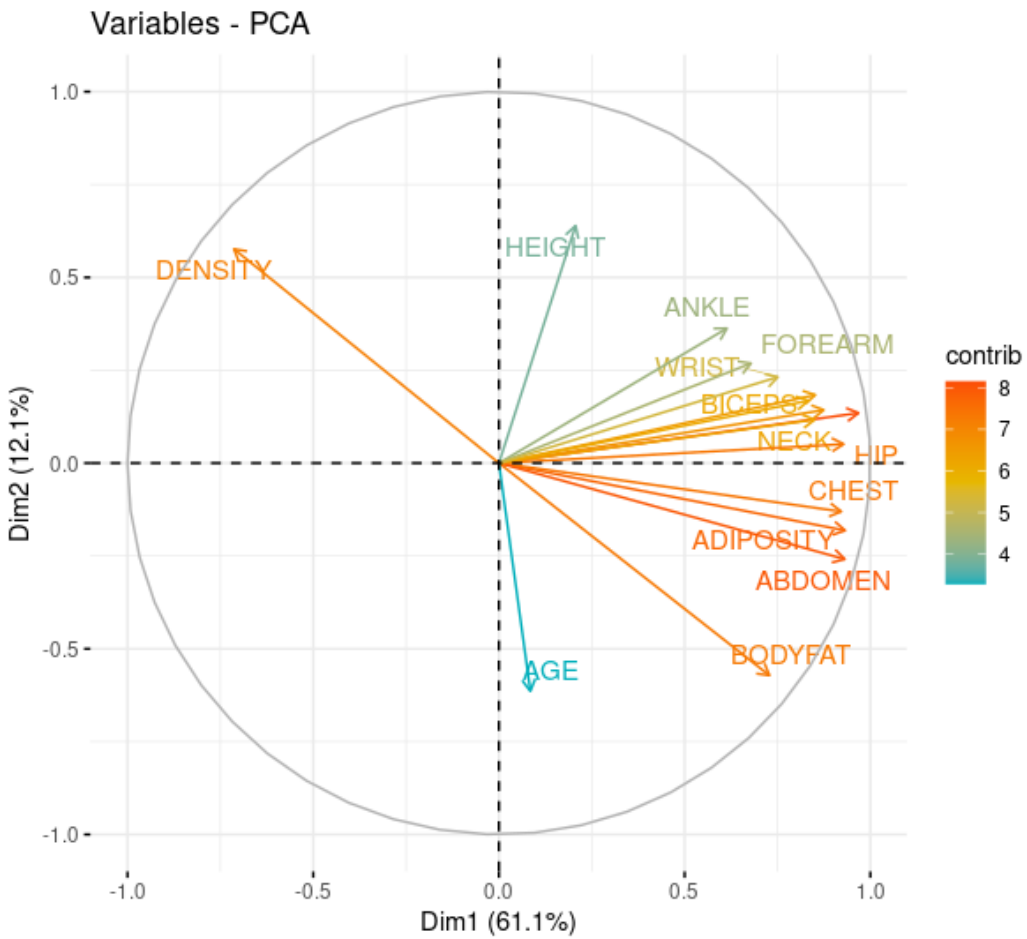
```
> prop_varianza <- pca$sdev ^ 2 / sum(pca$sdev ^ 2)
> prop_varianza_acum <- cumsum(prop_varianza)
```



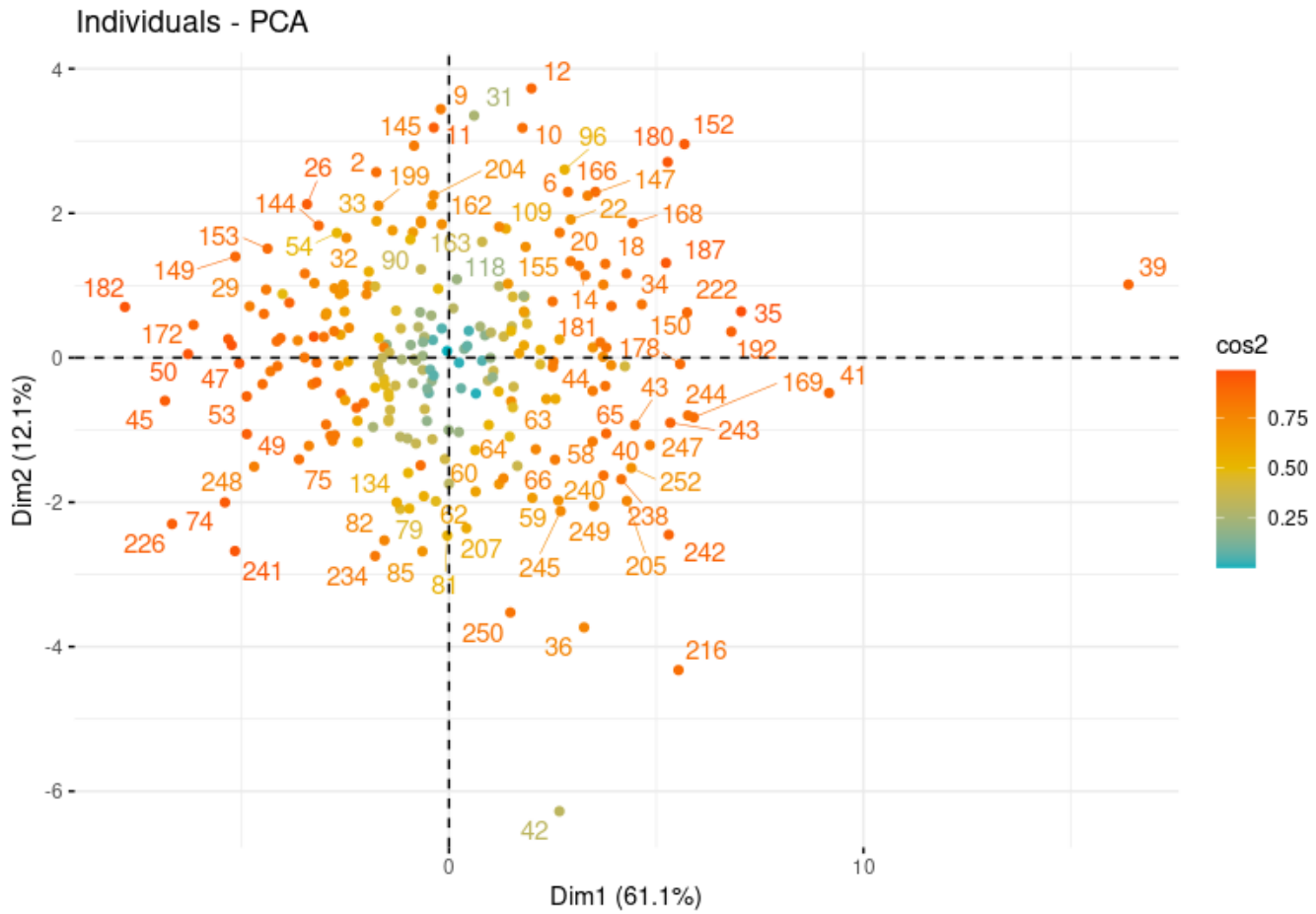
```
> library(factoextra)
> fviz_eig(pca)
```



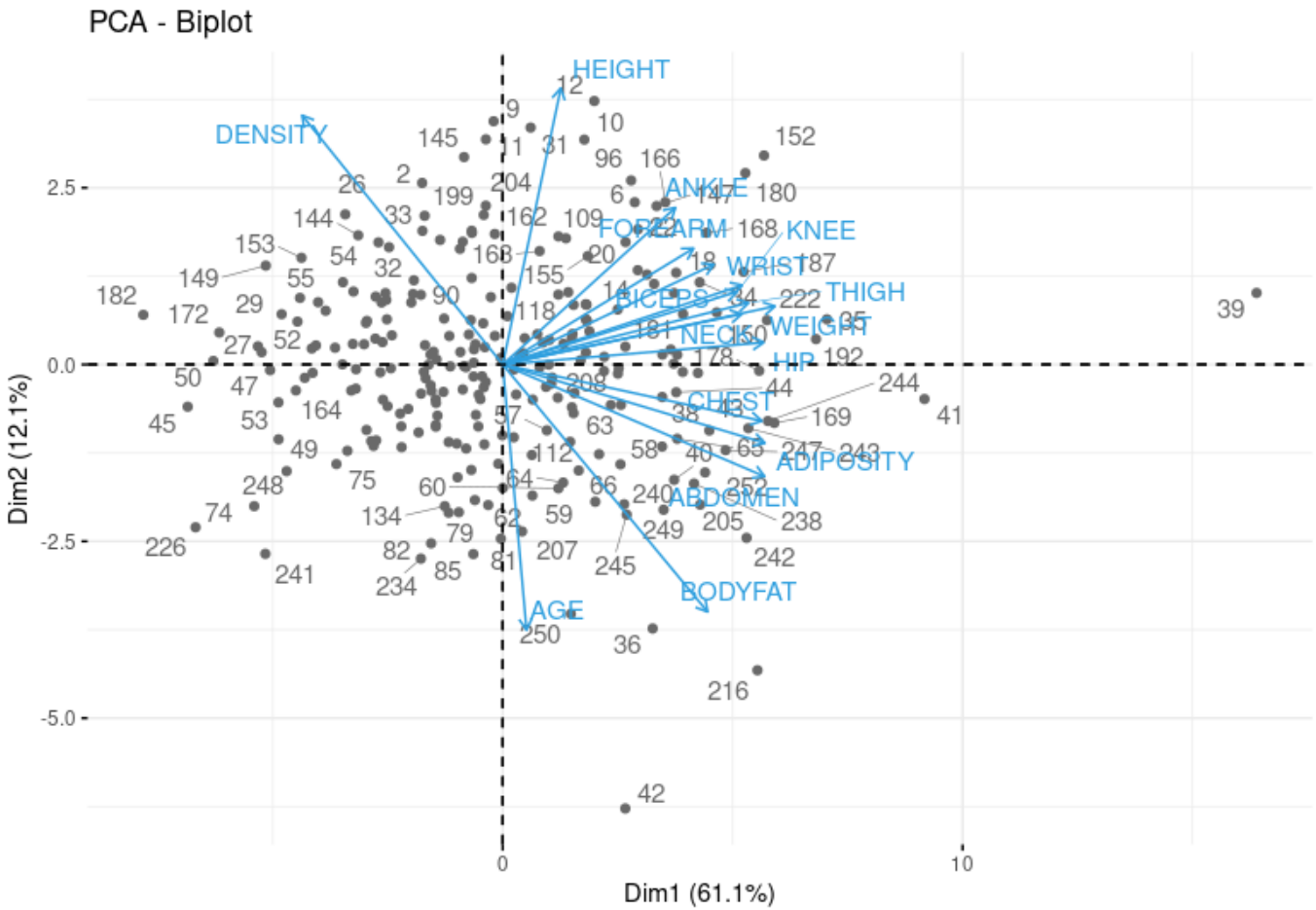
```
> fviz_pca_var(pca, col.var = "contrib", # Color by contributions to the PC
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE # Avoid text overlapping )
```

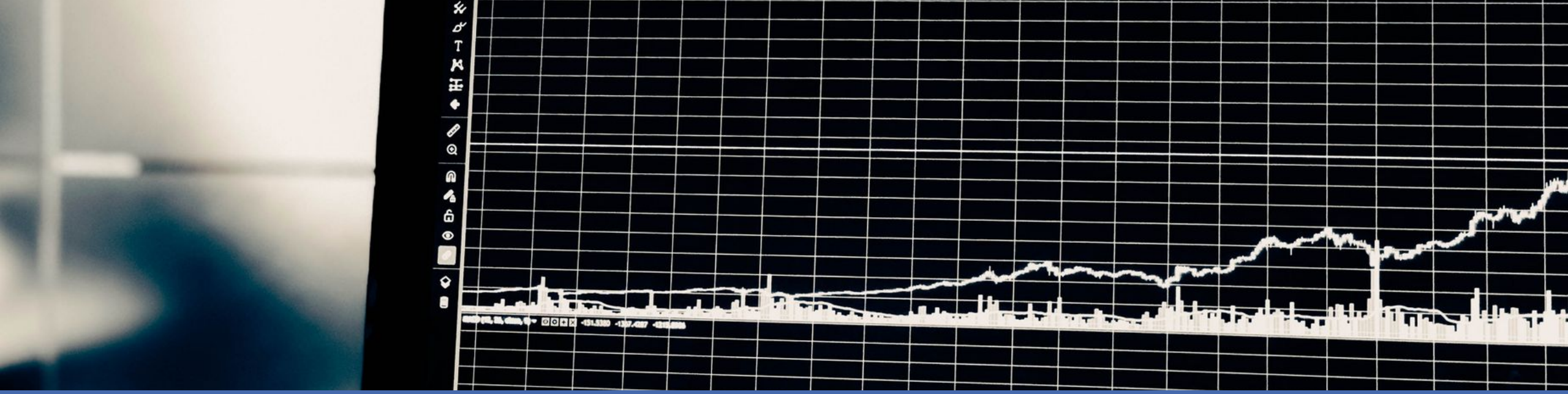


```
> fviz_pca_ind(pca, col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE # Avoid text overlapping)
```



```
> fviz_pca_biplot(pca, repel = TRUE, col.var = "#2E9FDF", # Variables color
                  col.ind = "#696969"  # Individuals color)
```





EXPECTATION - MAXIMIZATION

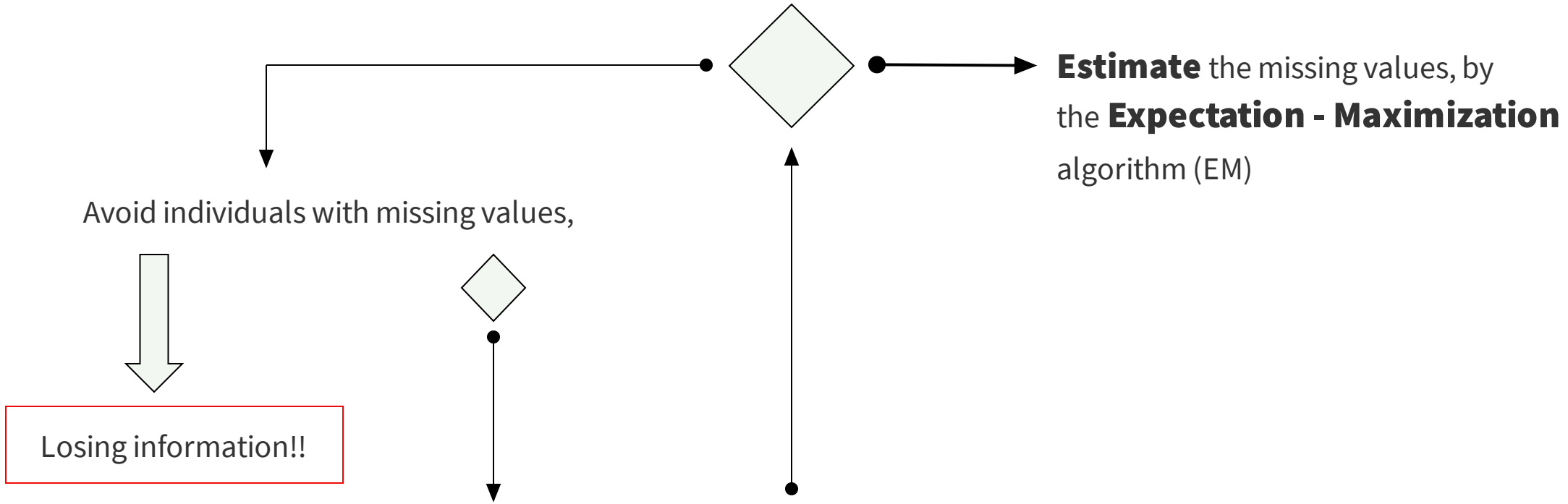
Francesc Contreras & Albert Pérez

- Ideal dataset: filled matrix
- Incomplete data encompass:
  - Typical missing values
  - Latent inferred variables

INCOMPLETE DATA?

Let's consider a data-set with missing values,

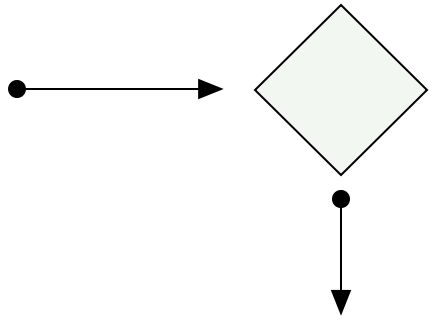
How do we approach this?

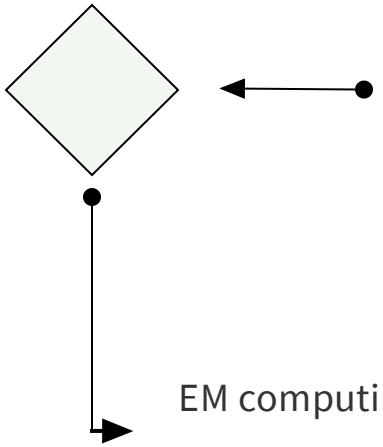


EXPECTATION - MAXIMIZATION ALGORITHM

There was an **old intiution**,

> "alternate between estimating unknows and the hidden variables"





Maximum Likelihood from Incomplete Data via the EM Algorithm

[Paper by A. P. Dempster; N. M. Laird; D. B. Rubin, 1977]

EM computing **the density distribution** and,

explained the relationship to the **Maximum Likelihood Estimation** method.

CONVERGENCE PROVED

EXPECTATION - MAXIMIZATION ALGORITHM

Estimation of the model parameters assuming the presence of missing values.

What is the EM algorithm?

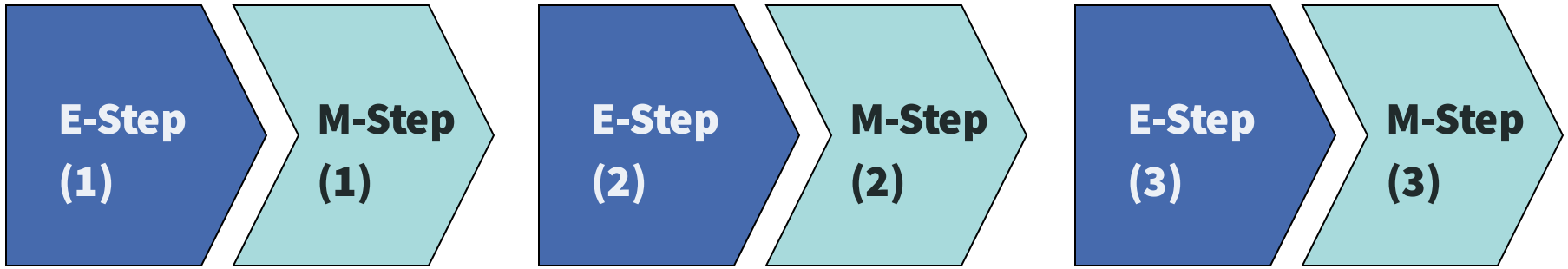
- **Iterative** method
- Based on the Maximum Likelihood Expectation (MLE)
- Allows the estimation of the model parameters, **including the latent variables as if they were observed**
- Two steps involved:
  - Expectation
  - Maximization

> "General approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data."
>
> [Original Paper EM definition by A. P. Dempster; N. M. Laird; D. B. Rubin, 1977]

ITERATION STEPS INVOLVED

1. Expectation Step (E-Step)
    - Estimate probabilities of latent states given current parameters (sample)
    - Obtain expected values as if they were observed
2. Maximization Step (M-Step)
    - Obtaion MLE of parameters given expected sufficient statistics from the E-Step

Then, the parameters found on the M-Step are used to begin another E-Step

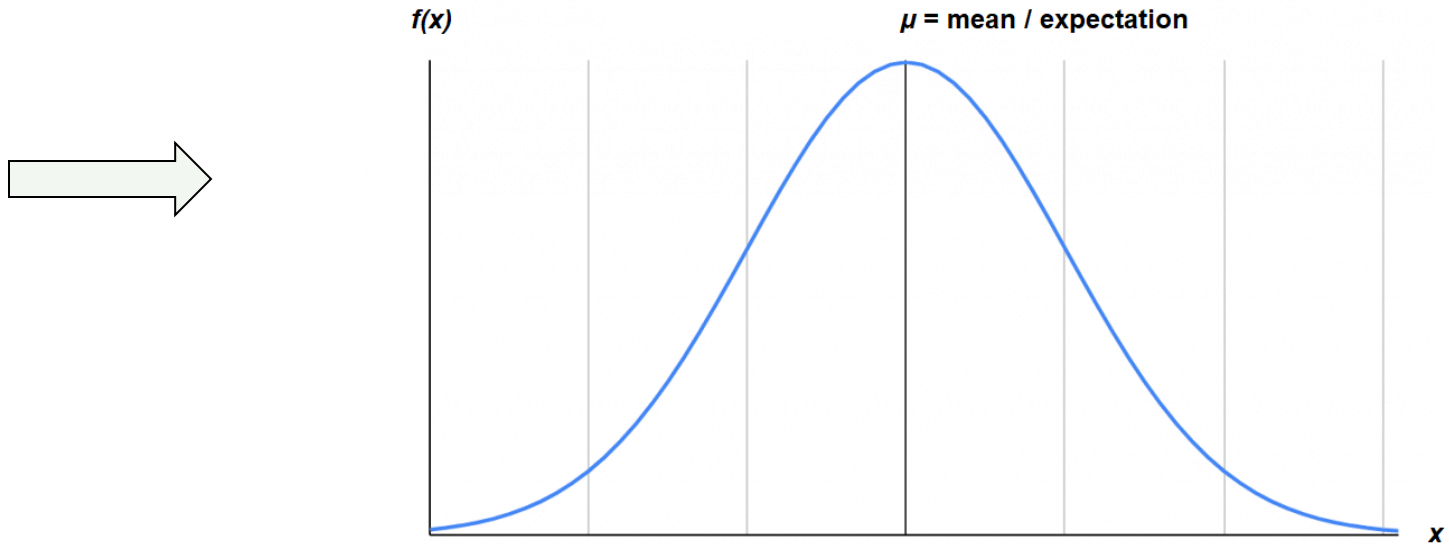


CLUSTERING APPROACH

Cluster classification and recognition?

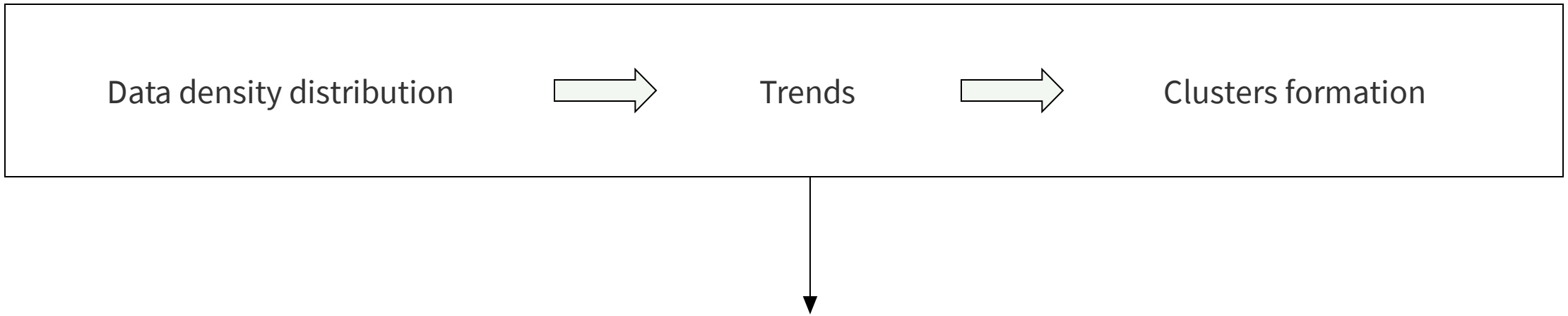
How can this be?

The Data



- Normal density distribution

CLUSTERING APPROACH



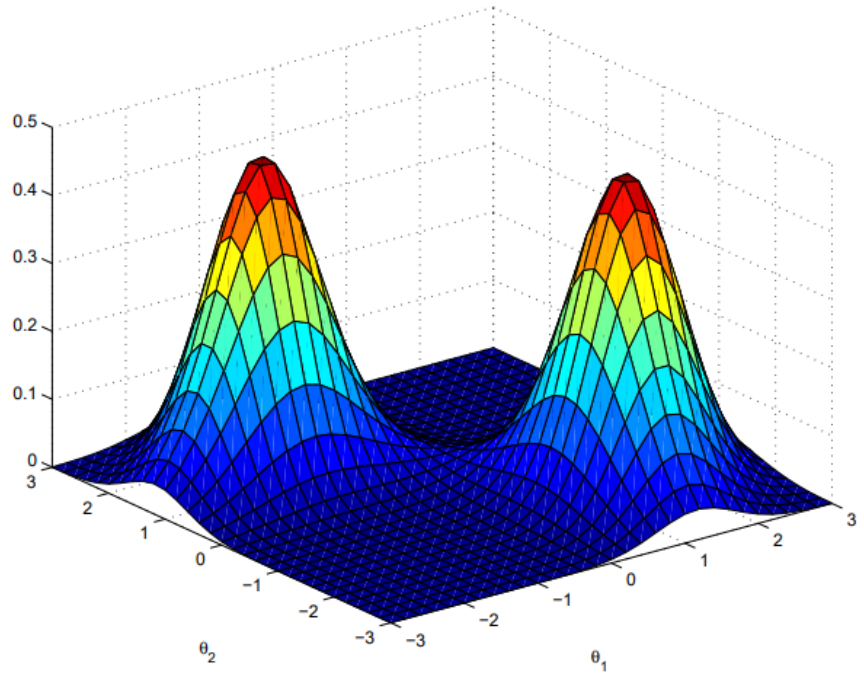
CLUSTERING APPROACH

Probabilistic Mixture Models

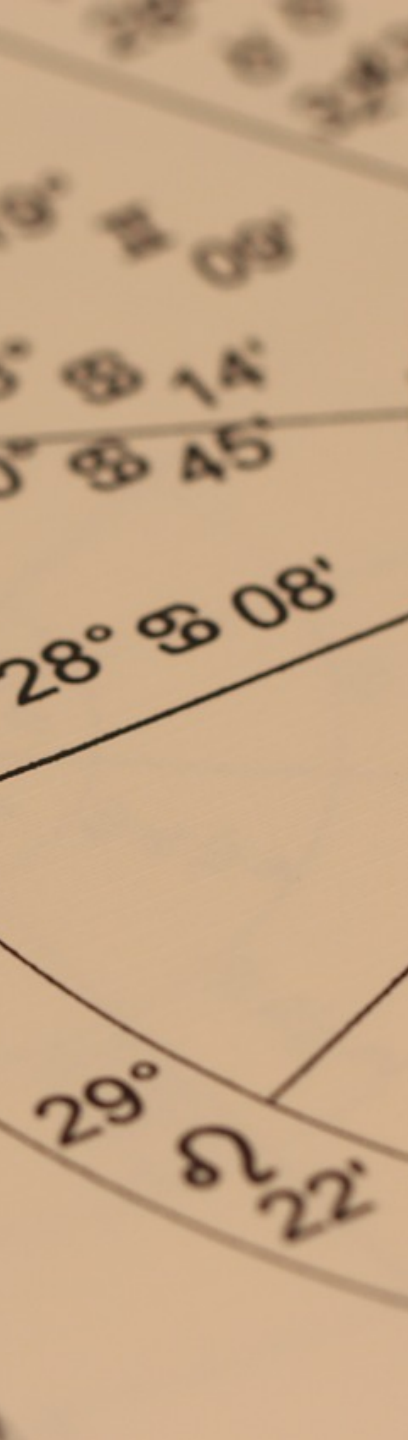
$$\Theta = \pi 1 f1(x) + (1 - \pi 1) f2(x)$$

where $fi(x) \sim$ **Normal distribution** expressed as $N(\mu i, \sigma i)$, and

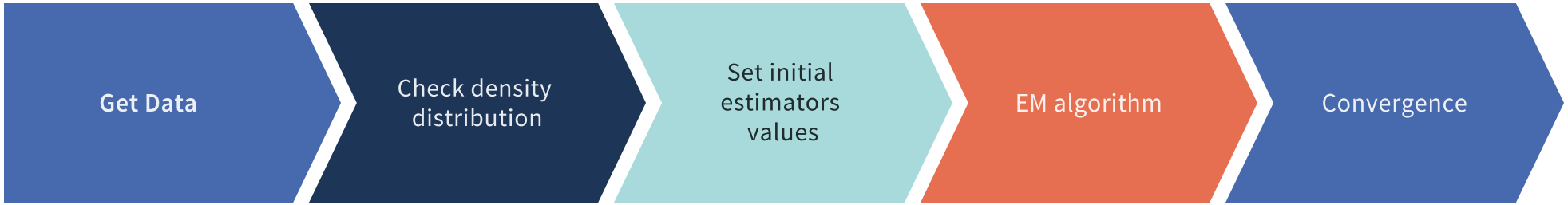
$0 \leq \pi 1 \leq 1$ being the proportion of individual of the sub-population



- Bimodal data distribution representation



ANALYSIS STEPS

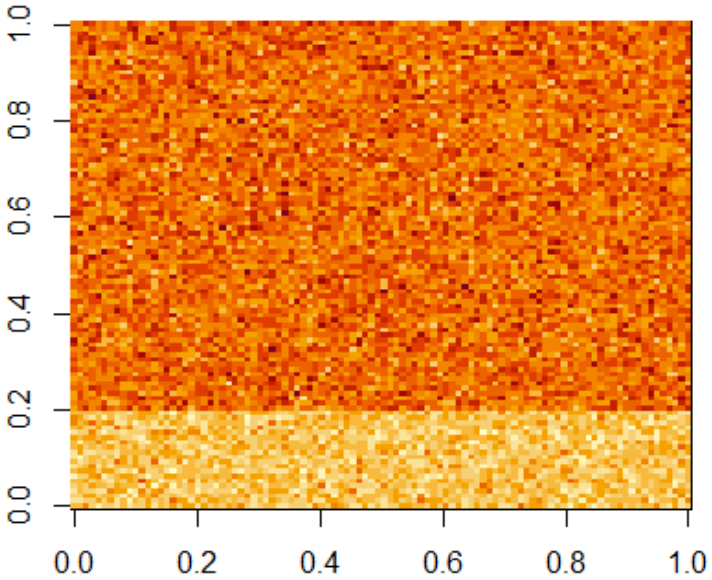


DATA-SET ANALYSIS

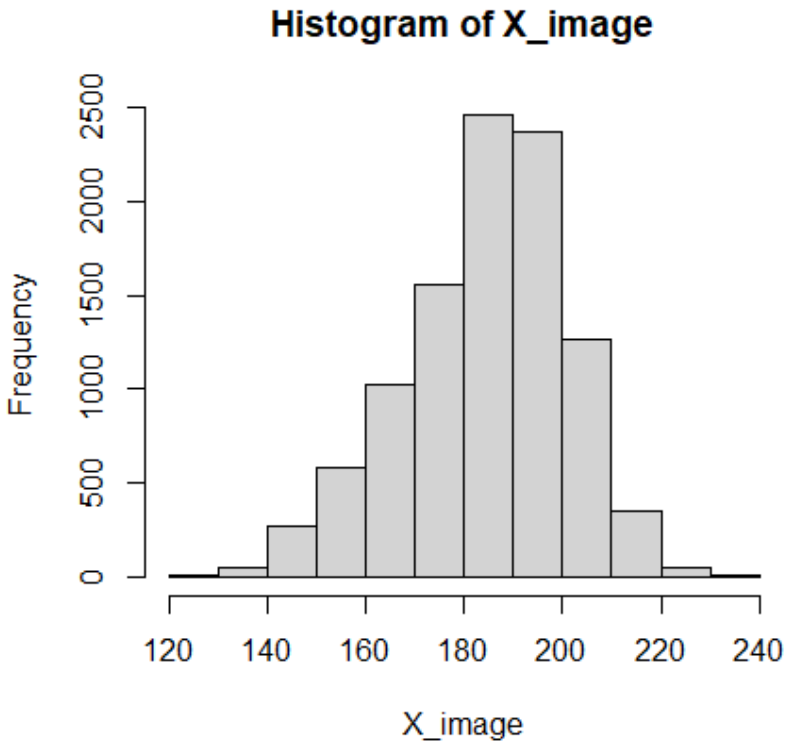
```
> data <- read.table(file.path(data_dir,"Picture1.data")) # read
> str(data) # show structure of the data
'data.frame':   10000 obs. of  1 variable:
 $ X: num  163 165 180 154 136 ...

> X <- data$X
> mean(X) # central expectation
[1] 2.377471
> var(X) # how much can the value be away from the expectation?
[1] 1.737588

> head(X)
       X
1 162.7507
2 165.1432
3 180.1964
4 154.2615
5 136.1654
6 159.7606
```

DATA-SET ANALYSIS

```
> class(X) # what kind of values
[1] "numeric"

> n <- length(X) # how many individuals?
> n
[1] 10000

> X_image<-matrix(X, nrow=sqrt(n), ncol=sqrt(n), byrow=F)
> class(X_image)
[1] "matrix" "array"

> image(X_image)
```



DATA-SET ANALYSIS



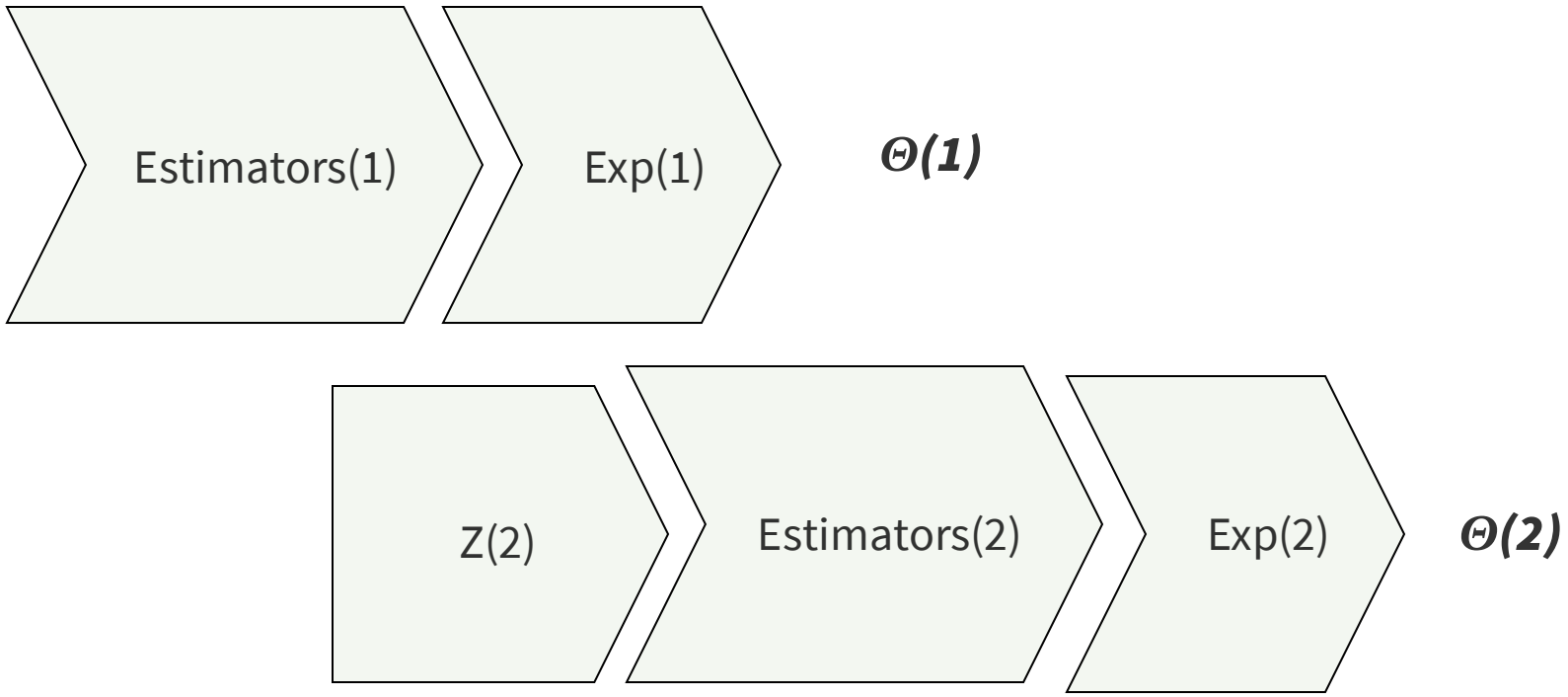

```
# setting the initial estimators parameters
> Mu1<-185 # expected value for the first cluster
> Mu2<-165 # expected value for the second cluster
> V1<-272 # variance of the initial whole data-set
> V2<-272
> pi1<-0.7 # mixture model proportion
> pi2<-1-pi1
```

DATA-SET ANALYSIS

```
for(0 to max_num_simulations) {
    if not first iteration then {
            Z1 <- round(Exp1, 4)
            Z2 <- round(Exp2, 4)

            # Maximization of the likelihood step
            MLE(Z1, Z2)
        }
    }

    # Calculate the expectations
    Exp1 <- c()
    for (i in 1:n) {
        f1 <- dnorm(x[i], mean=Mu1, sd=sqrt(v1))
        f2 <- dnorm(x[i], mean=Mu2, sd=sqrt(v2))
        Exp1[i] <- (pi1*f1) / (pi1*f1 + (1.0 -
pi1)*f2)
    }

    Exp2 <- 1 - Exp1
}
```

```
function MLE(Z1, Z2) {
      # Maximize the expectations from the previous
step
      Mu1<-round((t(Z1) %*% X)/sum(Z1),4)
      Mu2<-round((t(Z2) %*% X)/sum(Z2),4)

      # Adjust the variance
      B1<-X-mx1 %*% Mu1
      B2<-X-mx1 %*% Mu2
      V1<-round((t(B1) %*% diag(array(Z1)) %*%
B1)/sum(Z1),4)
      V2<-round((t(B2) %*% diag(array(Z2)) %*%
B2)/sum(Z2),4)

      # calculate the probability proportion
      pi1<-round(sum(Z1)/n, 4)
      pi2<-round(sum(Z2)/n, 4)
}
```

DATA-SET ANALYSIS



DATA-SET ANALYSIS

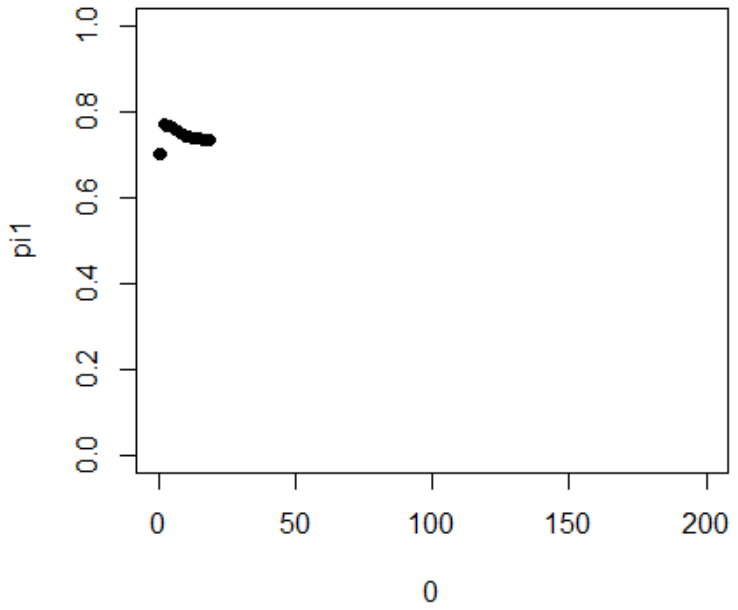
```
# iteration convergence control
if(abs(pi1-pi0) < epsi) {
    print("The algorithm has
converged!!")
}
```

$$|| \Theta(i) - \Theta(i+1) || < epsilon$$

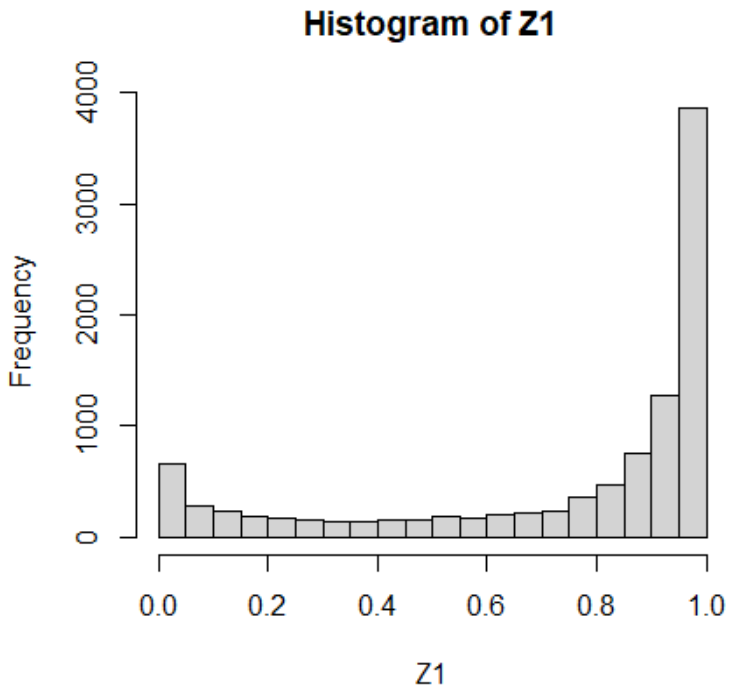
where $\Theta(i)$ is the estimated model at iteration ***i***, *and* **epsilon** the permissive difference between expectations.

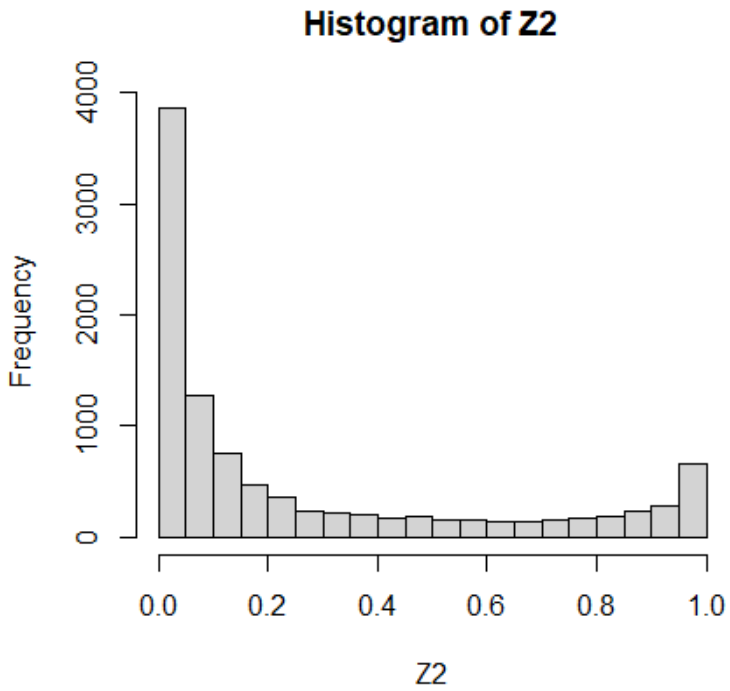
DATA-SET ANALYSIS

```
> Mu1; Mu2; V1; V2; pi1;
191.1512
165.6269
137.82
169.0573
0.7356
```



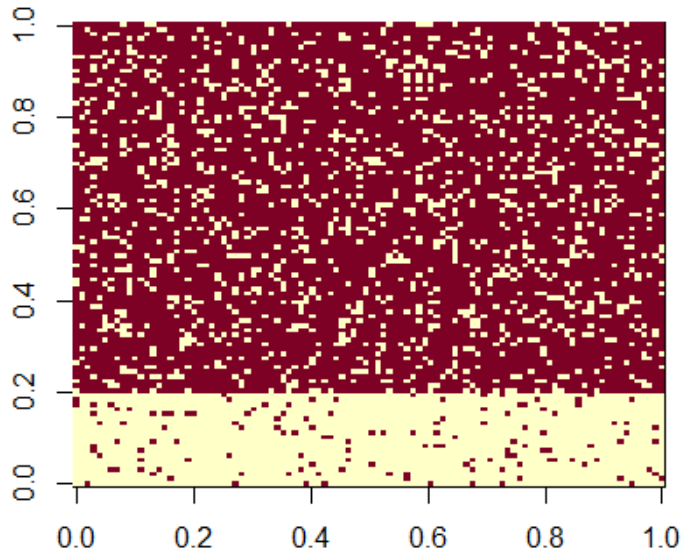
DATA-SET ANALYSIS





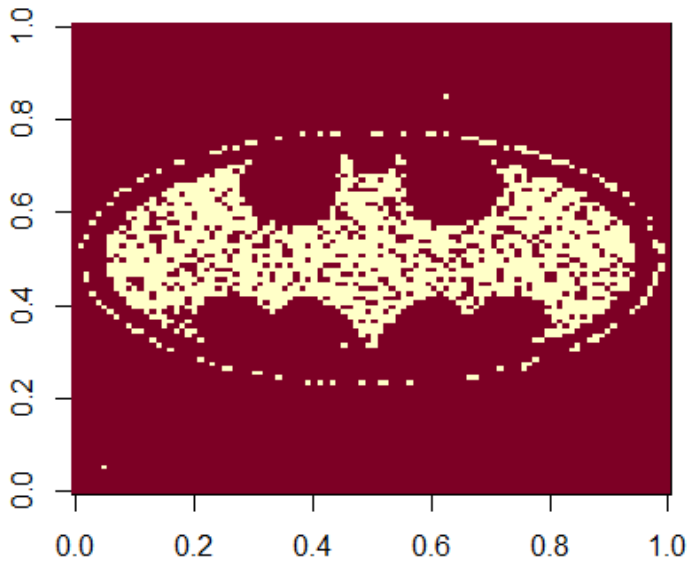
DATA-SET ANALYSIS

```
> Z11 <- array(rep(0,n))

# assign 0 or 1 in terms of Z1 probabilities
> for(i in 1:n) {
+  if(Z1[i] >= pi1) {
+    Z11[i] = 1
+  }
+ }

> Z11_image<-
matrix(Z11,nrow=sqrt(n),ncol=sqrt(n),byrow=F)
> image(Z11_image)
```



DATA-SET ANALYSIS

```
> Z11 <- array(rep(0,n))
> data2 <-
read.table(file.path(data_dir,"Batman.data"))
> X2 <- data2$X

# assign 0 or 1 in terms of Z1 probabilities
> for(i in 1:n) {
+  if(X2[i] >= pi1) {
+    Z11[i] = 1
+  }
+ }

> Z11_image<-
matrix(Z11,nrow=sqrt(n),ncol=sqrt(n),byrow=F)
> image(Z11_image)
```



WHY THE PRE-ANALYSIS IS SO IMPORTANT?

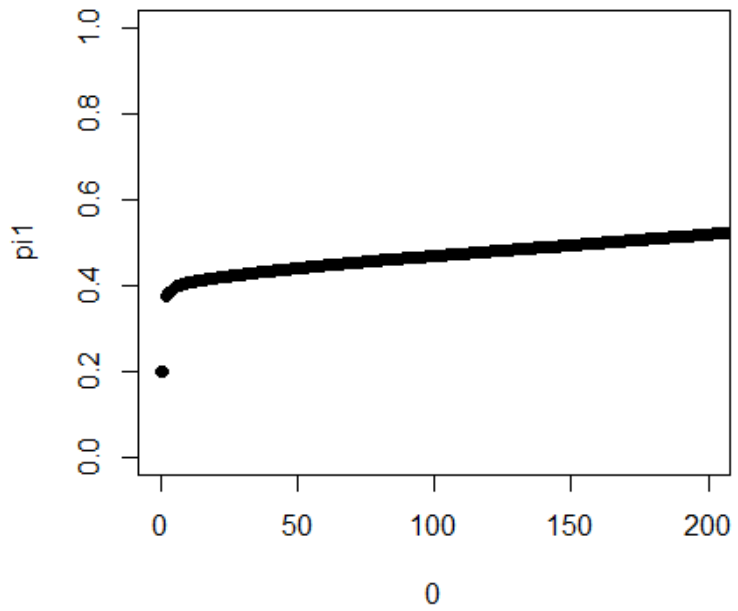
May the algorithm converge in a non-infinite time?

What happens if we do not estimate accurately the clusters parameters distribution?

1st Case

Assuming a deviated mixture model proportion

```
# setting the initial estimators parameters
> Mu1<-185 # expected value for the first cluster
> Mu2<-165 # expected value for the second cluster
> V1<-272 # variance of the initial whole data-set
> V2<-272
> pi1<-0.2 # mixture model proportion
> pi2<-1-pi1
```

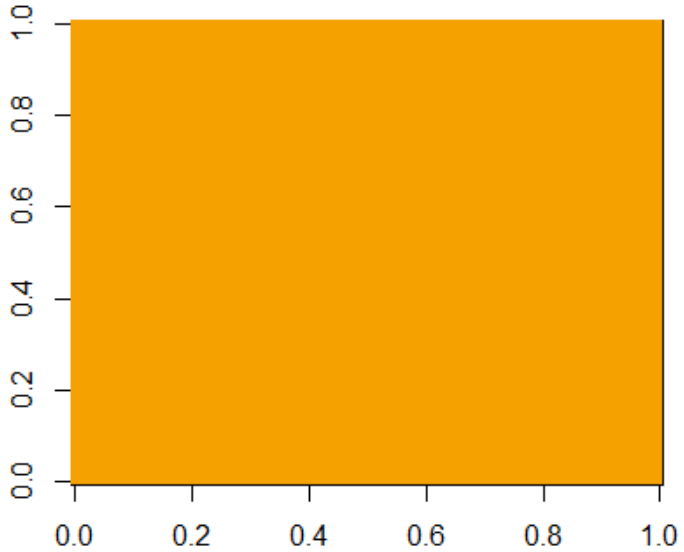


WHY THE PRE-ANALYSIS IS SO IMPORTANT?

2nd Case

Assuming a deviated expectation values

```
# setting the initial estimators parameters
> Mu1<-165 # expected value for the first cluster
> Mu2<-165 # expected value for the second cluster
> V1<-272 # variance of the initial whole data-set
> V2<-272
> pi1<-0.7 # mixture model proportion
> pi2<-1-pi1
```



No differentiation ⇨ No classification

OTHER CONSIDERATIONS

Why is important the **Convergence control**?

Do you really want to wait to **infinite time**?

Sorry, I don't have that **much!!**

```
# small value of epsilon for convenience
epsilon <- 0.0001
```

CONCLUSIONS

Many Thanks!!

REFERENCES

[1] Dr. Carles Comas. Big data project Master's Degree in Computer Engineering

[2] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society

[3] Hartley, H. (1958). Maximum likelihood estimation from incomplete data.

[4] Francesc Contreras, and Albert Pérez. Github Repository: https://github.com/elskater98/BigDataProject