Principal Component Analysis (PCA)

Problem:

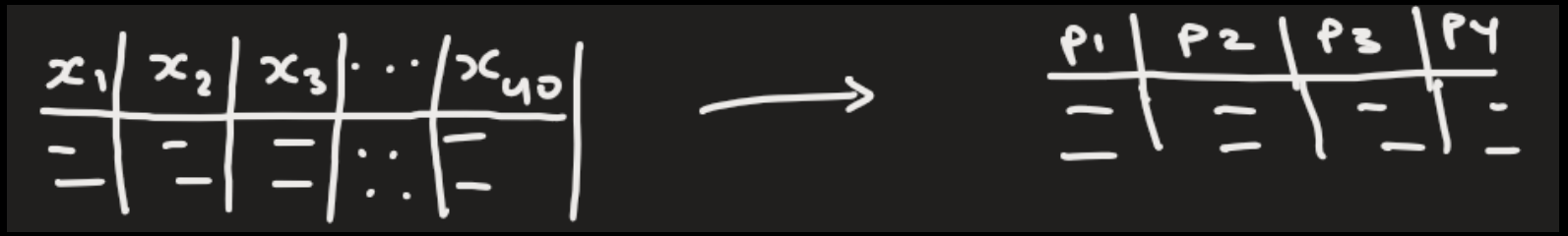
Imagine you are analyzing data from **1,000 sensors** in a smart city.
Each sensor records temperature, humidity, pressure, traffic density, energy usage, and other correlated measurements. Say there are 40 features.

But you face a big challenge:

- **Too many features → slower ML models**
- **Most features are correlated → redundant information**

Is there a way to **compress this high-dimensional data into a smaller number of powerful features (say 4)** that still explain most of the patterns?
In other words, can you reduce the number of columns without loosing important information ?

This is the real-world problem that **Principal Component Analysis (PCA)** solves.



Principal Component Analysis (PCA)

The feature that has most variance captures the most information that is contained in a feature space.

Here,
feature x1 has most variance (10 to 100),i.e. has more spread,
feature x2 has moderate amount of variance (8 to 11), and
feature x3 has 0 variance (all 5)

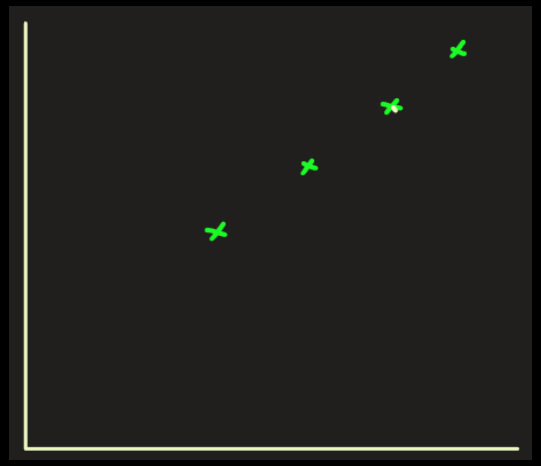
Feature x1 contains most information about our feature space, feature x2 contains the 2nd most information and feature x3 contains no information.

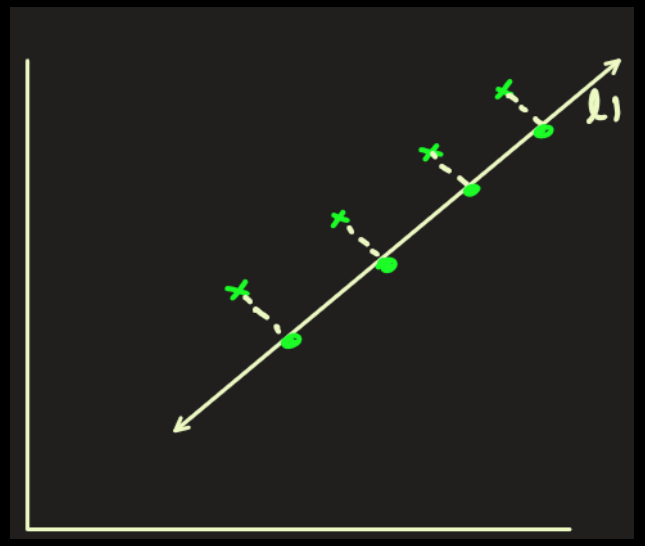
So, we can remove x3 from our feature space and would not loose any essential information.

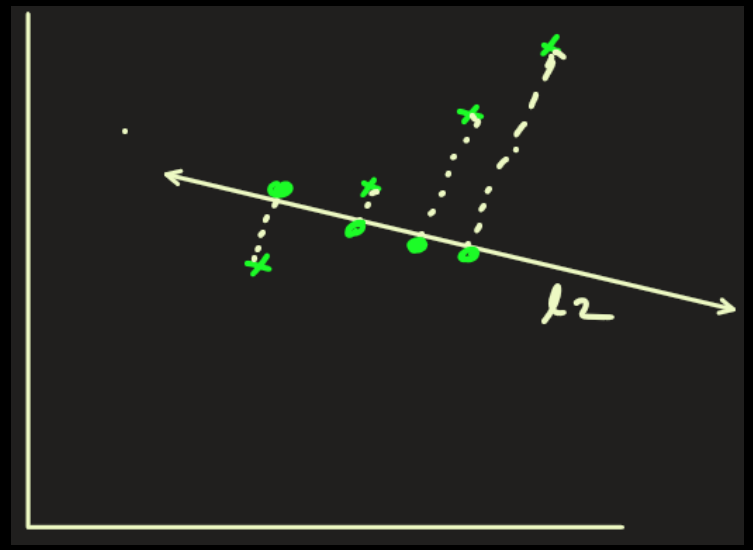
(age) x_1	commute time x_2	work hour x_3
15	11	5
20	10	5
10	9	5
70	8	5
50	10	5
100	9	5
60	8	5
26	10	5

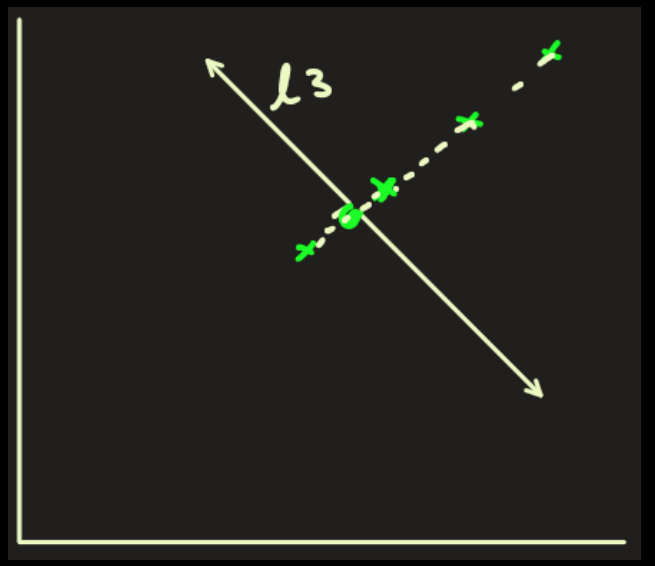
Principal Component Analysis (PCA)

Suppose I have 4 data points. I project these 4 points on 3 different lines as shown below.
We see that projected points on line1 has most spread (variance) and that on line3 has no spread, i.e. no variance



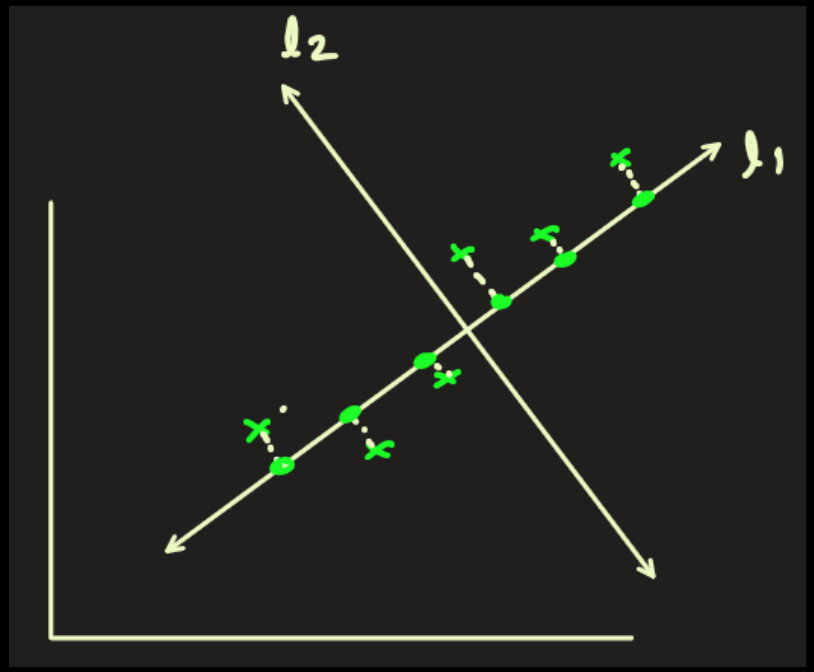


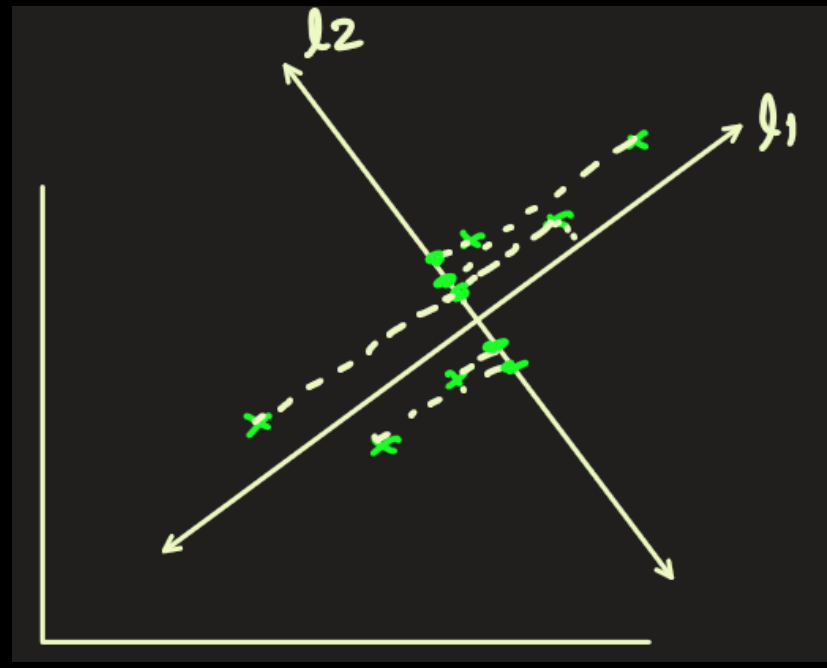




Principal Component Analysis (PCA)

Now let's take a look at another set of data points.
If we project the 2D data points on line1 and line2, we see that the projection points on line1 has more variance (aka spread) than that of line2.
So, the projected points on line1 contains the more information about your feature space.
The direction of line1 is along the axis that has most variance in the data.
The direction of line2 is along the axis that has least variance in the data. Also, note the line2 is perpendicular to line1

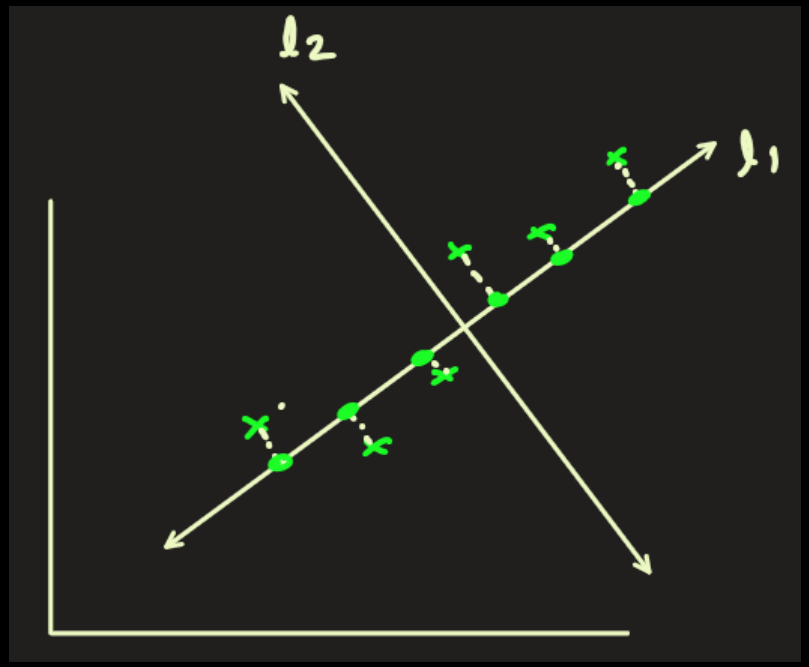


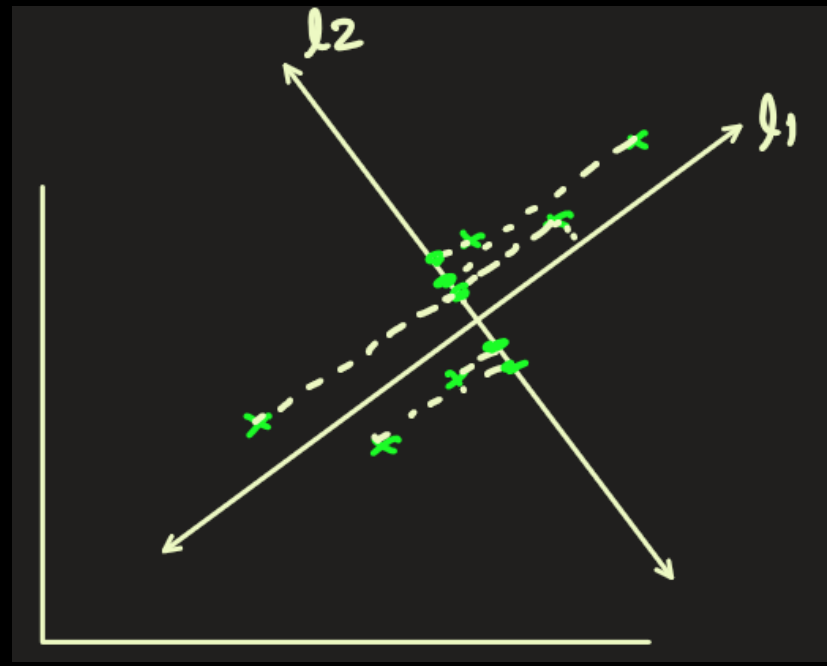


Principal Component Analysis (PCA)

These lines, line1 and line2, are called principal components (PC). They capture the most variance in your data.

The PC1 captures the maximum variance and PC2 captures the 2nd most variance.

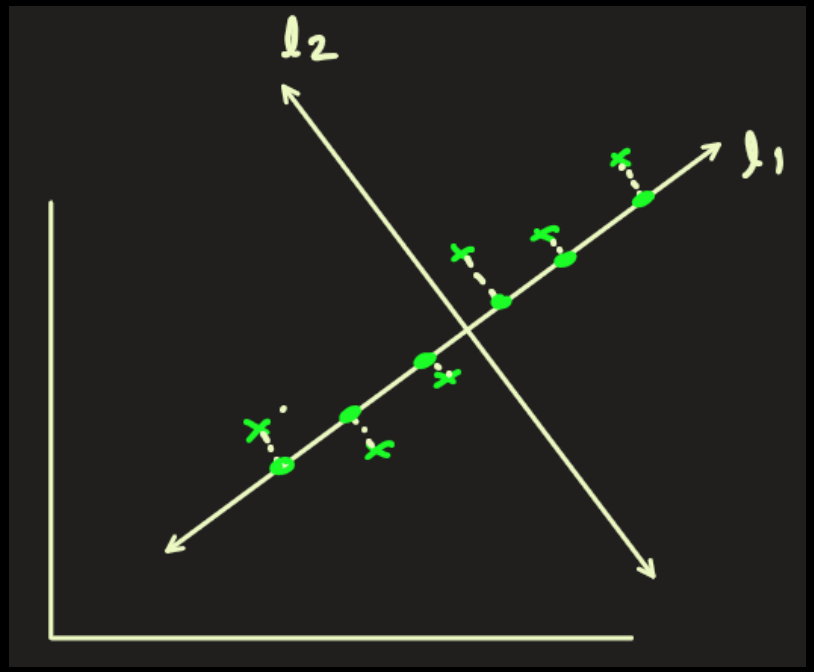


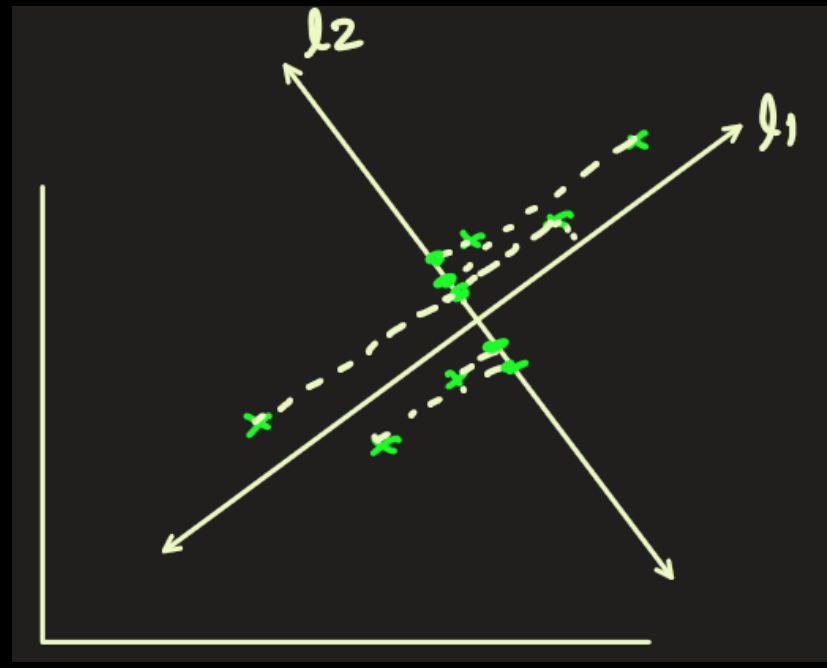


Principal Component Analysis (PCA)

Geometrically speaking, principal components represent the directions of the data that explain a **maximal amount of variance**, that is to say, the lines that capture most information of the data.

The relationship between variance and information here is following:
The larger the variance carried by a line -> the more information is contained along that line.





Principal components are new variables that are constructed as **linear combinations or mixtures of the initial variables**.

So, if you have features such as age, income, experience then PC is linear combination of age, income and experience. In other words,
PC1 = a1 x age + a2 x income + a3 x experience
PC2 = b1 x age + b2 x income + b3 x experience
PC3 = c1 x age + c2 x income + c3 x experience

Here a1, a2, a3, b1, etc. are numbers determined during PCA.

These combinations are done in such a way that the
1) principal components are uncorrelated and
2) most of the information within the initial variables is squeezed or compressed into the first PC1, maximum remaining information in the second PC2 and so on

Side Note:

Principal components are less interpretable and don’t have any real meaning since they are constructed as linear combinations of the initial variables.

How many components should we select in PCA ?

Let’s answer this by looking at an example:
Breast cancer dataset.

How many components should we select in PCA ?

Example: The breast cancer data set has 30 features and 1 target column. We apply PCA to feature columns only.

mean radi	mean text	mean peri	mean area	mean smc	mean com	mean con	mean con	mean sym	mean frac	radius err	texture er	perimeter	area error	smoothne	compactn	concavity	concave p	symmetry	fractal din	worst radi	worst text	worst peri	worst area	worst smc	worst com	worst con	worst con	worst sym	w
17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	
20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275	
19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613	
11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	
20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	
12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985	
18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063	
13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196	
13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378	
12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366	
16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948	
15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792	
19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176	
15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809	
13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596	
14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218	
14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029	

How many components should we select in PCA ?

First find all the 30 Principal Components. Then, select the top number of component that would explain the maximum amount of variance.

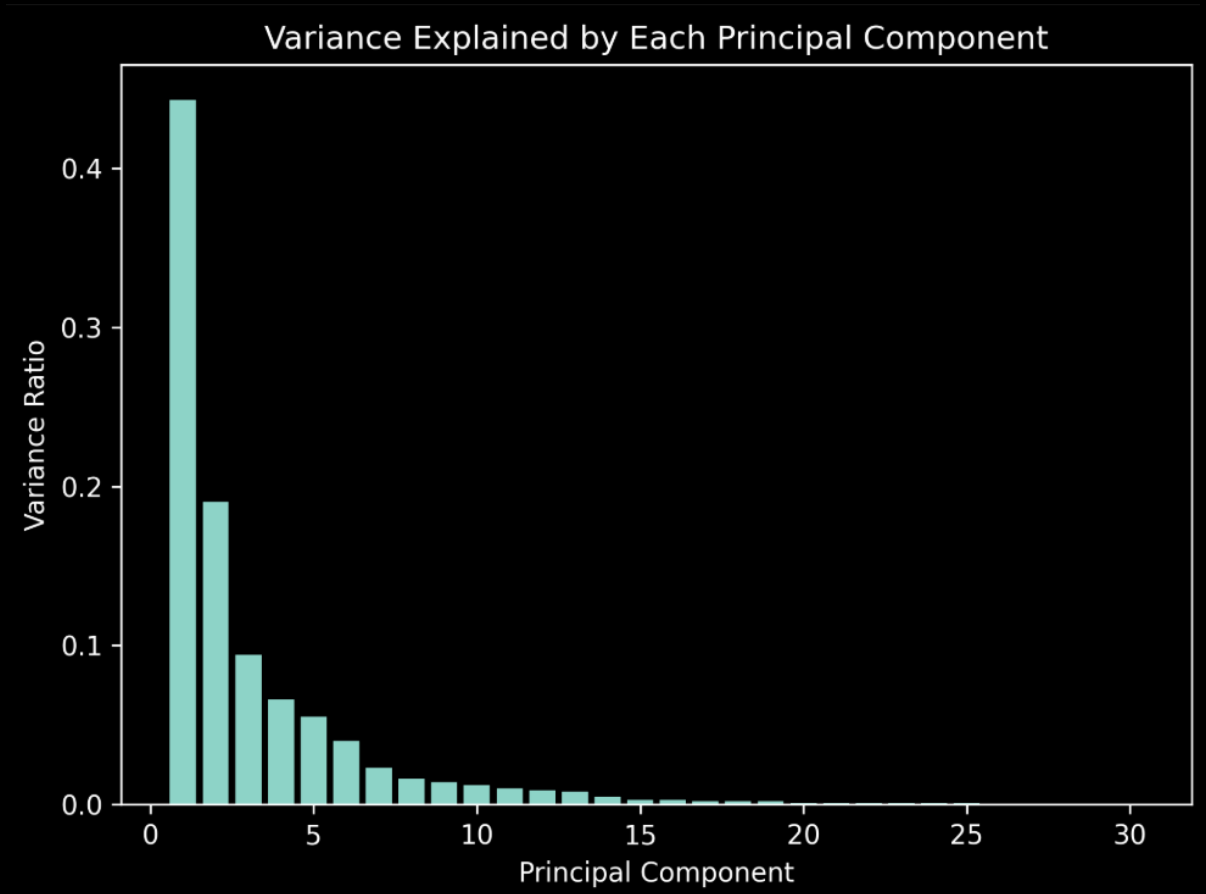
For example, in **breast cancer dataset** there are 30 features. From the graph, we see that

- 5 principal components will explain about 86% + of the variance in the data.
- 10 principal components will explain about 95% + of the variance in the data.

So, 10 component might be a good number.

This way, we reduced our number of columns from 30 → 10.

Principal component	Explained variance (%)
PC1	44.3 %
PC2	19 %
PC3	9.4 %
PC4	6.6 %
PC5	5.5 %
and so on…	….



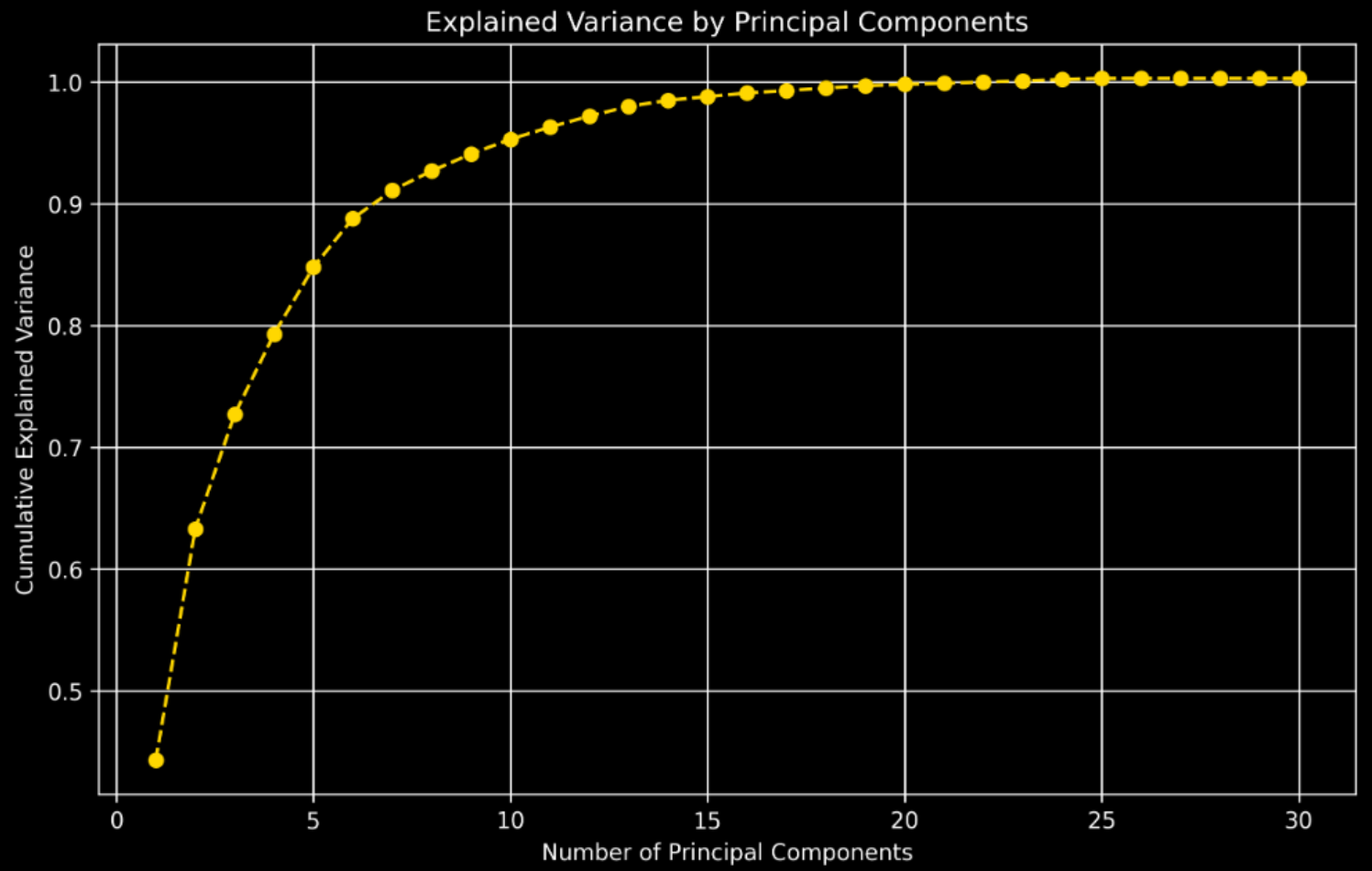
How many components should we select in PCA ?

Side Note:
Alternatively, we can plot the cumulative variance and read the number of components based on how much variance is required.

From the graph, we see that
- 5 principal components will explain about 86% of the variance in the data.

- 10 principal components will explain about 95% of the variance in the data.

So, 10 component seems a good number.



EXTRA