Principal Components

Overview

- Principal Components
- Principal Components Applications
- Principal Components Regression
- Partial Least Squares Regression

Principal Components

Transformation to explain

the variance-covariance

structure of the variables

Principal Components

Transformation to explain

the variance-covariance

structure of the *predictors*

Principal Components

The transformation is a set of

linear combinations

Principal Components

A linear combination of a set of variables

is a new variable

Example:

$$
\begin{aligned}
X &= [X_1, X_2, \ldots, X_p] \\
Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p
\end{aligned}
$$

Principal Components

A linear combination of a set of variables

is a new variable

Example:

$$
\begin{aligned}
X &= [X_1, X_2, \dots, X_p] \\
Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Z_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Z_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

Principal Components - Fundamentals

	X_1	X_2	X_3	$\ldots$	X_p
n					

Principal Components - Fundamentals

X_1	X_2	X_3	$\ldots$	X_p

n

σ_1^2	σ_2^2	σ_3^2	$\ldots$	σ_p^2

$$\sum_{i=1}^{p} \sigma_i^2$$

Principal Components - Fundamentals

	X_1	X_2	X_3	$\dots$	X_p
n					

σ_1^2	σ_2^2	σ_3^2	$\dots$	σ_p^2

$$\sum_{i=1}^{p} \sigma_i^2$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,p} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \cdots & \sigma_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{bmatrix}$$

Principal Components - Fundamentals

	X_1	X_2	X_3	$\dots$	X_p
n					

σ_1^2	σ_2^2	σ_3^2	$\dots$	σ_p^2

$$\sum_{i=1}^{p} \sigma_i^2$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,p} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \cdots & \sigma_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{bmatrix}$$

Principal Components - Fundamentals

.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,p} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \cdots & \sigma_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{bmatrix}$$

$\vec{e}_1$	$\vec{e}_2$	$\vec{e}_3$	$\ldots$	$\vec{e}_p$

λ_1	λ_2	λ_3	$\ldots$	λ_p

Principal Components - Fundamentals

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,p} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \cdots & \sigma_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{bmatrix}$$

$\vec{e_1}$	$\vec{e_2}$	$\vec{e_3}$	$\dots$	$\vec{e_p}$
e_{11}	e_{21}	e_{31}	$\dots$	e_{p1}
e_{12}	e_{22}	e_{32}	$\dots$	e_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
e_{1p}	e_{2p}	e_{3p}	$\dots$	e_{pp}

λ_1	λ_2	λ_3	$\dots$	λ_p

Principal Components - Fundamentals

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,p} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \cdots & \sigma_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{bmatrix}$$

$\overrightarrow{e_1}$	$\overrightarrow{e_2}$	$\overrightarrow{e_3}$	$\dots$	$\overrightarrow{e_p}$
e_{11}	e_{21}	e_{31}	$\dots$	e_{p1}
e_{12}	e_{22}	e_{32}	$\dots$	e_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
e_{1p}	e_{2p}	e_{3p}	$\dots$	e_{pp}

λ_1	λ_2	λ_3	$\dots$	λ_p

$$\sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{p} \lambda_i$$

Principal Components - Fundamentals

.

	PC_1	PC_2	PC_3	$\dots$	PC_p
n					

$$
\begin{aligned}
X &= [X_1, X_2, \dots, X_p] \\
PC_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
PC_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
PC_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

Principal Components - Fundamentals

.

	PC_1	PC_2	PC_3	$\dots$	PC_p
n					

var	λ_1	λ_2	λ_3	$\dots$	λ_p

$$
\begin{aligned}
PC_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
PC_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
PC_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

$$\sum_{i=1}^{p} \lambda_i$$

Principal Components - Fundamentals

.

	PC_1	PC_2	PC_3	$\ldots$	PC_p
n					

var	λ_1	λ_2	λ_3	$\ldots$	λ_p

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$

$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_i^2$$

Principal Components - Fundamentals

	X_1	X_2	X_3	$\dots$	X_p
n					

$$\Sigma = \begin{bmatrix} 1 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & 1 & \sigma_{2,3} & \cdots & \sigma_{2,p} \\ \sigma_{3,1} & \sigma_{3,2} & 1 & \cdots & \sigma_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p-1} & 1 \end{bmatrix}$$

data scaled

Principal Components - Fundamentals

	PC_1	PC_2	PC_3	$\dots$	PC_p
n					

var	λ_1	λ_2	λ_3	$\dots$	λ_p

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$

$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_i^2$$

data scaled

Principal Components - Fundamentals

	PC_1	PC_2	PC_3	$\ldots$	PC_p
n					
var	λ_1	λ_2	λ_3	$\ldots$	λ_p

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$

$$\sum_{i=1}^{p} \lambda_i = p$$

data scaled

Principal Components - Fundamentals

```
> m1 = eigen(var(d1))
> names(m1)
[1] "values"  "vectors"
> str(m1)
List of 2
 $ values : num [1:4] 7011.11 201.99 42.11 6.16
 $ vectors: num [1:4, 1:4] -0.0417 -0.9952 -0.0463 -0.0752 0.0448 ...
 - attr(*, "class")= chr "eigen"
```

Principal Components - Fundamentals

```
> m1=prcomp(d1)
> names(m1)
[1] "sdev"      "rotation" "center"   "scale"    "x"
```

```
# "sdev": square-root of eigenvalues
```

```
# "rotation": matrix of eigenvectors
# "center"   "scale": # mean and sd of original data -unscaled-
# "x": transformed data set
```

Principal Components - Fundamentals

.

```
> m1=prcomp(d1)
> names(m1)
[1] "sdev"      "rotation" "center"   "scale"    "x"
```

principal
components

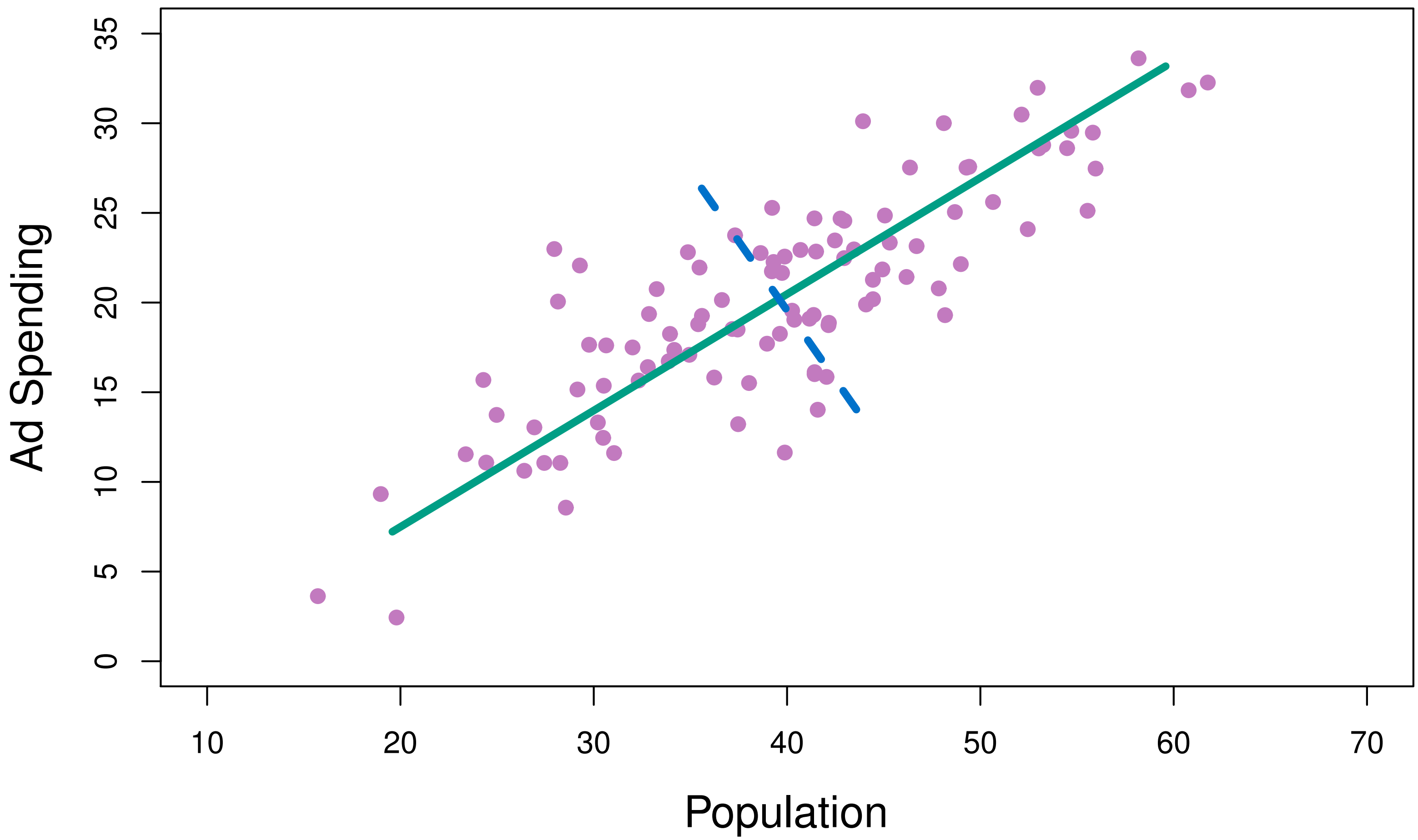
```
# "sdev": square-root of eigenvalues
```

```
# "rotation": matrix of eigenvectors
# "center"   "scale": # mean and sd of original data -unscaled-
# "x": transformed data set
```

Finding Principal Components

- Find 1st PC, having the largest variance
- Find 2nd PC, having the largest variance while orthogonal to the 1st PC
- Find 3rd PC, having the largest variance while orthogonal to the 1st and 2nd PCs

Finding Principal Components

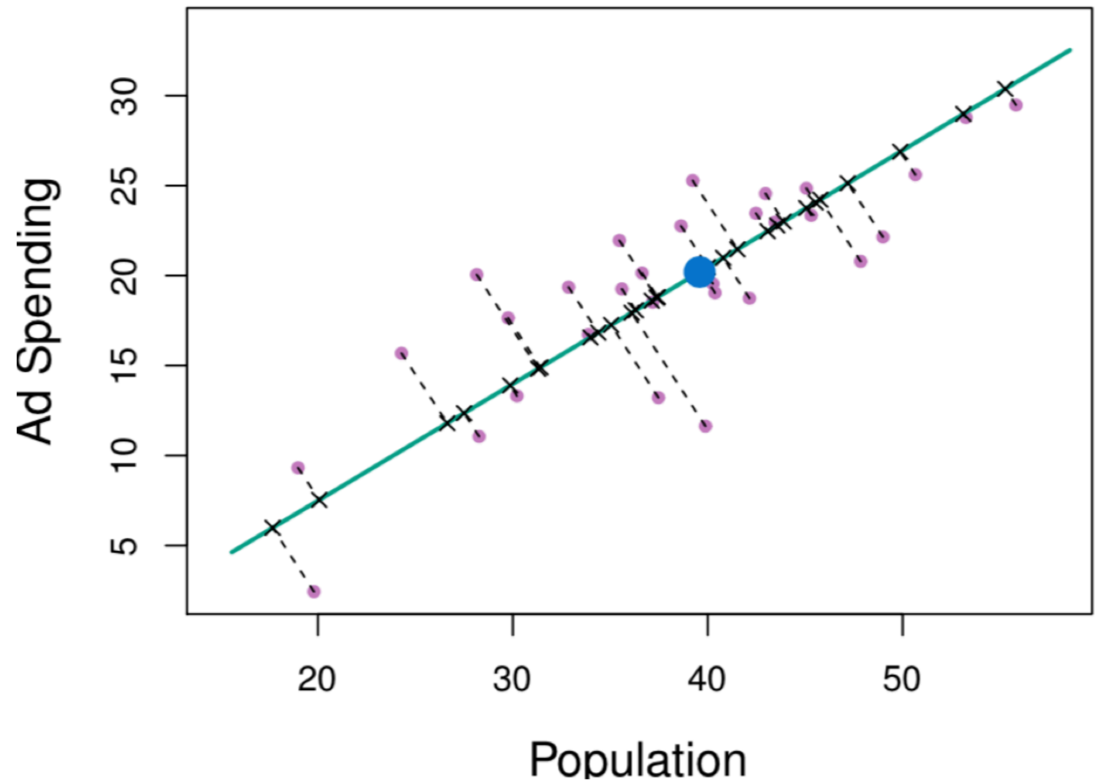


Orthogonal least squares vs OLS

- Find the straight line closest to a set of points
- How to measure the distance from the line to the points?
- OLS uses vertical distances

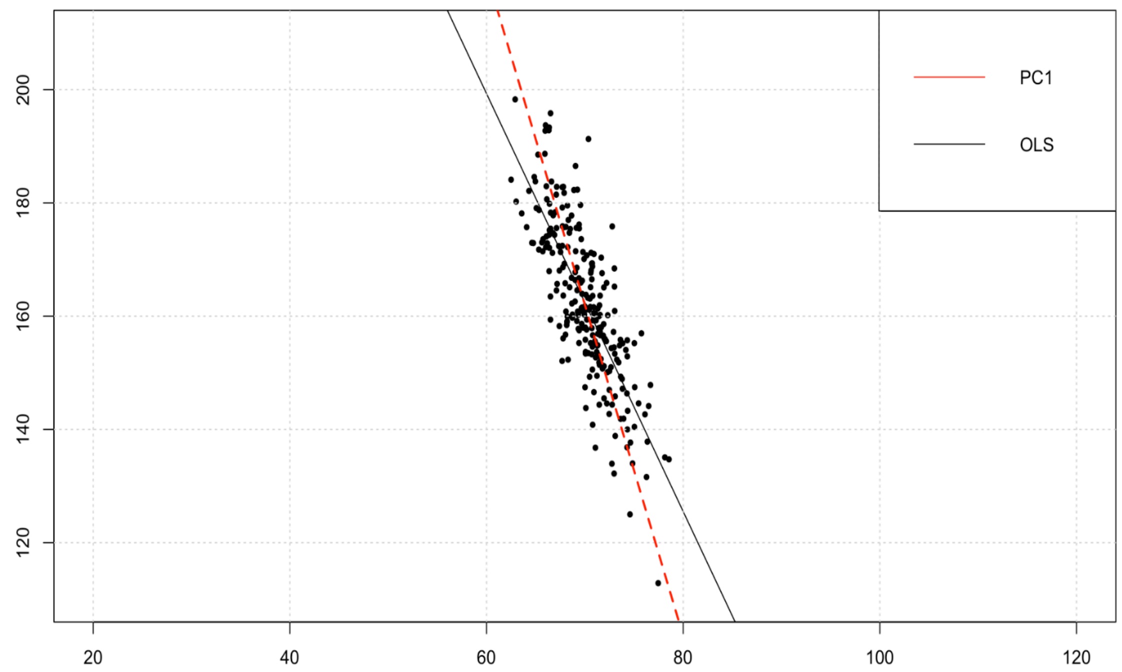
Orthogonal least squares vs OLS

- First PC is the solution to the Orthogonal least squares line
- It is the closest line to the data



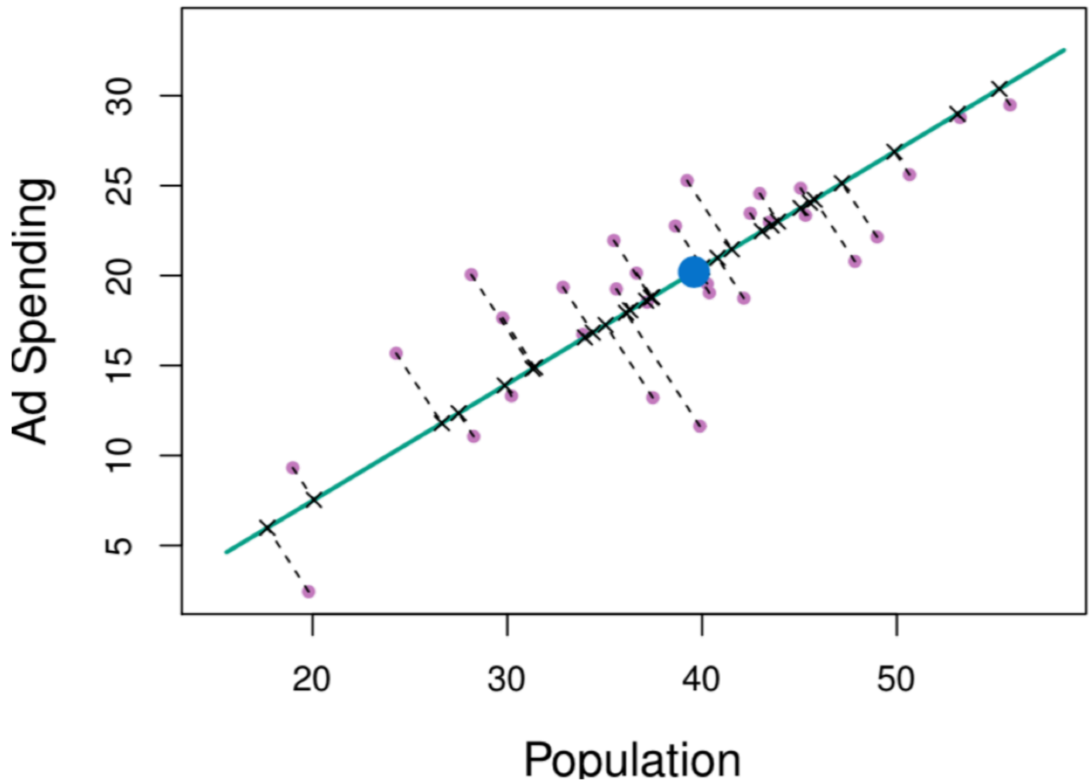
Orthogonal least squares vs OLS

- First PC is the solution to the Orthogonal least squares line
- It is the closest line to the data



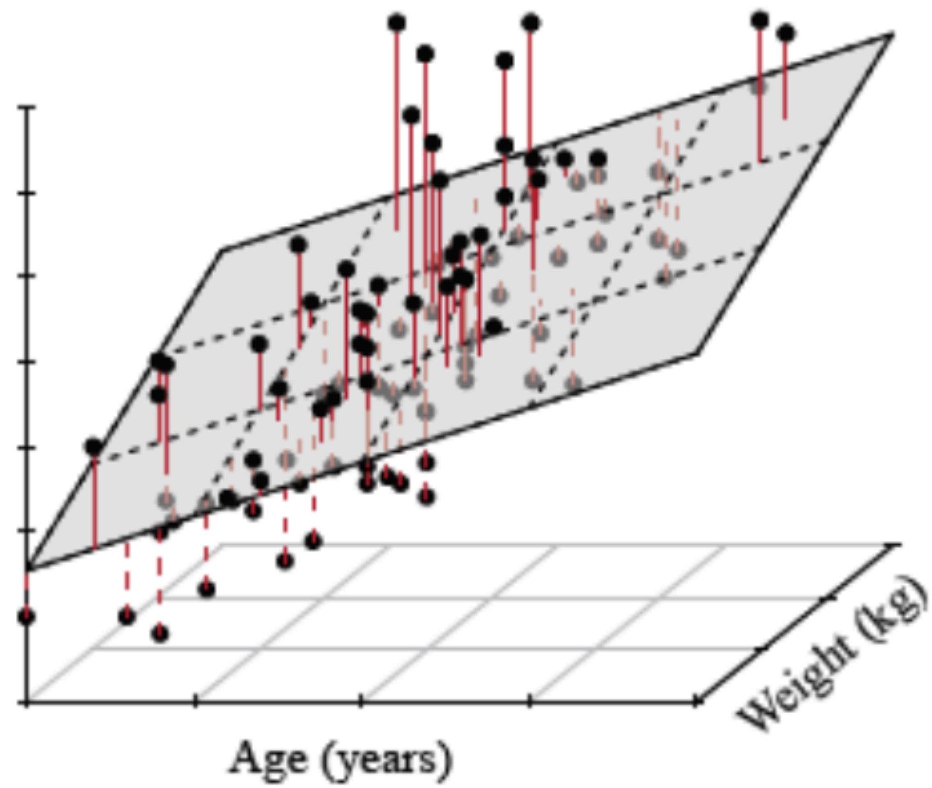
Orthogonal least squares vs OLS

- First PC is the solution to the Orthogonal least squares line
- The closest line is given by the 1st eigenvector of the covariance matrix



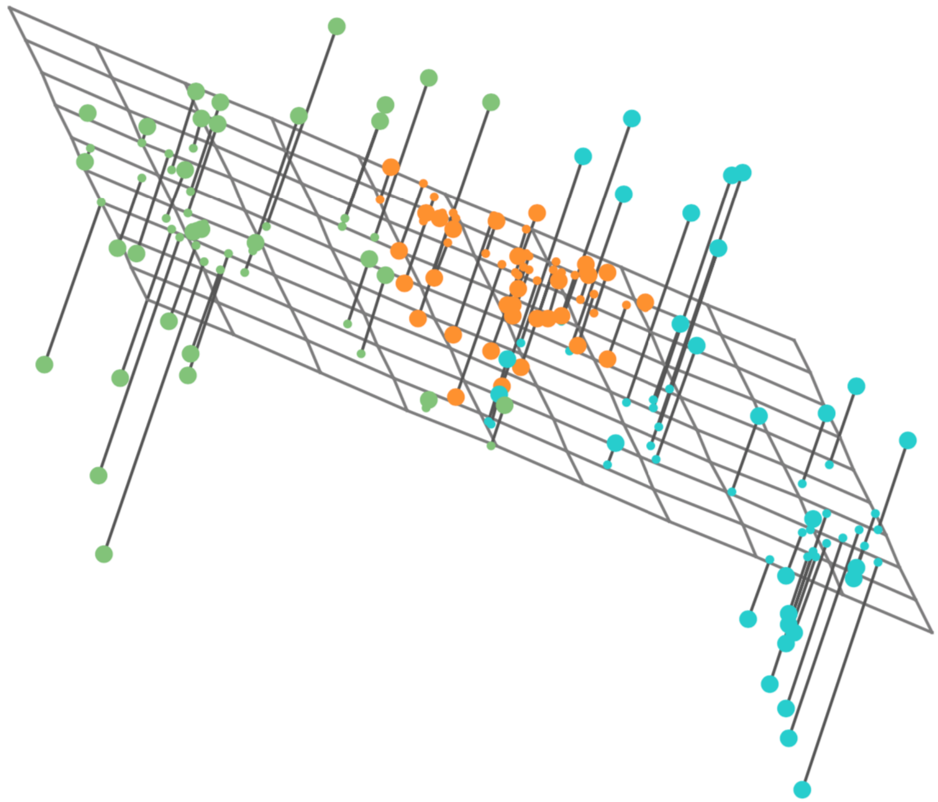
Ordinary Least Squares (OLS) plane

- Closest plane to the data points
- Closest measured by vertical distances



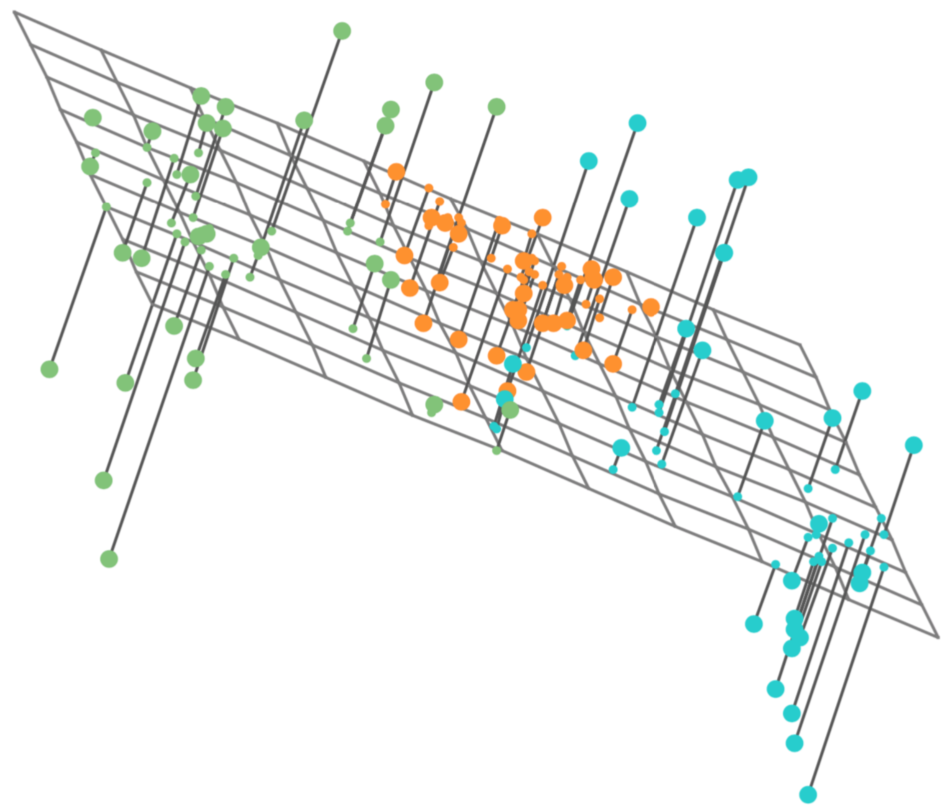
Orthogonal least squares vs OLS

- First and second PCs are the solution to the Orthogonal least squares plane
- It is the closest plane to the data



Orthogonal least squares vs OLS

- First and second PCs are the solution to the Orthogonal least squares plane
- The closest plane is given by the 1st and 2nd eigenvectors of the covariance matrix



Finding Principal Components

- Find 1st PC, having the largest variance
- Find 2nd PC, having the largest variance while orthogonal to the 1st PC
- Find 3rd PC, having the largest variance while orthogonal to the 1st and 2nd PCs

Finding Principal Components

$$\text{for each eigenvector } \vec{e}_i \qquad e_{i1}^2 + e_{i2}^2 + \cdots + e_{ip}^2 = 1$$

$$
\begin{aligned}
X &= [X_1, X_2, \ldots, X_p] \\
PC_1 &= \vec{e}_1 X \\
PC_2 &= \vec{e}_2 X \\
&\vdots \\
PC_p &= \vec{e}_p X
\end{aligned}
$$

Finding Principal Components

.

$$\text{for each eigenvector } \overrightarrow{e}_i \qquad e_{i1}^2 + e_{i2}^2 + \cdots + e_{ip}^2 = 1$$

$$X = [X_1, X_2, \ldots, X_p]$$

$$\begin{aligned} PC_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ PC_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ PC_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

Finding Principal Components

.

$$\text{for each eigenvector } \overrightarrow{e}_i \qquad e_{i1}^2 + e_{i2}^2 + \cdots + e_{ip}^2 = 1$$

$$X = [X_1, X_2, \ldots, X_p]$$

for i^{th} row of the data set

$$\begin{aligned} PC_{1i} &= e_{11}X_{1i} + e_{12}X_{2i} + \cdots + e_{1p}X_{pi} \\ PC_{2i} &= e_{21}X_{1i} + e_{22}X_{2i} + \cdots + e_{2p}X_{pi} \\ &\vdots \\ PC_{pi} &= e_{p1}X_{1i} + e_{p2}X_{2i} + \cdots + e_{pp}X_{pi} \end{aligned}$$

Finding 1st Principal Component

- Center the data, then
- look for a linear combination of the predictors values that has the largest variance

Finding 1st Principal Component

To find the first PC solve

$$\max_{e_{11},\ldots,e_{1p}} \quad \frac{1}{n}\sum_{i=1}^{n} PC_{1i}^2 \qquad \text{st.} \qquad \sum_{i=1}^{p} e_{1j}^2 = 1$$

Finding 1st Principal Component

To find the first PC solve

$$\max_{e_{11},\ldots,e_{1p}} \quad \frac{1}{n}\sum_{i=1}^{n} PC_{1i}^2 \qquad \text{st.} \qquad \sum_{i=1}^{p} e_{1j}^2 = 1$$

$$\max_{e_{11},\ldots,e_{1p}} \quad \frac{1}{n}\sum_{i=1}^{n}(e_{11}X_{1i} + e_{12}X_{2i} + \cdots + e_{1p}X_{pi})^2$$

$$\text{st.} \qquad \sum_{i=1}^{p} e_{1j}^2 = 1$$

Finding Principal Components

- Data X must be centered
- centering does not change data variance
- A data vector X is
    - centered if $\overline{X} = 0$
    - scaled if Var(X) = 1

Finding Principal Components

- Data X must be centered
- centering does not change data variance
- A data vector X is
  - centered if $\sum_{i=1}^{n} X = 0$
  - scaled if Var(X) = 1

Finding Principal Components

- Data X must be centered
- must be *scaled* if variables in different scales

```
##              Murder Assault UrbanPop Rape
## Alabama        13.2     236       58 21.2
## Alaska         10.0     263       48 44.5
## Arizona         8.1     294       80 31.0
## Arkansas        8.8     190       50 19.5
## California      9.0     276       91 40.6
## Colorado        7.9     204       78 38.7
```

Principal Components - Applications

- Data Visualization
- Principal Components Regression
- Clustering
- Multicollinearity prevention
- Outliers identification

Principal Components – Data Visualization

Example: Principal Components Classification

- Predict tumor outcome (benign or malign)
  of patients based on tissue measurements
- Collect lab data of tissue measurements
  related to cancer tumors
- Use PCs to build a classification model to
  predict if a patient has a benign or
  would develop a malign tumor

Principal Components – Data Visualization

	<-------	-----------	-----------	-------	average	values	-----------	-----------	-----------	----------->	<-------	--------	-----------	--------	worst	values	-----------	-----------	-----------
out	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry	fractal_dir	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364
M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985
M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366
M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948
M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792
M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176
M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809
M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596
M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218
M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029
M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	20.96	31.48	136.8	1315	0.1789	0.4233	0.4784	0.2073	0.3706
M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	27.32	30.88	186.8	2398	0.1512	0.315	0.5372	0.2388	0.2768
B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	15.11	19.26	99.7	711.2	0.144	0.1773	0.239	0.1288	0.2977
B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	14.5	20.49	96.09	630.5	0.1312	0.2776	0.189	0.07283	0.3184
B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	10.23	15.66	65.13	314.9	0.1324	0.1148	0.08867	0.06227	0.245
M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	18.07	19.08	125.1	980.9	0.139	0.5954	0.6305	0.2393	0.4667
M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	29.17	35.59	188	2615	0.1401	0.26	0.3155	0.2009	0.2822

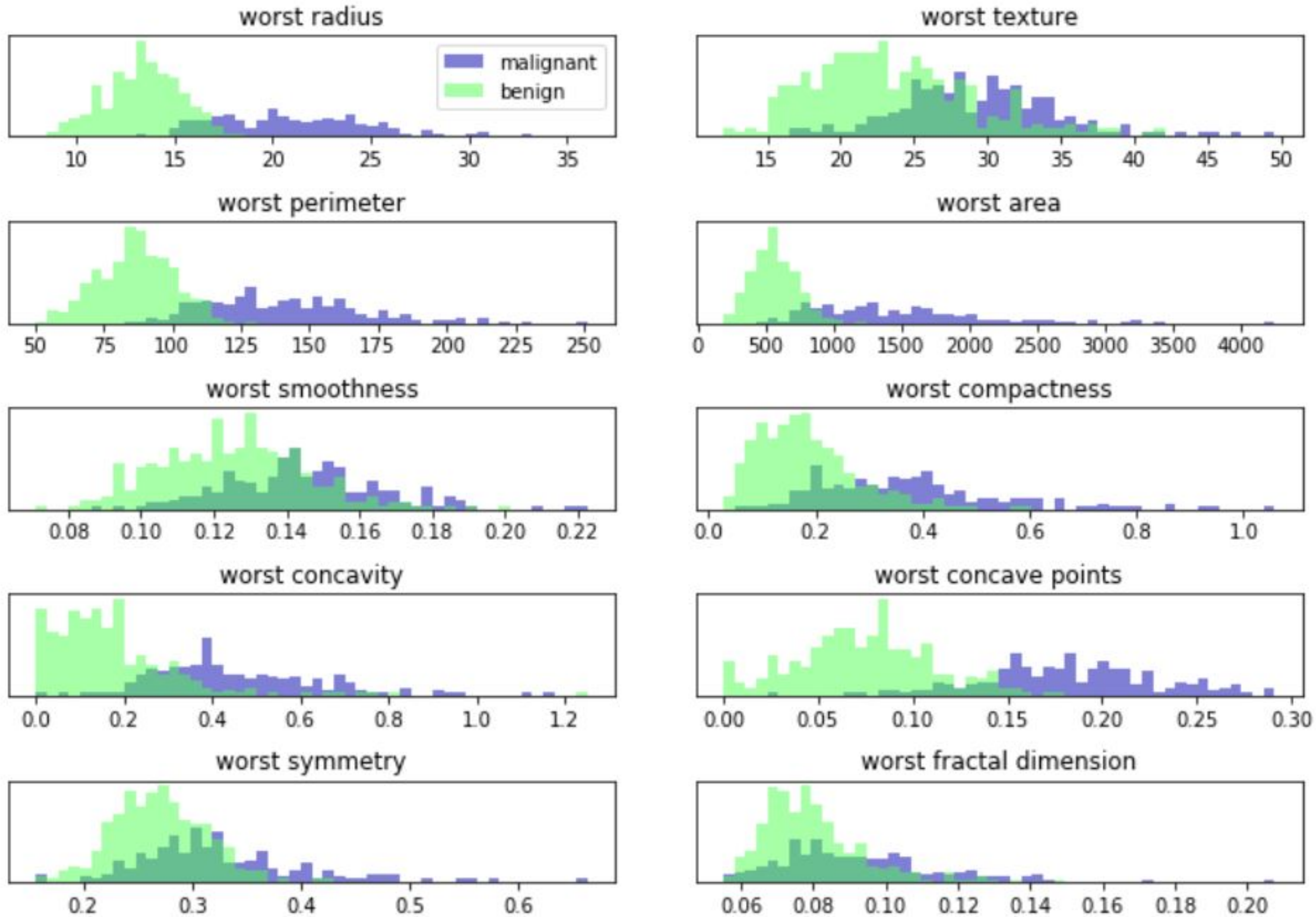
Principal Components – Data Visualization

	<-------	-----------	-----------	-------	average	values	-----------	-----------	-----------	----------->	<-------	--------	-----------	--------	worst	values	-----------	-----------	-----------
out	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry	fractal_dir	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364
M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985
M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366
M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948
M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792
M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176
M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809
M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596
M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218
M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029
M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	20.96	31.48	136.8	1315	0.1789	0.4233	0.4784	0.2073	0.3706
M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	27.32	30.88	186.8	2398	0.1512	0.315	0.5372	0.2388	0.2768
B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	15.11	19.26	99.7	711.2	0.144	0.1773	0.239	0.1288	0.2977
B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	14.5	20.49	96.09	630.5	0.1312	0.2776	0.189	0.07283	0.3184
B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	10.23	15.66	65.13	314.9	0.1324	0.1148	0.08867	0.06227	0.245
M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	18.07	19.08	125.1	980.9	0.139	0.5954	0.6305	0.2393	0.4667
M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	29.17	35.59	188	2615	0.1401	0.26	0.3155	0.2009	0.2822

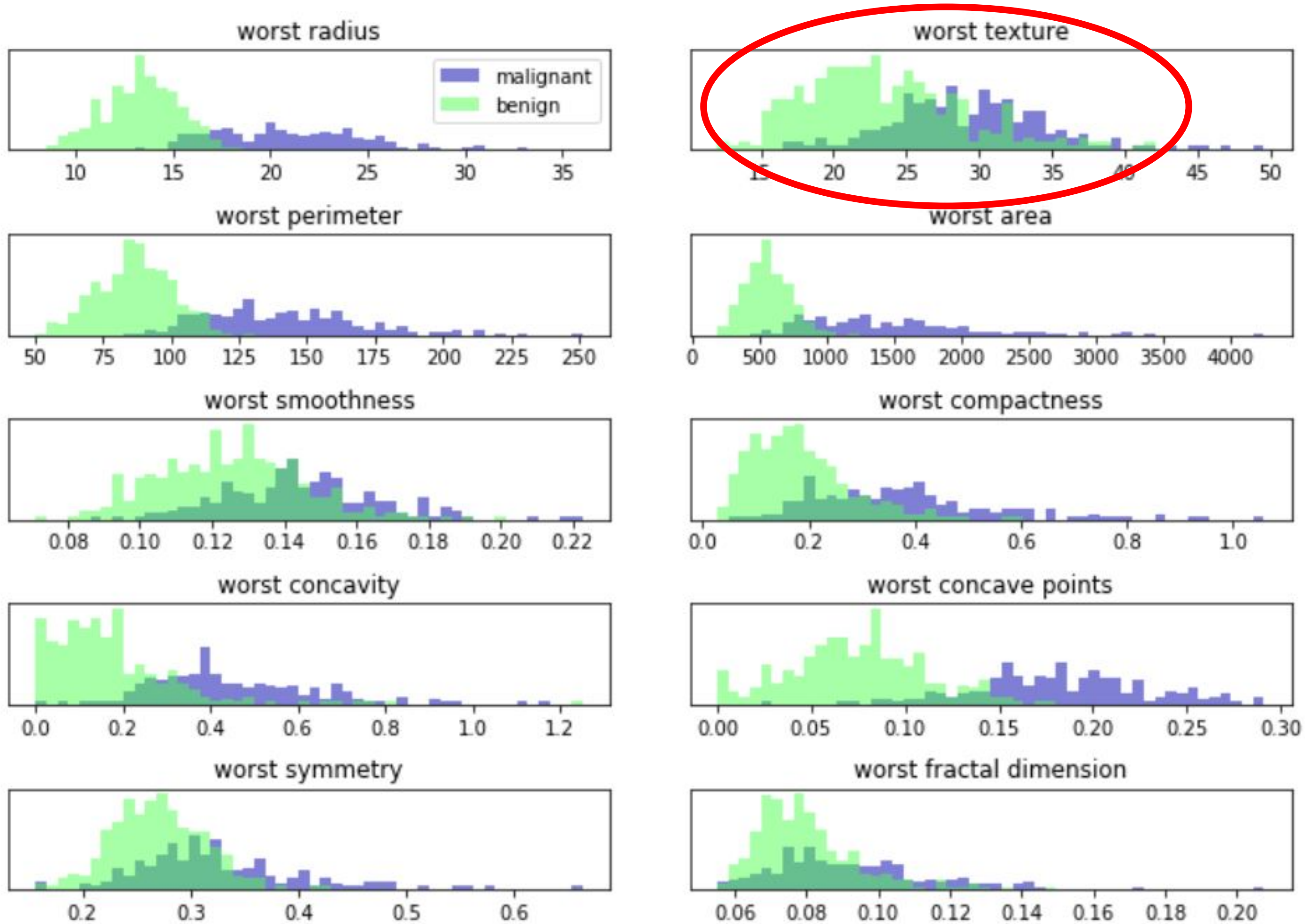
Principal Components – Data Visualization

	<-------	-----------	------------	-------	average	values	------------	------------	------------	----------->	<-------	--------	------------	--------	worst	values	------------	------------	------------
out	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry	fractal_di	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364
M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985
M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366
M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948
M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792
M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176
M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809
M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596
M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218
M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029
M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	20.96	31.48	136.8	1315	0.1789	0.4233	0.4784	0.2073	0.3706
M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	27.32	30.88	186.8	2398	0.1512	0.315	0.5372	0.2388	0.2768
B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	15.11	19.26	99.7	711.2	0.144	0.1773	0.239	0.1288	0.2977
B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	14.5	20.49	96.09	630.5	0.1312	0.2776	0.189	0.07283	0.3184
B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	10.23	15.66	65.13	314.9	0.1324	0.1148	0.08867	0.06227	0.245
M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	18.07	19.08	125.1	980.9	0.139	0.5954	0.6305	0.2393	0.4667
M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	29.17	35.59	188	2615	0.1401	0.26	0.3155	0.2009	0.2822

Principal Components – Data Visualization

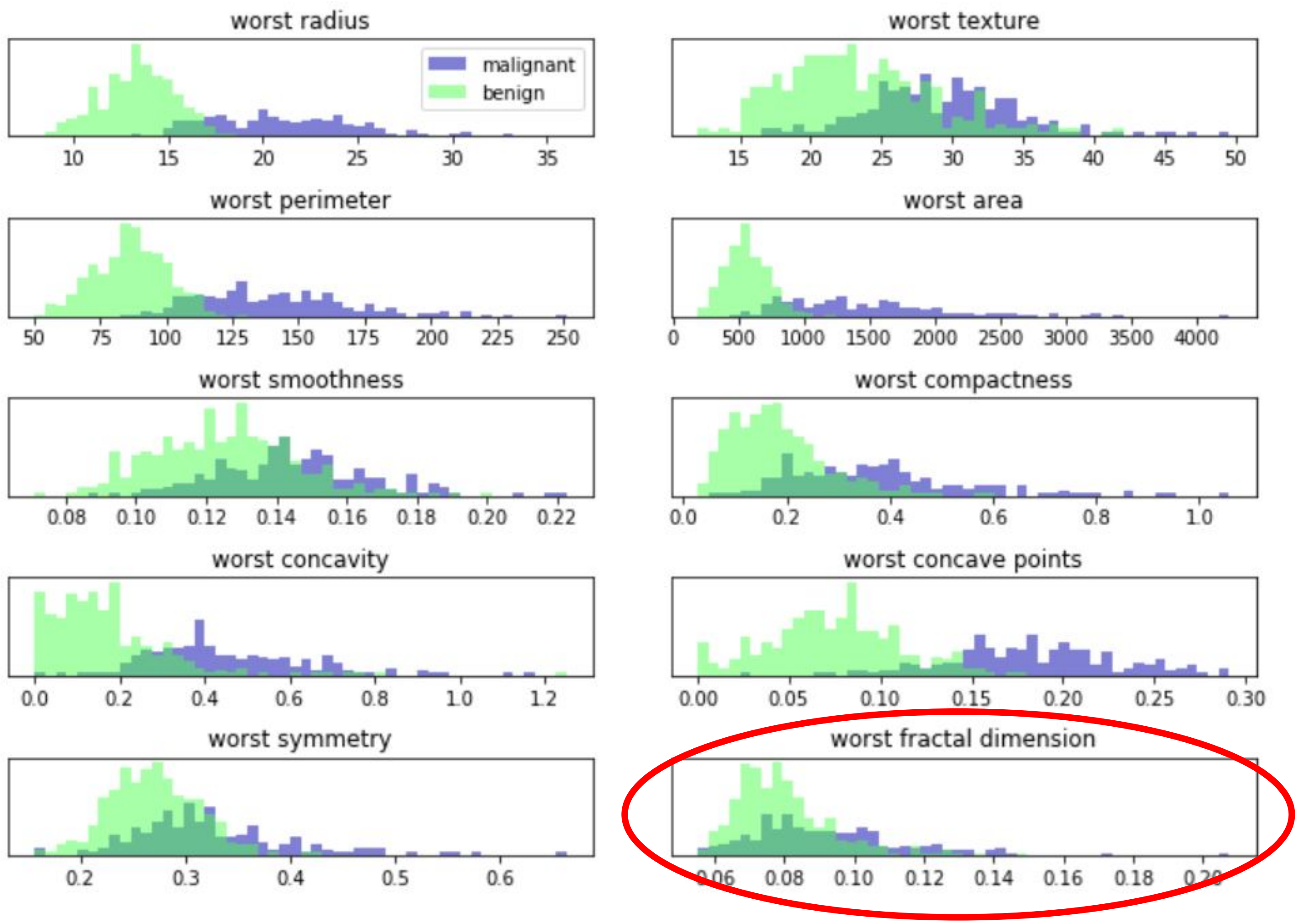


Principal Components – Data Visualization



not good
classifier

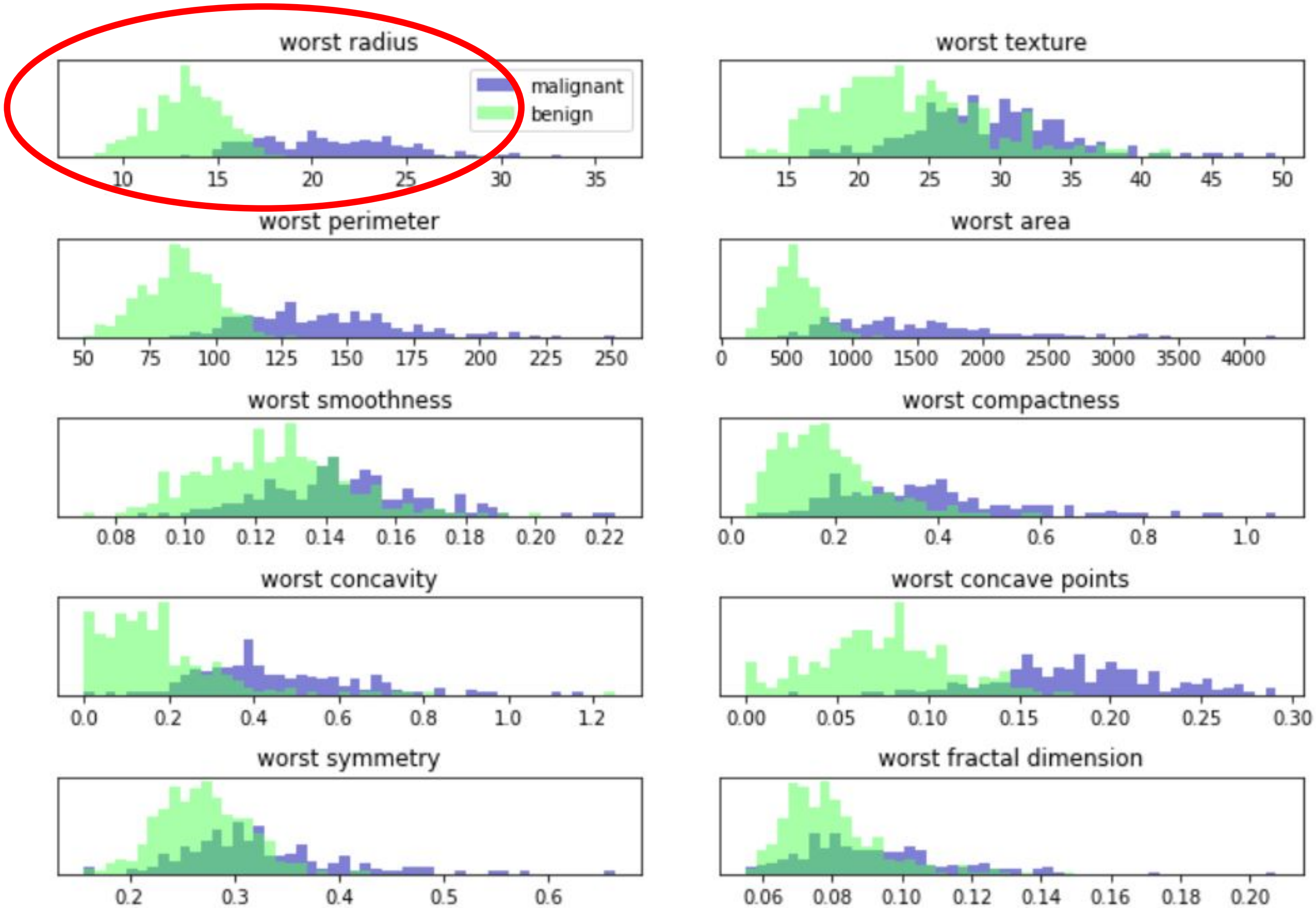
Principal Components – Data Visualization



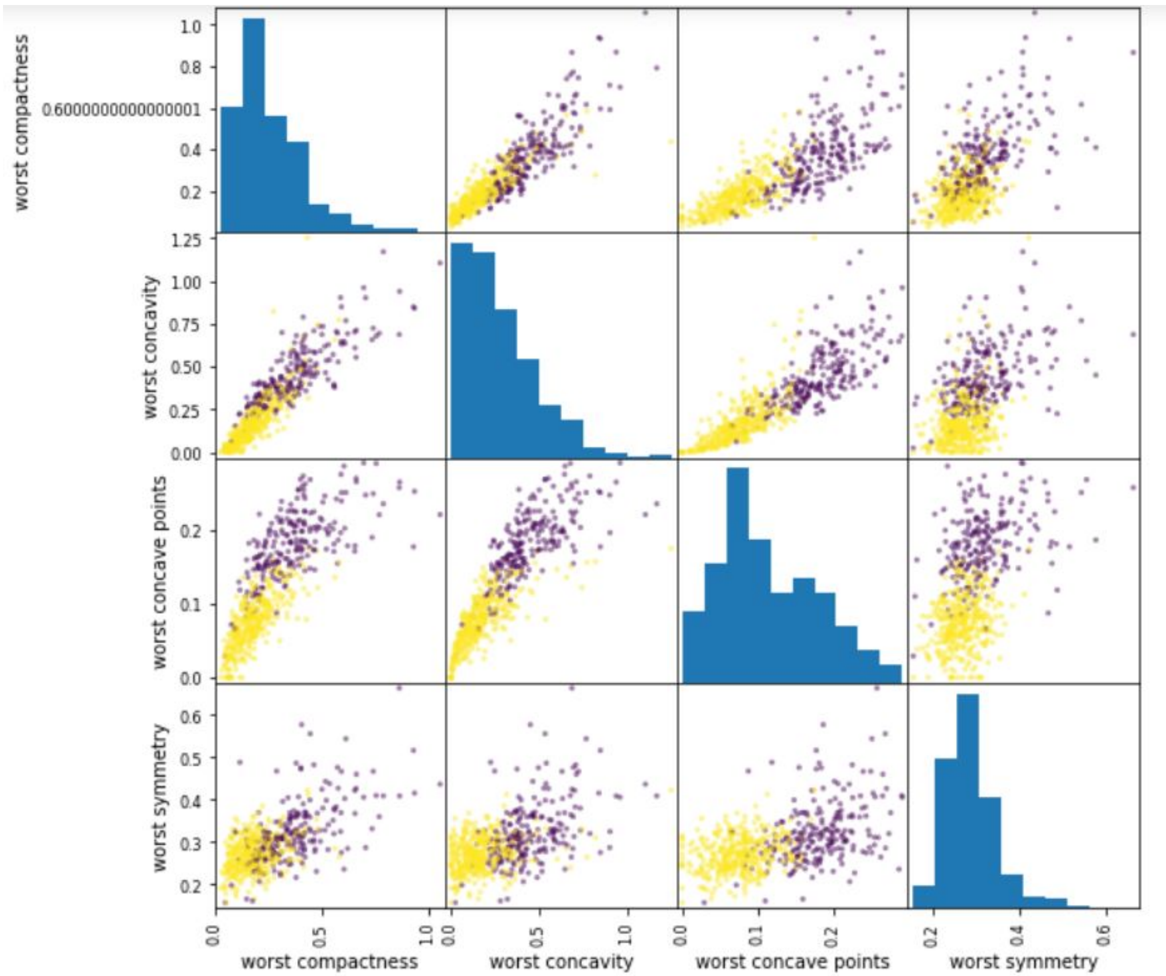
not good classifier

Principal Components – Data Visualization

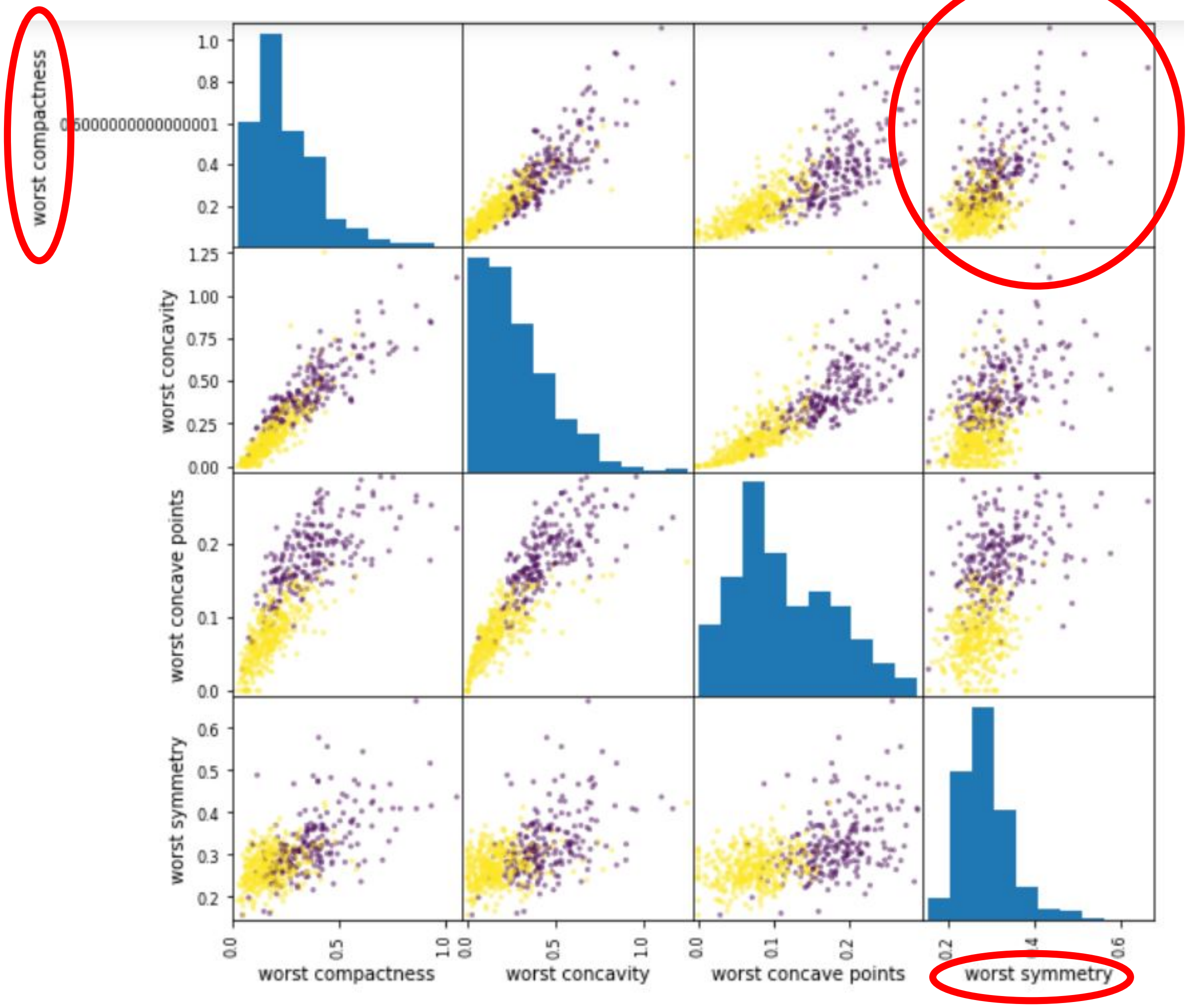
good classifier



Principal Components – Data Visualization

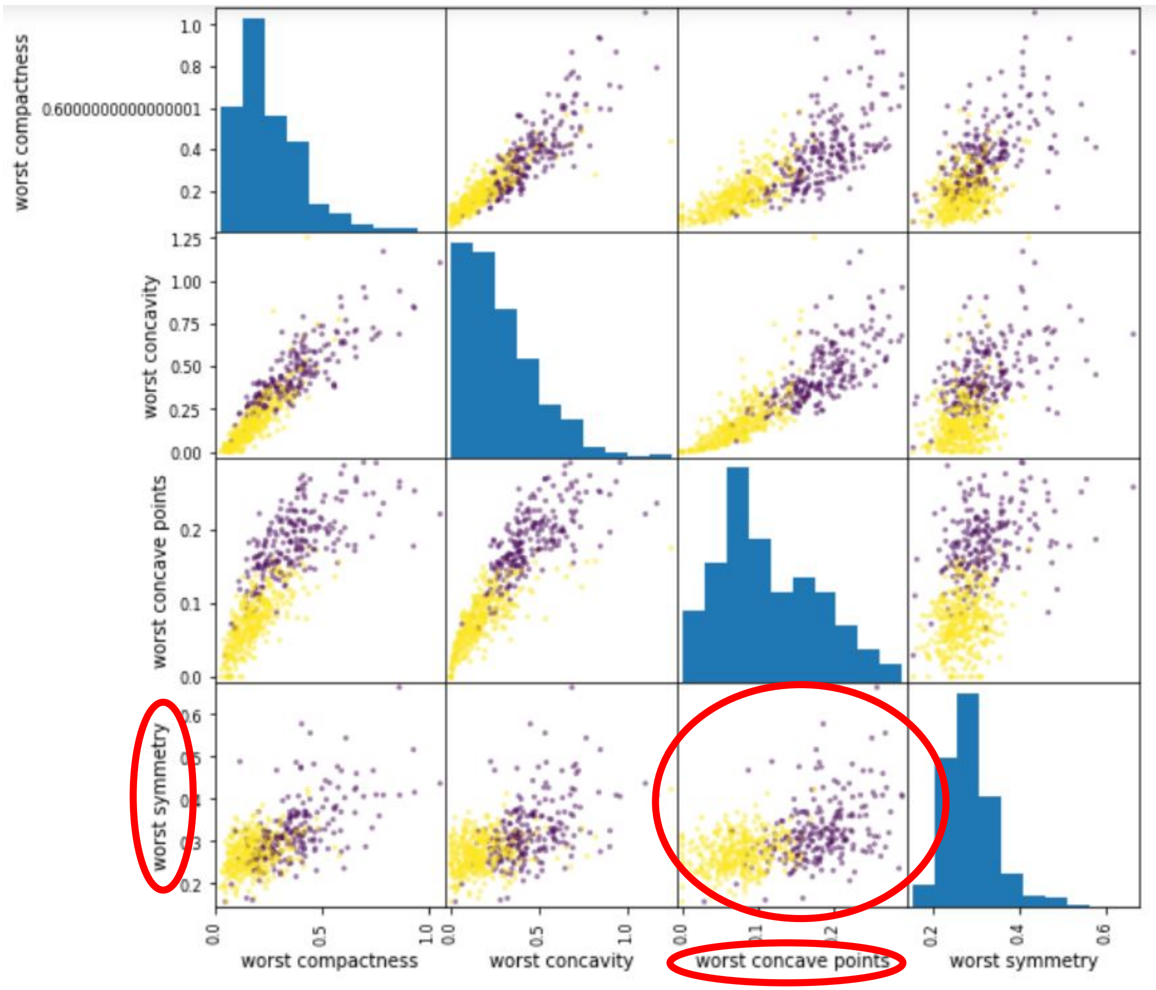


Principal Components – Data Visualization



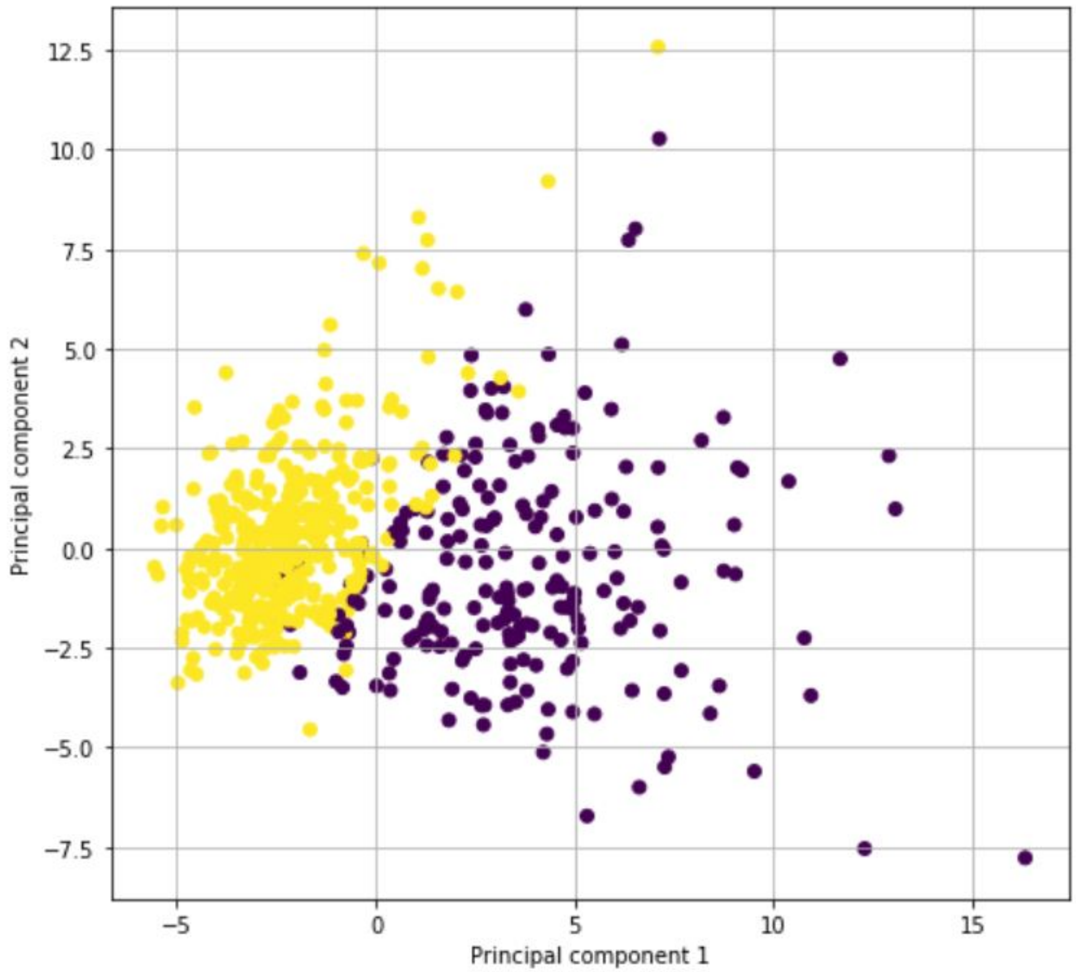
not good classifiers

Principal Components – Data Visualization

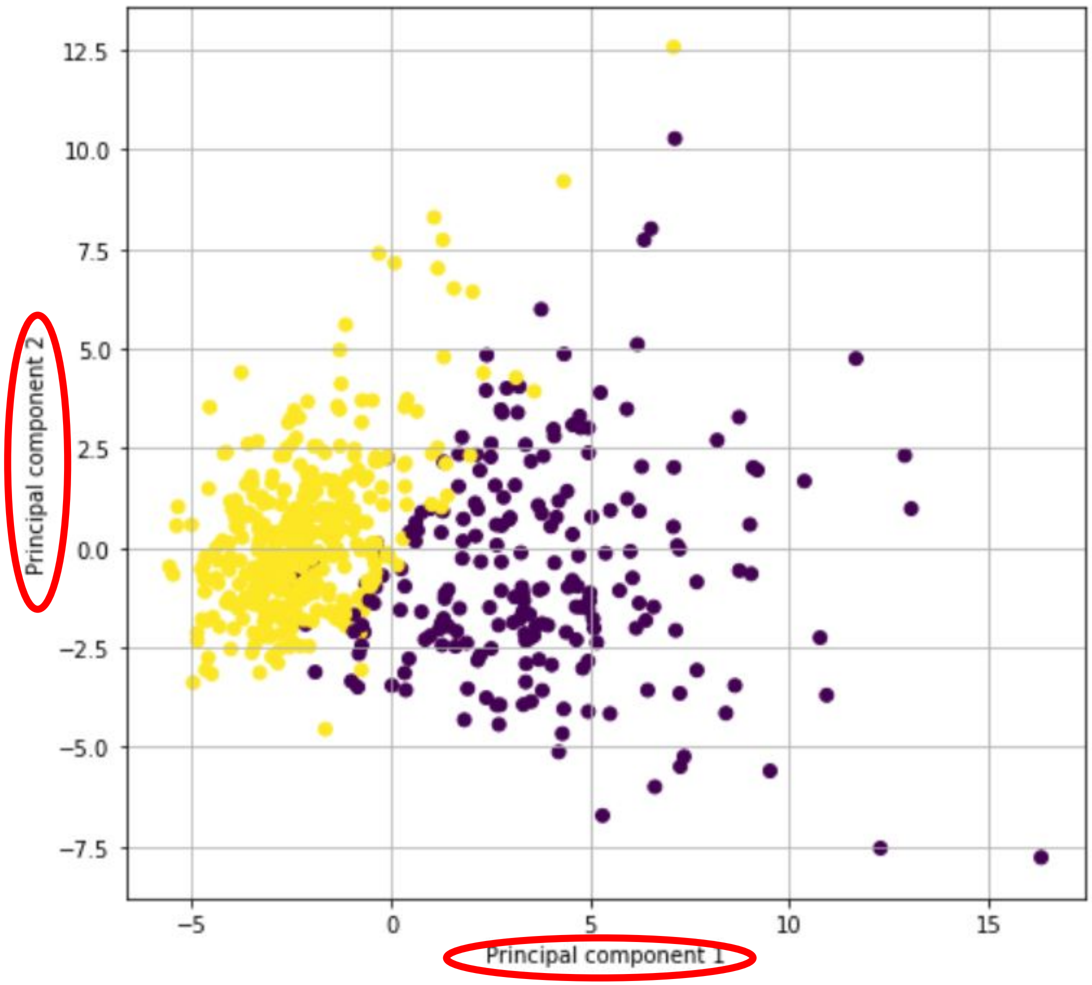


good
classifiers

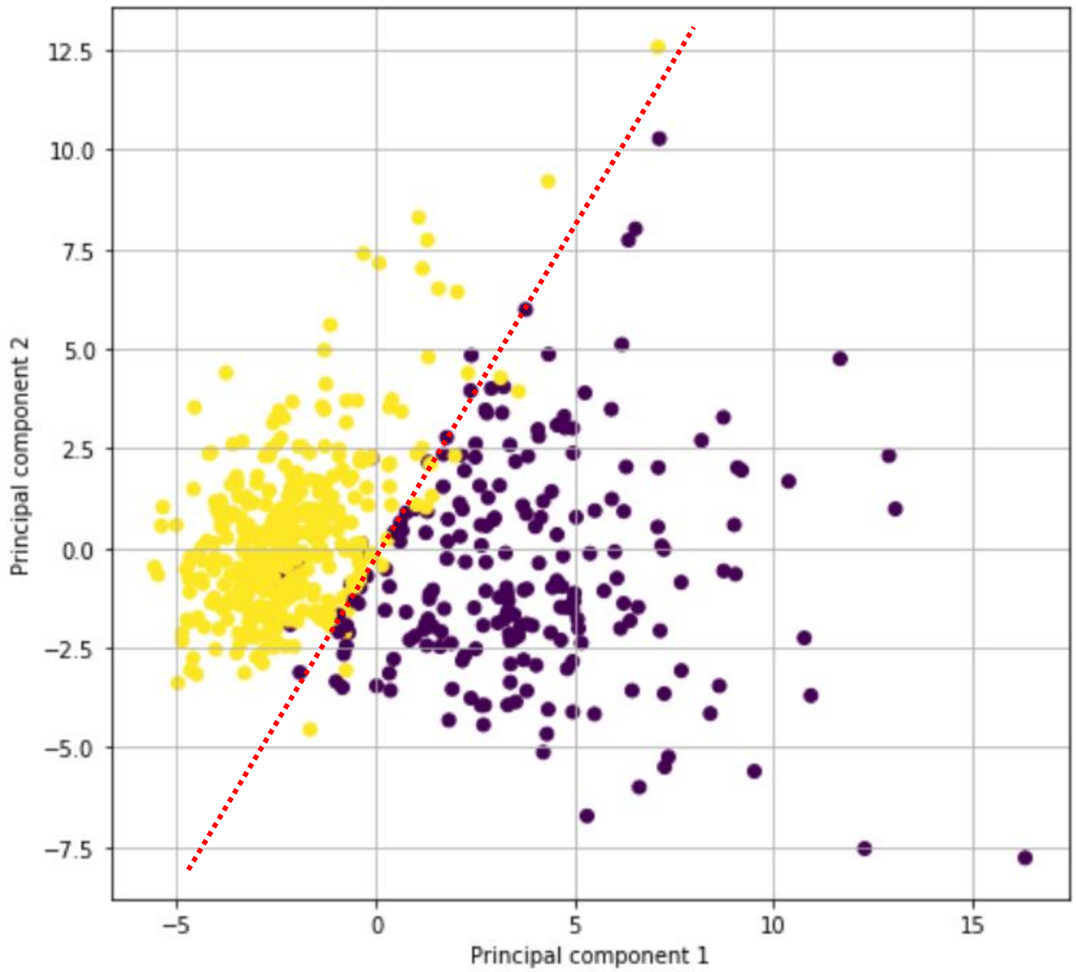
Principal Components – Data Visualization



Principal Components – Data Visualization



Principal Components – Data Visualization



Principal Components Regression

- PC Analysis can be used to substitute correlated variables with a set of uncorrelated ones
- The uncorrelated variables can be used as predictors in a regression model

Principal Components Regression

- PC Analysis can be used to substitute correlated variables with a subset of uncorrelated ones
- The uncorrelated variables can be used as predictors in a regression model

Principal Components Regression

- The uncorrelated variables can then be used as predictors in a regression model
- The resulting model is called Principal Components regression (PCR)
- The number of PCs in the PCR model is chosen by cross-validation

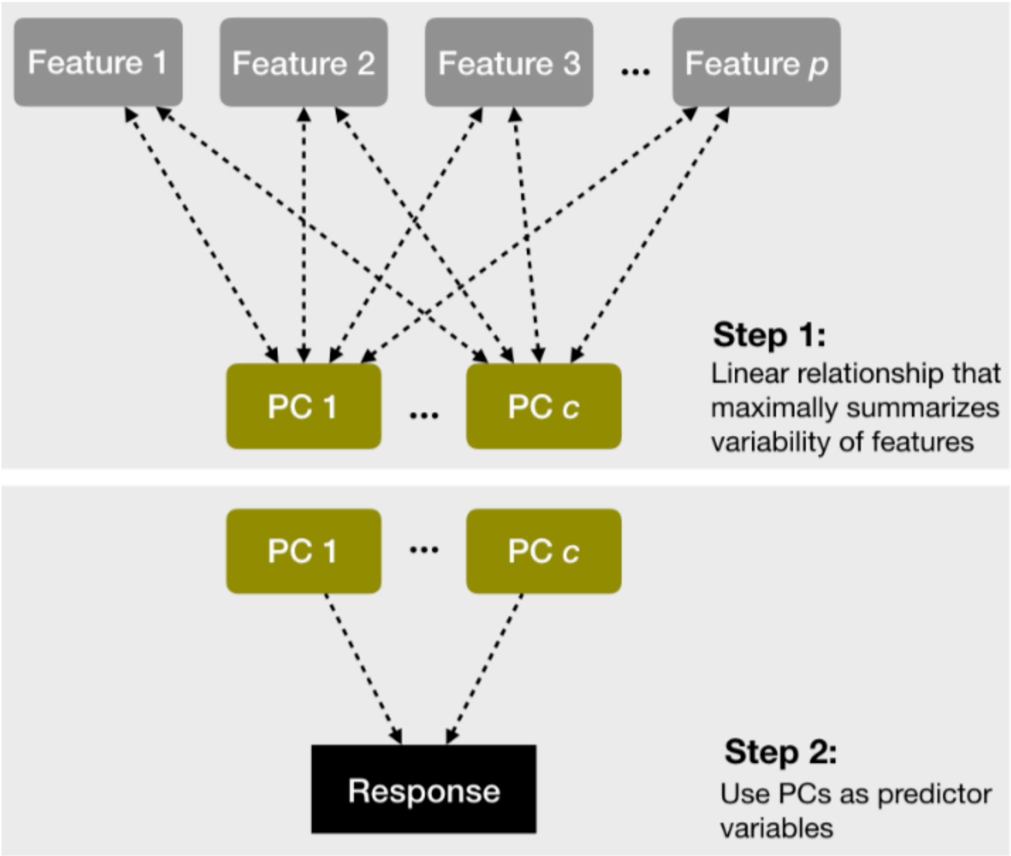
Partial least squares Regression

- PCR does not guarantee that the selected PCs are associated with y
- The model may be better than linear regression but still poor
- An alternative is PLS

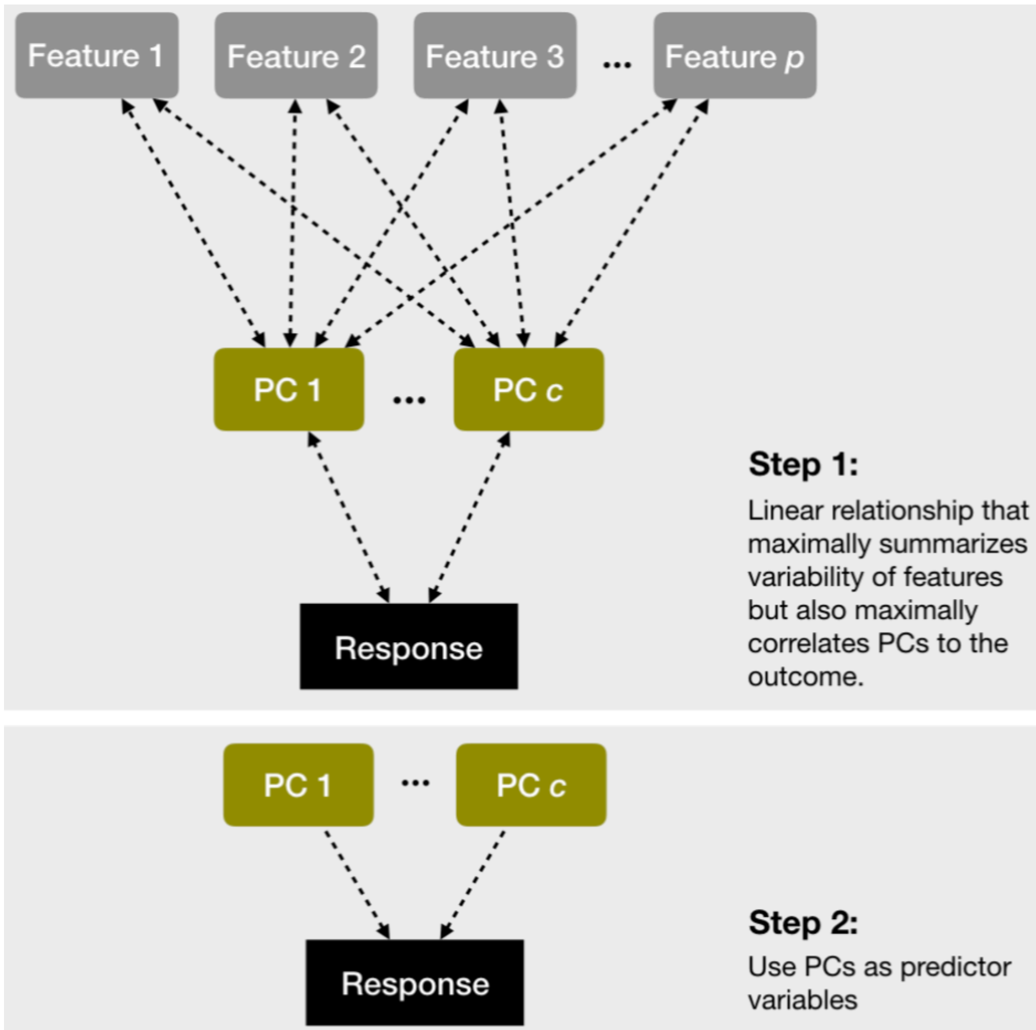
Partial least squares Regression

- PLS identifies components that are linear combination of the original predictors but also are related to y

Principal Components Regression



Partial least squares Regression



Finding 1st PLS Component

$$
\begin{aligned}
X &= [X_1, X_2, \ldots, X_p] \\
Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Z_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Z_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

Finding 1st PLS Component

.

$$
\begin{aligned}
X &= [X_1, X_2, \ldots, X_p] \\
Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p
\end{aligned}
$$

$$
\begin{aligned}
Y &= a_{11}X_1 \\
Y &= a_{12}X_2 \\
&\vdots \\
Y &= a_{1p}X_p
\end{aligned}
$$

Finding 2nd PLS Component

.

$$
\begin{aligned}
X &= [X_1, X_2, \ldots, X_p] \\
Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Z_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p
\end{aligned}
$$

For 2nd component Z_2

regress $X_1, X_2, X_3, \ldots, X_p$ on to Z_1
Use the residuals to fit the model

$$Z_2 = a_{21}r_1 + a_{22}r_2 + \cdots + a_{2p}r_p$$