Data Fabrication

For the purpose of explaining the math behind PCA and how to refer back to the original coefficients a linear model will generate "artificial ideal" data

Model: $z = x + y + 3$

Note:

z : Response variable

x, y : Predictors or features

Coefficients : $a_x = 1$ $a_y = 1$

Intercept $= 3$

ORIGINAL DATA

Data points	x	y	z
P_0	0	0	3
P_1	1	1	5
P_2	2	2	7

Scaling the data

Each predictor will be scaled by $X_s = \frac{X - M_x}{\sigma_x}$

$\sigma = \left[\frac{1}{N}\sum (X - M_x)^2\right]^{1/2}$

M_x and $M_y = 1$

σ_x and $\sigma_y = \left[\frac{1}{3}(1+0+1)\right]^{1/2}$

$= \sqrt{\frac{2}{3}}$

$X_s = \frac{X - 1}{\sqrt{\frac{2}{3}}}$

Scaled Data

Data Points	Xs	Ys	z
P_{0s}	$-\sqrt{\frac{3}{2}}$	$-\sqrt{\frac{3}{2}}$	3
P_{1s}	0	0	5
P_{2s}	$\sqrt{\frac{3}{2}}$	$\sqrt{\frac{3}{2}}$	7

$P_{0s} = \frac{-1}{\sqrt{\frac{2}{3}}} = -\sqrt{\frac{3}{2}}$

$P_{01} = 0$

$P_{02} = \sqrt{\frac{3}{2}}$

Algebraically transforming to the scaled model

$z = x + y + 3$

1. subtract the means

$z = (X - M_x) + (Y - M_y) + (3 + M_x + M_y)$

2. divide by σ_x and σ_y

$$\frac{z}{\sigma_x \sigma_y} = \underbrace{\frac{(X - M_x)}{\sigma_x \sigma_y}}_{x \text{ axis shift}} + \underbrace{\frac{(Y - M_y)}{\sigma_x \sigma_x}}_{y \text{ axis shift}} + \underbrace{\frac{(3 + M_x + M_y)}{\sigma_x \sigma_y}}_{z \text{ axis shift}}$$

3. Multiply by σ_x and σ_y

$z = \sigma_x X_s + \sigma_y Y_s + (3 + M_x + M_y)$

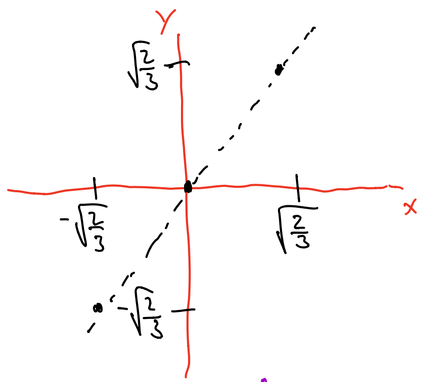
scaled coefficients

Intercept after

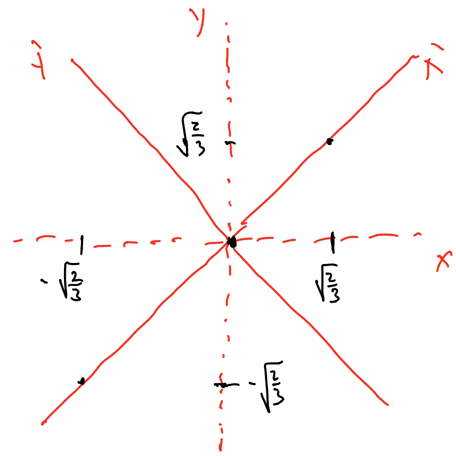
Scaled coefficients ... scaling

Apply PCA

- PCA will transform the data. Note that PCA excludes the response variable z.
- Visualizing the scaled data and rotating it



Apply PCA so that the least # of predictors accounts for most of the variation



* Remember we have artificially ideal data

From the graph it can be observed that all the variation in the data is accounted for by x' (which would be PC1). Now we need to get the data in terms of x', y'

Scaled + Rotated

Data points	x'	y'	z
P_0'	$-\sqrt{3}$	0	3
P_1'	0	0	5
P_2'	$\sqrt{3}$	0	7

We can use the rotation matrix for two dimensions to calculate x' and y' (search rotation matrix in wikipedia)

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_s \\ y_s \end{bmatrix} \Rightarrow \begin{aligned} x' &= x_s\cos\theta - y_s\sin\theta \\ y' &= x_s\sin\theta + y_s\cos\theta \end{aligned}$$

In the figures the axis are rotated by $45°$ which is equivalent to rotating the data points by $-45°$. The θ in the equation refers to the rotation of the data points defined by (x,y).

Note: $\cos(\theta<0) = \cos\theta$ Even function $\qquad \cos(-45) = \frac{\sqrt{2}}{2}$

$\sin(\theta<0) = -\sin\theta$ odd function $\qquad \sin(-45) = -\frac{\sqrt{2}}{2}$

Formulas used to obtain values

$$x' = \frac{\sqrt{2}}{2}x_s + \frac{\sqrt{2}}{2}y_s$$

$$y' = -\frac{\sqrt{2}}{2}x_s + \frac{\sqrt{2}}{2}y_s$$

Solve for x and y in terms of x' and y' $\Rightarrow$

$$x_s = \frac{\sqrt{2}}{2}x' - \frac{\sqrt{2}}{2}y'$$

$$y_s = \frac{\sqrt{2}}{2}y' + \frac{\sqrt{2}}{2}x'$$

$$z = \sigma_x\left[\frac{\sqrt{2}}{2}x' - \frac{\sqrt{2}}{2}y'\right] + \sigma_y\left[\frac{\sqrt{2}}{2}y' + \frac{\sqrt{2}}{2}x'\right] + (3 + m_x + m_y)$$

$z = \sigma_x \frac{\sqrt{2}}{2} x' + \frac{\sigma_x\sqrt{2}}{2} x' + \cancel{\sigma_y \frac{\sqrt{2}}{2} y'} - \cancel{\sigma_y \frac{\sqrt{2}}{2} y'} + 3 + \mu_x + \mu_y$

Plugging in σ_x and $\sigma_y = \sqrt{\frac{2}{3}}$, μ_x and $\mu_y = 1$

$z = \frac{2}{\sqrt{3}} x' + 5$

* plugging in for x' for each data point and solving for z we get the same values as expected since rotating along the z axis (2 dimension xy) does not affect the z value

* From the data points we can also build a line by using the point slope formula since $y' = 0$. If y' wasn't 0 then the 3 data points can be use to find the equation for the plane $ax + by + cz = d$. by using the three points.

Perform linear regression on the data

Since we are using ideal artificial data we would obtain the same line equation as shown above

$z = \frac{2}{\sqrt{3}} x' + 5$

Refer back to the original model

1. Rotate back to the scaled data

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{matrix} x'\cos\theta - y'\sin\theta \\ x'\sin\theta + y'\cos\theta \end{matrix}$$

Solve for x' in terms of x and y $\quad \theta = 45°$

$x' = \frac{\sqrt{2}}{2} x_s + \frac{\sqrt{2}}{2} y_s$

$$z = \frac{2}{\sqrt{3}} \frac{\sqrt{2}}{2} x_s + \frac{2}{\sqrt{3}} \frac{\sqrt{2}}{2} y_s + 5$$

Do reverse operation as for scaling data $x_s = \frac{x - \mu}{\sigma}$

multiply by σ and add μ

$$x = x_s \sigma + \mu$$

$$z = \frac{\sqrt{\frac{2}{3}}}{\sigma_x} (x_s \sigma_x + \mu_x) + \frac{\sqrt{\frac{2}{3}}}{\sigma_y} (y_s \sigma_x + \mu_y) \sigma_y + 5 - \frac{\sqrt{\frac{2}{3}}\mu_x}{\sigma_x} - \frac{\sqrt{\frac{2}{3}}\mu_y}{\sigma_y}$$

$$= \frac{\sqrt{\frac{2}{3}}}{\sigma_x} (x) + \frac{\sqrt{\frac{2}{3}}}{\sigma_y} (y) + 5 - \frac{\sqrt{\frac{2}{3}}\mu_x}{\sigma_x} - \frac{\sqrt{\frac{2}{3}}\mu_y}{\sigma_y}$$

* Note $\sqrt{\frac{2}{3}}$ are the scaled coeficients

Plug in $\sigma_y, \sigma_x = \sqrt{\frac{2}{3}}$ and $\mu_x, \mu_y = 1$

$z = x + y + 3$ back to original model

In a real model we would not have returned to the original model because we would had droped some principal components and our linea regression would not had been perfect.