Name: Chee Kitt Win
Student Number: 45589140
STAT3006 Assignment 3

Note: All python code has been attached

1.

1. a) $\tau_1(x) = P(Y=1 \mid X=x)$

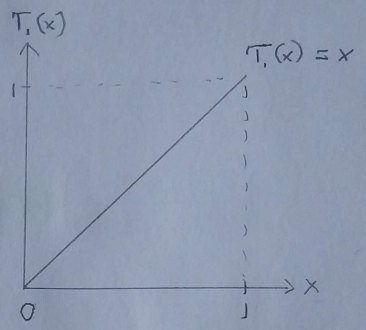
$= \dfrac{f(x \mid Y=1)P(Y=1)}{f(x)}$

$= \dfrac{f(x \mid Y=1)P(Y=1)}{\sum_{c=0}^{1} f(x \mid Y=c)P(Y=c)}$

$= \dfrac{(2x)(\frac{1}{2})}{(2x)\frac{1}{2} + (2-2x)\frac{1}{2}}$

$= \dfrac{x}{x + (1-x)}$

$= x$ //



b) $L(r^*) = \frac{1}{2} - \frac{1}{2}E(|2\tau_1(x) - 1|)$

$= \frac{1}{2} - \frac{1}{2}E(|2x-1|)$

$= \frac{1}{2} - \frac{1}{2}\int_0^1 |2x-1| f(x)\, dx$ → From part a), $f(x) = 1$

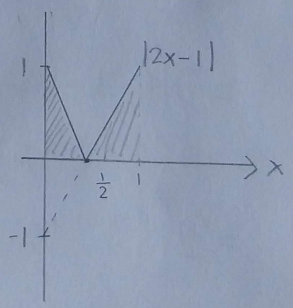
$= \frac{1}{2} - \frac{1}{2}\left(\int_{1/2}^{1}(2x-1)\,dx + \int_0^{1/2} -(2x-1)\,dx\right)$

$= \frac{1}{2} - \frac{1}{2}\left([x^2 - x]_{1/2}^{1} + [-x^2 + x]_0^{1/2}\right)$

$= \frac{1}{2} - \frac{1}{2}\left((0 - (-\frac{1}{4})) + (\frac{1}{4} - 0)\right)$

$= \frac{1}{2} - \frac{1}{2}\left(\frac{1}{4} + \frac{1}{4}\right)$

$= \frac{1}{4}$ //



c) $L(r^*) = \frac{1}{2} - \frac{1}{2}E(|2\pi_1(x) - 1|)$

$\pi_1(x) = P(y=1|X=x)$

$E(|2\pi_1(x) - 1|)$

$= E\left(\left|\frac{2x\pi_1 - 1 + x + \pi_1 - 2\pi_1 x}{1 - x - \pi_1 + 2\pi_1 x}\right|\right)$

$= E\left(\left|\frac{-1 + x + \pi_1}{1 - x - \pi_1 + 2\pi_1 x}\right|\right)$

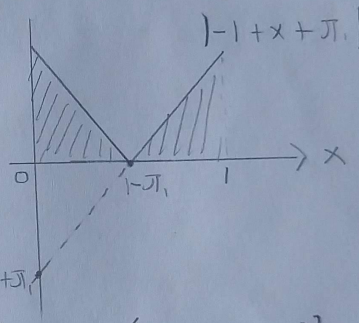
$= \int_0^1 \left|\frac{-1 + x + \pi_1}{1 - x - \pi_1 + 2\pi_1 x}\right| f(x)\, dx$

$= \int_0^1 \left|\frac{-1 + x + \pi_1}{1 - x - \pi_1 + 2\pi_1 x}\right| (2)\underbrace{(1 - x - \pi_1 + 2\pi_1 x)}_{\text{Note this is } \geq 0}\, dx$

$= 2\int_0^1 |-1 + x + \pi_1|\, dx$

$= 2\left(\int_0^{1-\pi_1} -(-1 + x + \pi_1)\, dx + \int_{1-\pi_1}^1 -1 + x + \pi_1\, dx\right)$

$= 2\left(\left[-x + \frac{1}{2}x^2 + \pi_1 x\right]_0^{1-\pi_1} + \left[-x + \frac{1}{2}x^2 + \pi_1 x\right]_{1-\pi_1}^1\right)$



~~$= 2\left(\left(1 - \pi_1 - \frac{1}{2}(1-\pi_1)^2 - \pi_1(1-\pi_1)\right)\right.$~~

~~$\left. + \left(-1 + \frac{1}{2} + \pi_1\right) - (-($~~

$= 2\left[-\left(\pi_1 - 1 + \frac{1}{2}(1-\pi_1)^2 + \pi_1(1-\pi_1)\right) + \left(-1 + \frac{1}{2} + \pi_1 - \left(\pi_1 - 1 + \frac{1}{2}(1-\pi_1)^2 + \pi_1(1-\pi_1)\right)\right)\right]$

$= 2\left(-2\left(\pi_1 - 1 + \frac{1}{2} - \pi_1 + \frac{1}{2}\pi_1^2 + \pi_1 - \pi_1^2\right) + \left(-\frac{1}{2} + \pi_1\right)\right)$

$= -4\left(-\frac{1}{2} - \frac{1}{2}\pi_1^2 + \pi_1\right) - 1 + 2\pi_1$

$= 2 + 2\pi_1^2 - 4\pi_1 - 1 + 2\pi_1$

$= 1 + 2\pi_1^2 - 2\pi_1$

$L(r^*) = \frac{1}{2} - \frac{1}{2}(1 + 2\pi_1^2 - 2\pi_1)$

$= \frac{1}{2} - \frac{1}{2} - \pi_1^2 + \pi_1$

$= \underline{\pi_1 - \pi_1^2}$ #

d) From part b), we have

$$L(r^*) = E\left(\min\{\eta_1(x), 1-\eta_1(x)\}\right)$$

$$= E\left(\min\{P(Y=1|X=x), P(Y=0|X=x)\}\right)$$

$$= E\left(\min\left\{\frac{f(x|Y=1)\pi_1}{f(x)}, \frac{f(x|Y=0)(1-\pi_1)}{f(x)}\right\}\right)$$

$$= \int_{\mathcal{X}} \min\left\{\frac{f(x|Y=1)\pi_1}{f(x)} f(x), \frac{f(x|Y=0)(1-\pi_1)}{f(x)} f(x)\right\} dx$$

$$= \int_{\mathcal{X}} \min\left\{f(x|Y=1)\pi_1, f(x|Y=0)(1-\pi_1)\right\} dx$$

Also from part b), we have

$$L(r^*) = \frac{1}{2} - \frac{1}{2} E\left(|2\eta_1(X)-1|\right)$$

$$= \frac{1}{2} - \frac{1}{2}\int_0^1 |2P(Y=1|X=x)-1| f(x)\, dx$$

$$= \frac{1}{2} - \frac{1}{2}\int_0^1 |P(Y=1|X=x) + (1-P(Y=0|X=x)) - 1| f(x)\, dx$$

$$= \frac{1}{2} - \frac{1}{2}\int_0^1 |P(Y=1|X=x) - P(Y=0|X=x)| f(x)\, dx$$

$$= \frac{1}{2} - \frac{1}{4}\int_0^1 |2P(Y=1|X=x)f(x) - 2P(Y=0|X=x)f(x)|\, dx$$

$$= \frac{1}{2} - \frac{1}{4}\int_{\mathcal{X}} \left|\frac{2f(x|Y=1)(\frac{1}{2})f(x)}{f(x)} - \frac{2f(x|Y=0)(\frac{1}{2})f(x)}{f(x)}\right| dx$$

$$= \frac{1}{2} - \frac{1}{4}\int_{\mathcal{X}} |f(x|Y=1) - f(x|Y=0)|\, dx$$

e) From part d), we have

$$L(r^*) = \int_x \min\{(1-\pi_1) f(x|y=0),\ \pi_1 f(x|y=1)\}\, dx$$

Case 1: $(1-\pi_1) f(x|y=0) \le \pi_1 f(x|y=1)$

Since both L.H.S and R.H.S ≥ 0,

$\Rightarrow (1-\pi_1) \int_x f(x|y=0)\, dx \le \pi_1 \int_x f(x|y=1)\, dx$

$\Rightarrow 1-\pi_1 \le \pi_1$

$\Rightarrow \pi_1 \ge \frac{1}{2}$

Also, $L(r^*) = \int_x (1-\pi_1) f(x|y=0)\, dx$

$= 1-\pi_1$

$\therefore$ In this case, $\min(L(r^*)) = 0$ when $\pi_1 = 1$

$\max(L(r^*)) = \frac{1}{2}$ when $\pi_1 = \frac{1}{2}$

Case 2: $\pi_1 f(x|y=1) \le (1-\pi_1) f(x|y=0)$

Since both L.H.S and R.H.S ≥ 0,

$\Rightarrow \pi_1 \int_x f(x|y=1)\, dx \le (1-\pi_1) \int_x f(x|y=0)\, dx$

$\Rightarrow \pi_1 \le 1-\pi_1$

$\Rightarrow \pi_1 \le \frac{1}{2}$

Also, $L(r^*) = \int_x \pi_1 f(x|y=1)\, dx$

$= \pi_1$

$\therefore$ In this case, $\min(L(r^*)) = 0$ when $\pi_1 = 0$

$\max(L(r^*)) = \frac{1}{2}$ when $\pi_1 = \frac{1}{2}$

2.

2. a) $\Gamma_y(x;\theta) = P(Y=y \mid X=x)$

$= \dfrac{f(x,y)}{f(x)}$

Given $\begin{cases} f(x_{i1} \mid Y_i = y) = \phi(x_{i1}; 0, \delta_y^2) \quad \text{---} \ (1) \\ f(x_{ir} \mid X_{i1}=x_{i1}, X_{i2}=x_{i2}, \ldots X_{i,r-1}=x_{i,r-1}, Y_i=y) = \phi(x_{ir}; \beta_y x_{i,r-1}, \delta_y^2) \quad \text{---} \ (2) \end{cases}$

$f(x_{i1}, Y_i=y) = f(x_{i1} \mid Y_i=y) P(Y_i=y)$

$= \phi(x_{i1}; 0, \delta_y^2) \pi_y \quad \text{---} \ (3)$

$f(x_i, y) = f(x_{i1}, x_{i2}, \ldots, y)$

$= f(x_{i1}, Y_i=y) f(x_{i2} \mid x_{i1}, Y_i=y) f(x_{i3} \mid x_{i1}, x_{i2}, Y_i=y) \ldots$

$\ldots f(x_{i,100} \mid x_{i1}, \ldots, x_{i,99}, Y_i=y)$

$= \pi_y f(x_{i1}, Y_i=y) \prod_{r=2}^{100} \phi(x_{ir}; \beta_y x_{i,r-1}, \delta_y^2)$

$= \pi_y \phi(x_{i1}; 0, \delta_y^2) \prod_{r=2}^{100} \phi(x_{ir}; \beta_y x_{i,r-1}, \delta_y^2) \quad \text{---} \ (4)$

$= \dfrac{f(x,y)}{\sum_{c=1}^{3} f(x, Y=c)}$

from (4), $= \dfrac{\pi_y \phi(x_1; 0, \delta_y^2) \prod_{r=2}^{100} \phi(x_r; \beta_y x_{r-1}, \delta_y^2)}{\sum_{c=1}^{3} \left(\pi_c \phi(x_1; 0, \delta_c^2) \prod_{r=2}^{100} \phi(x_r; \beta_c x_{r-1}, \delta_c^2) \right)}$ //

Bayes classifier : $r(x) = \underset{y \in \{1,2,3\}}{\arg\max}\ \Gamma_y(x;\theta)$

$= \underset{y \in \{1,2,3\}}{\arg\max} \left(\dfrac{\pi_y \phi(x_1; 0, \delta_y^2) \prod_{r=2}^{100} \phi(x_r; \beta_y x_{r-1}, \delta_y^2)}{\sum_{c=1}^{3} \left(\pi_c \phi(x_1; 0, \delta_c^2) \prod_{r=2}^{100} \phi(x_r; \beta_c x_{r-1}, \delta_c^2) \right)} \right)$ //

(Denominator is a constant, so can be ignored when considering $\arg\max$)

b) $L(\theta; Z_n) = \prod_{i=1}^{n} f(z_i; \theta)$

$= \prod_{i=1}^{n} f(x_i, y_i; \theta)$

from (4) in part a),

$= \prod_{i=1}^{n} \left(\prod_{k=1}^{3} \left[\pi_k \phi(x_{i1}; 0, \delta_k^2) \prod_{r=2}^{100} \phi(x_{ir}; \beta_k x_{i,r-1}, \delta_k^2) \right]^{[\![y_i = k]\!]} \right)$

$= \prod_{i=1}^{n} \left(\prod_{k=1}^{3} \left[\pi_k \left(\frac{1}{\sqrt{2\pi\delta_k^2}} e^{-\frac{1}{2}\left(\frac{x_{i1}}{\delta_k}\right)^2} \right) \prod_{r=2}^{100} \left(\frac{1}{\sqrt{2\pi\delta_k^2}} e^{-\frac{1}{2}\left(\frac{x_{ir} - \beta_k x_{i,r-1}}{\delta_k}\right)^2} \right) \right]^{[\![y_i = k]\!]} \right)$

$= \prod_{i=1}^{n} \left(\prod_{k=1}^{3} \left[\pi_k \left(\frac{1}{\sqrt{2\pi\delta_k^2}} \right)^{100} e^{-\frac{1}{2}\left(\frac{x_{i1}}{\delta_k}\right)^2 - \frac{1}{2\delta_k^2} \sum_{r=2}^{100} (x_{ir} - \beta_k x_{i,r-1})^2} \right]^{[\![y_i = k]\!]} \right)$

$\log L(\theta; Z_n) = \sum_{i=1}^{n} \sum_{k=1}^{3} \left(\log \left[\pi_k^n \left(\frac{1}{\sqrt{2\pi\delta_k^2}} \right)^{100} e^{-\frac{1}{2\delta_k^2}(x_{i1})^2 - \frac{1}{2\delta_k^2} \sum_{r=2}^{100} (x_{ir} - \beta_k x_{i,r-1})^2} \right]^{[\![y_i = k]\!]} \right)$

$= \sum_{i=1}^{n} \sum_{k=1}^{3} \left(\left[\log \pi_k^n + \log \left(\frac{1}{\sqrt{2\pi\delta_k^2}} \right)^{100} + \left(-\frac{1}{2\delta_k^2} \left(x_{i1}^2 + \sum_{r=2}^{100} (x_{ir} - \beta_k x_{i,r-1})^2 \right) \right) \right]^{[\![y_i = k]\!]} \right)$

$= \log \pi_y^n + \log \left(\frac{1}{\sqrt{2\pi\delta_y^2}} \right)^{100n} + \sum_{i=1}^{n} \left(-\frac{1}{2\delta_y^2}(x_{i1})^2 - \frac{1}{2\delta_y^2} \sum_{r=2}^{100} (x_{ir} - \beta_y x_{i,r-1})^2 \right)$

$= \log \pi_y^n + \log \left(\frac{1}{\sqrt{2\pi\delta_y^2}} \right)^{100n} - \frac{1}{2\delta_y^2} \sum_{i=1}^{n} (x_{i1})^2 - \frac{1}{2\delta_y^2} \sum_{i=1}^{n} \sum_{r=2}^{100} (x_{ir} - \beta_y x_{i,r-1})^2$

$= \sum_{i=1}^{n} \sum_{k=1}^{3} \left[\left(\log \pi_k^n + \log \left(\frac{1}{\sqrt{2\pi\delta_k^2}} \right)^{100} - \frac{1}{2\delta_k^2} \left(x_{i1}^2 + \sum_{r=2}^{100} (x_{ir} - \beta_k x_{i,r-1})^2 \right) \right) \left([\![y_i = k]\!] \right) \right]$

Let $\mathcal{L}(\theta; z_n) = \log L(\theta; z_n) - \lambda g(\theta; z_n)$

where $g(\theta; z_n) = \sum_{k=1}^{3} \pi_k - 1$ is the constraint function

Using the method of Lagrange multipliers to maximize $\log L(\theta; z_n)$ with respect to π_k, set $\nabla \mathcal{L}(\theta; z_n) = 0$

$$\Rightarrow \frac{\partial}{\partial \pi_k}\left(\log L(\theta; z_n)\right) = \lambda \frac{\partial}{\partial \pi_k}\left(\sum_{k=1}^{3} \pi_k - 1\right)$$

$$\Rightarrow \sum_{i=1}^{n} \frac{n}{\pi_k} [\![y_i = k]\!] = \lambda (1)$$

$$\Rightarrow n \sum_{i=1}^{n} [\![y_i = k]\!] = \pi_k \lambda \quad \text{---- (1)}$$

$$\Rightarrow n \sum_{k=1}^{3} \sum_{i=1}^{n} [\![y_i = k]\!] = \sum_{k=1}^{3} \pi_k \lambda$$

$$\Rightarrow n^2 = \lambda_k (1)$$

$$\Rightarrow \lambda_k = n^2 \quad \text{---- (2)}$$

(2) → (1) $\quad n \sum_{i=1}^{n} [\![y_i = k]\!] = n^2 \pi_k$

$$\Rightarrow \underline{\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} [\![y_i = k]\!] \quad \text{for } k \in \{1, 2, 3\}}$$

Now setting $\frac{\partial}{\partial \beta_k}\left(\log L(\theta; z_n)\right) = 0$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k}\left[\sum_{k=1}^{3}\left(\left(-\frac{1}{2\sigma_k^2} \sum_{r=2}^{100} (x_{ir} - \beta_k x_{i,r-1})^2\right)[\![y_i = k]\!]\right)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k}\left(-\frac{1}{2\sigma_k^2} \sum_{r=2}^{100} (-2 x_{ir} \beta_k x_{i,r-1} + \beta_k^2 x_{i,r-1}^2)[\![y_i = k]\!]\right) = 0$$

$$\Rightarrow -\frac{1}{2\sigma_k^2} \sum_{i=1}^{n} \sum_{r=2}^{100} (-2 x_{ir} x_{i,r-1} + 2\beta_k x_{i,r-1}^2)[\![y_i = k]\!] = 0$$

$$\Rightarrow -\sum_{i=1}^{n} \sum_{r=2}^{100} x_{ir} x_{i,r-1} [\![y_i = k]\!] + \beta_k \sum_{i=1}^{n} \sum_{r=2}^{100} x_{i,r-1}^2 [\![y_i = k]\!] = 0$$

$$\Rightarrow \hat{\beta}_k = \frac{\sum_{i=1}^{n} \sum_{r=2}^{100} x_{ir} x_{i,r-1} [\![y_i = k]\!]}{\sum_{i=1}^{n} \sum_{r=2}^{100} x_{i,r-1}^2 [\![y_i = k]\!]} \quad \text{for } k \in \{1, 2, 3\}$$

Now setting $\frac{\partial}{\partial \sigma_k^2}\left(\log L(\theta; z_n)\right) = 0$.

$$\Rightarrow 0 = \frac{\partial}{\partial \sigma_k^2} \sum_{i=1}^{n} \sum_{k=1}^{3} \left(100 \log\left(2\pi\sigma_k^2\right)^{-1/2} - \frac{1}{2}\left(\sigma_k^2\right)^{-1}\left(x_{i1}^2 + \sum_{r=2}^{100}\left(x_{ir} - B_k x_{i,r-1}\right)^2\right)\right)[\![y_i = k]\!]$$

$$\Rightarrow 0 = \sum_{i=1}^{n} \left(100\left(2\pi\sigma_k^2\right)^{\frac{1}{2}}\left(-\frac{1}{2}\right)\left(2\pi\sigma_k^2\right)^{-3/2}(2\pi) - \frac{1}{2}(-1)\left(\sigma_k^2\right)^{-2}\left(x_{i1}^2 + \sum_{r=2}^{100}\left(x_{ir} - B_k x_{i,r-1}\right)^2\right)\right)[\![y_i = k]\!]$$

$$\Rightarrow 0 = \sum_{i=1}^{n} \left(-50(2\pi)\left(2\pi\sigma_k^2\right)^{-1} + \frac{1}{2\sigma_k^4}\left(x_{i1}^2 + \sum_{r=2}^{100}\left(x_{ir} - B_k x_{i,r-1}\right)^2\right)\right)[\![y_i = k]\!]$$

$$\Rightarrow 0 = -50 \sum_{i=1}^{n} \frac{1}{\sigma_k^2}[\![y_i = k]\!] + \frac{1}{2\sigma_k^4} \sum_{i=1}^{n} \left(x_{i1}^2 + \sum_{r=2}^{100}\left(x_{ir} - B_k x_{i,r-1}\right)^2\right)[\![y_i = k]\!]$$

$$\Rightarrow 50\sigma_k^2 \sum_{i=1}^{n} [\![y_i = k]\!] = \frac{1}{2} \sum_{i=1}^{n} \left(x_{i1}^2 + \sum_{r=2}^{100}\left(x_{ir} - B_k x_{i,r-1}\right)^2\right)[\![y_i = k]\!]$$

$$\Rightarrow \hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n}\left(x_{i1}^2 + \sum_{r=2}^{100}\left(x_{ir} - B_k x_{i,r-1}\right)^2\right)[\![y_i = k]\!]}{100 \sum_{i=1}^{n} [\![y_i = k]\!]} \quad \text{for } k \in \{1,2,3\}$$

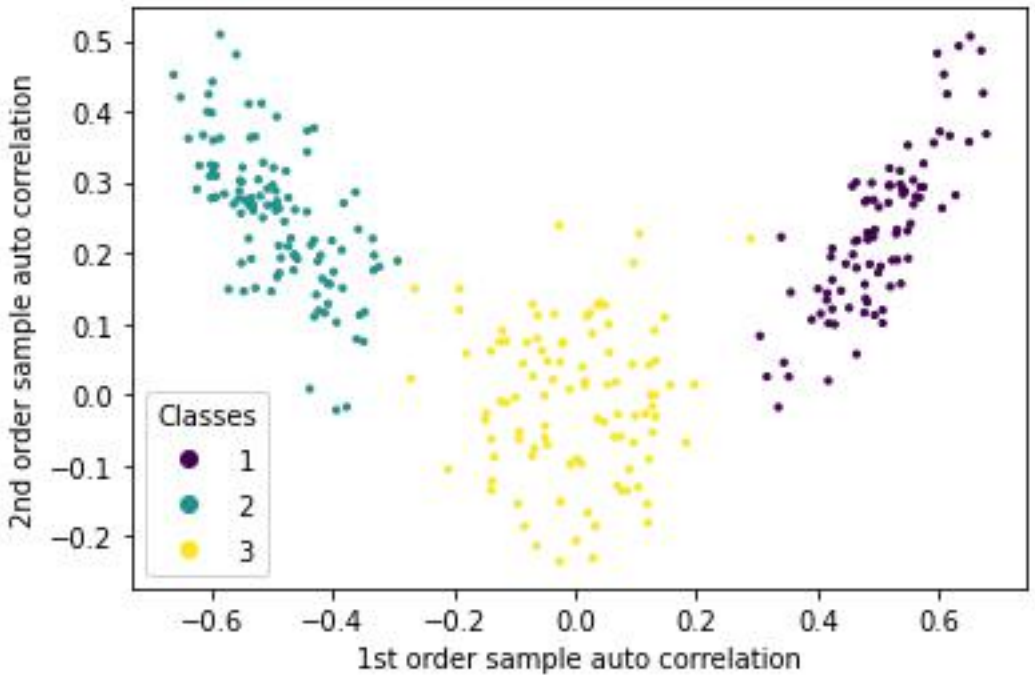
2. d)

I used the normalized form of the sample auto covariance (sample auto correlation). *Figure 1* shows a plot of the 2nd order sample auto correlation vs the 1st order sample auto correlation. Each point represents a time series.

Formula used for sample auto covariance[1]

$$\hat{R}(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} (X_t - \mu)(X_{t+k} - \mu)$$

Figure 1



[1] https://en.wikipedia.org/wiki/Autocorrelation

e) $\Gamma_y(x;\hat{\theta}) = \dfrac{\hat{\pi}_y \frac{1}{\sqrt{2\pi\hat{\sigma}_y^2}} e^{-\frac{1}{2}\left(\frac{x_1}{\hat{\sigma}_y}\right)^2} \prod_{r=2}^{100} \frac{1}{\sqrt{2\pi\hat{\sigma}_y^2}} e^{-\frac{1}{2}\left(\frac{x_r - \hat{B}_y x_{r-1}}{\hat{\sigma}_y}\right)^2}}{\sum_{c=1}^{3}\left(\hat{\pi}_c \frac{1}{\sqrt{2\pi\hat{\sigma}_c^2}} e^{-\frac{1}{2}\left(\frac{x_1}{\hat{\sigma}_c}\right)^2} \prod_{r=2}^{100} \frac{1}{\sqrt{2\pi\hat{\sigma}_c^2}} e^{-\frac{1}{2}\left(\frac{x_r - \hat{B}_c x_{r-1}}{\hat{\sigma}_c}\right)^2}\right)}$

$$= \frac{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_y^2}}\right)^{100} \hat{\pi}_y \, e^{-\frac{1}{2\hat{\sigma}_y^2}\left((x_1)^2 + \sum_{r=2}^{100}(x_r - \hat{B}_y x_{r-1})^2\right)}}{\sum_{c=1}^{3}\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_c^2}}\right)^{100} \hat{\pi}_c \, e^{-\frac{1}{2\hat{\sigma}_c^2}\left((x_1)^2 + \sum_{r=2}^{100}(x_r - \hat{B}_c x_{r-1})^2\right)}}$$

for $y \in \{1,2,3\}$

where $\hat{\theta} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{B}_1, \hat{B}_2, \hat{B}_3, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2)$

$= (0.3, 0.3667, 0.3333, 0.5278, -0.4930, 0.0068, 1.0198, 1.0042, 3.9780)$ (from code)

Bayes classifier:

$$\underset{y\in\{1,2,3\}}{\arg\max}\, \Gamma_y(x;\hat{\theta}) = \underset{y\in\{1,2,3\}}{\arg\max} \left(\frac{1}{\sqrt{2\pi\hat{\sigma}_y^2}}\right)^{100} \hat{\pi}_y \, e^{-\frac{1}{2\hat{\sigma}_y^2}\left(x_1^2 + \sum_{r=2}^{100}(x_r - \hat{B}_y x_{r-1})^2\right)}$$

Note: Denominator of $\Gamma_y(x;\hat{\theta})$ is a constant

The in sample empirical risk is 0 (See code)

3. a) and b)

3a)

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{d+1}} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - y_i(\alpha + \beta x_i))^2 + \lambda \|\beta\|_2^2 \right)$$

let $\theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, let $V_i = \begin{pmatrix} y_i \\ y_i x_i \end{pmatrix}$, let $A = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$ where I is the $d \times d$ identity matrix

To find $\hat{\theta}$, we solve

$$\frac{\partial}{\partial \theta} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - \theta^T V_i)^2 + \lambda \theta^T A \theta \right) = 0$$

Matrix Cookbook eqn 84 and 85

$$\Rightarrow \frac{1}{n} \sum_{i=1}^{n} (-2(1 - \theta^T V_i) V_i) + \lambda 2 A\theta = 0$$

$$\Rightarrow 2\lambda n A\theta - 2 \sum_{i=1}^{n} (1 - \theta^T V_i) V_i = 0$$

$$\Rightarrow \lambda n A\theta - \sum_{i=1}^{n} V_i + \sum_{i=1}^{n} (\theta^T V_i) V_i = 0$$

Note $\theta^T V_i = V_i^T \theta$ = some scalar

$$\Rightarrow \lambda n A\theta + \sum_{i=1}^{n} (V_i^T \theta) V_i = \sum_{i=1}^{n} V_i$$

Note $(V_i^T \theta) V_i = V_i (V_i^T \theta)$

And $V_i (V_i^T \theta) = (V_i V_i^T) \theta$ by matrix associativity

$$\Rightarrow \lambda n A\theta + \sum_{i=1}^{n} V_i V_i^T \theta = \sum_{i=1}^{n} V_i$$

$\sum_{i=1}^{n} (V_i V_i^T \theta) = \left(\sum_{i=1}^{n} (V_i V_i^T) \right) \theta$ by matrix distributivity

$$\Rightarrow \left(\lambda n A + \sum_{i=1}^{n} V_i V_i^T \right) \theta = \sum_{i=1}^{n} V_i$$

$$\Rightarrow \hat{\theta} = \left(\lambda n A + \sum_{i=1}^{n} V_i V_i^T \right)^{-1} \sum_{i=1}^{n} V_i$$

b) Here $d = 2$. The linear least squares SVM therefore has the form

$$r(x; \theta) = \text{sign}(\alpha + \beta^T x)$$

$$= \text{sign}(\alpha + \beta_1 x_1 + \beta_2 x_2)$$

Using the data given and the estimator from part a), we have

$\alpha = 0.02731$
$\beta_1 = -0.29184$
$\beta_2 = 0.00737$

$$\Rightarrow r(x; \theta) = \text{sign}(0.02731 - 0.29184 x_1 + 0.00737 x_2)$$

Note: The lambda chosen in this section is the lambda used to calculate the alpha, beta1 and beta2 values.

One strategy to pick lambda is by picking the lambda that minimizes the sample empirical risk. I did this over the range 0 to 10 (in steps of 0.1) which in this case gives lambda = 1.313131. This range seems reasonable, since from *Figure 2*, the sample empirical risk increases with larger and larger lambda values. Testing higher values of lambda eventually resulted in the decision boundary not even passing through the data points.

Figure 2

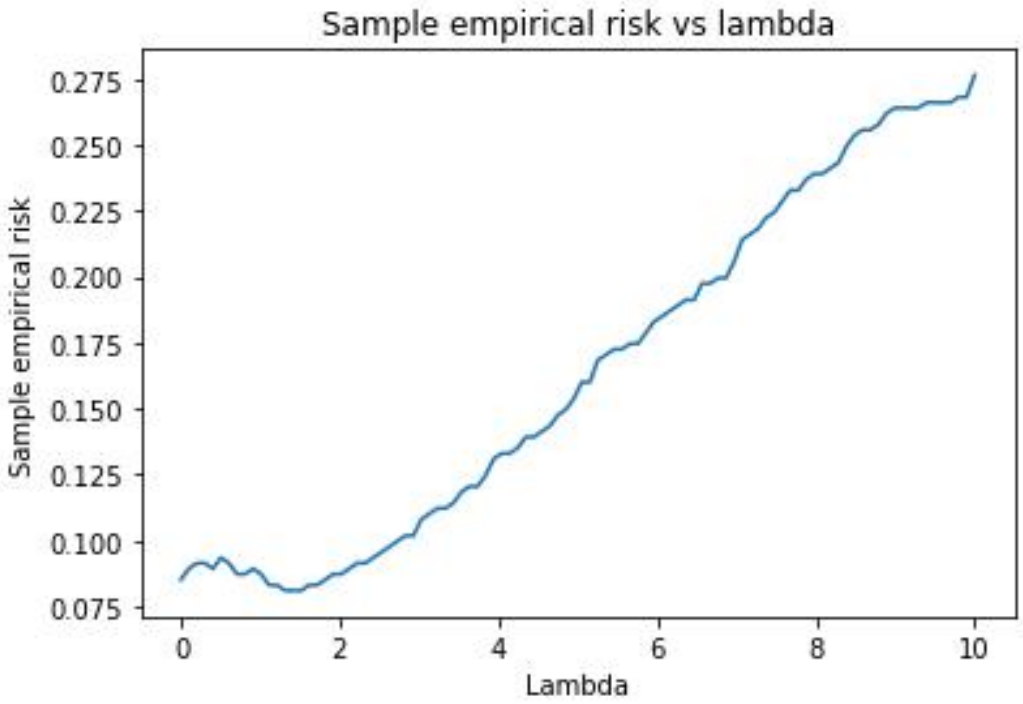
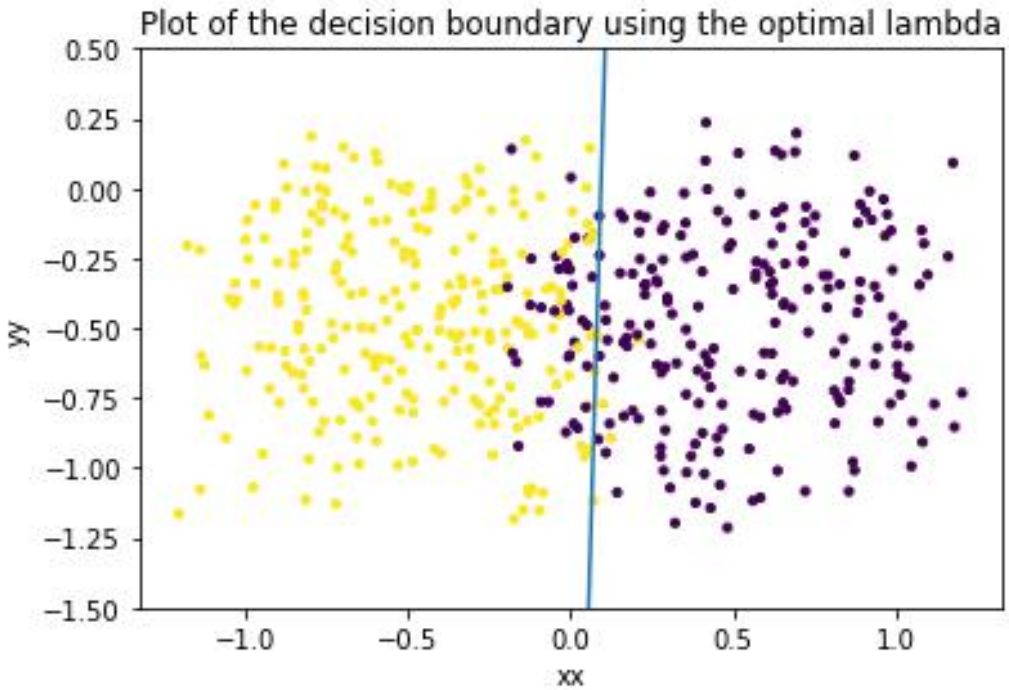


Figure 3



3. c) and d)

Using the data given, estimator from part a), and the same strategy used in part b) to find lambda, the decision boundary is

$$0.35425 - 0.20413x_1 - 0.11664x_2 - 0.12353x_3 - 0.017010x_4 = 0$$

and the estimate of the in-sample empirical classification risk is 0.020408.

Figure 4: Pink represents genuine banknotes and purple, forged.

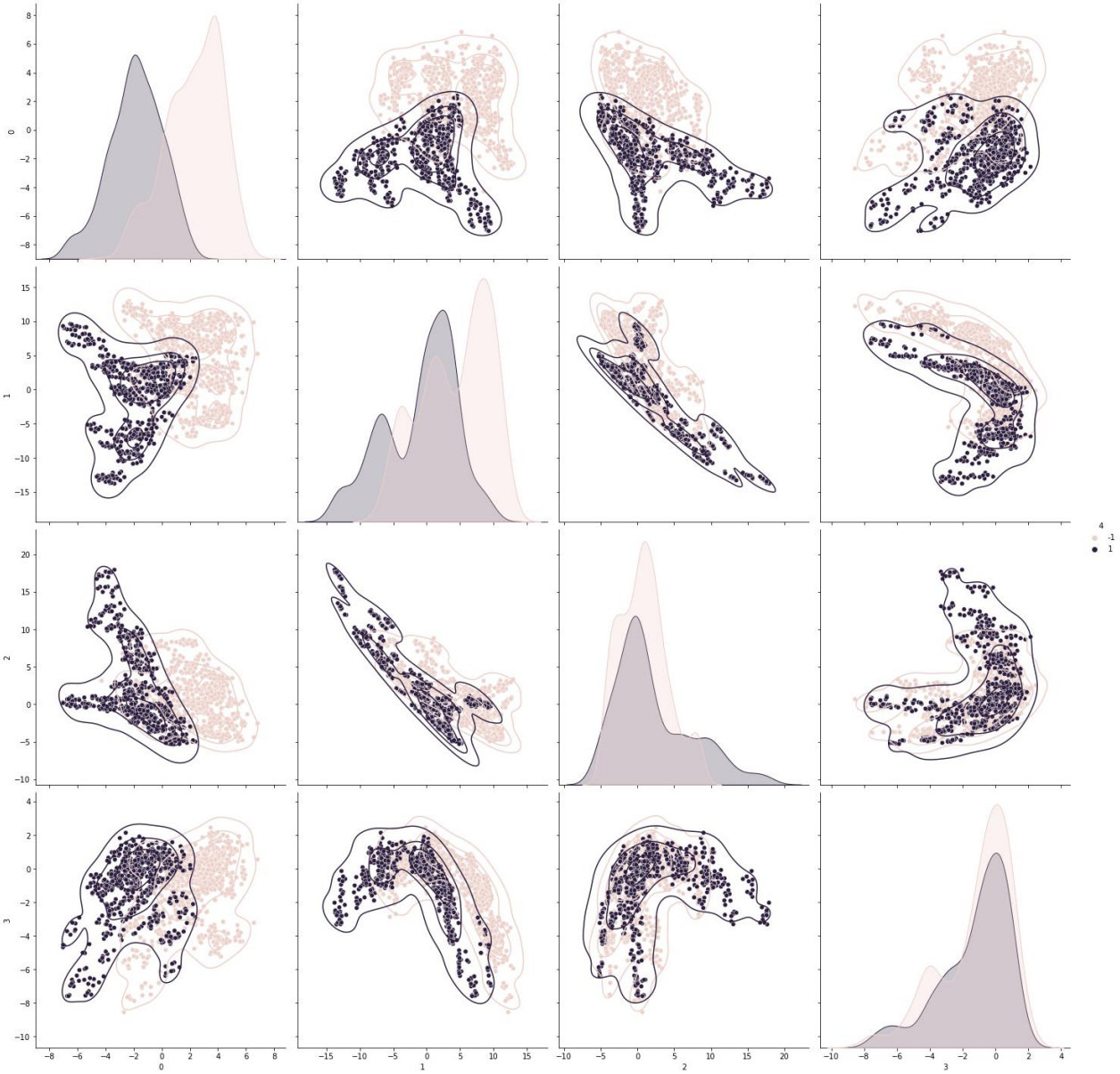


Figure 5: Orange represents correctly classified data and blue, misclassifications.

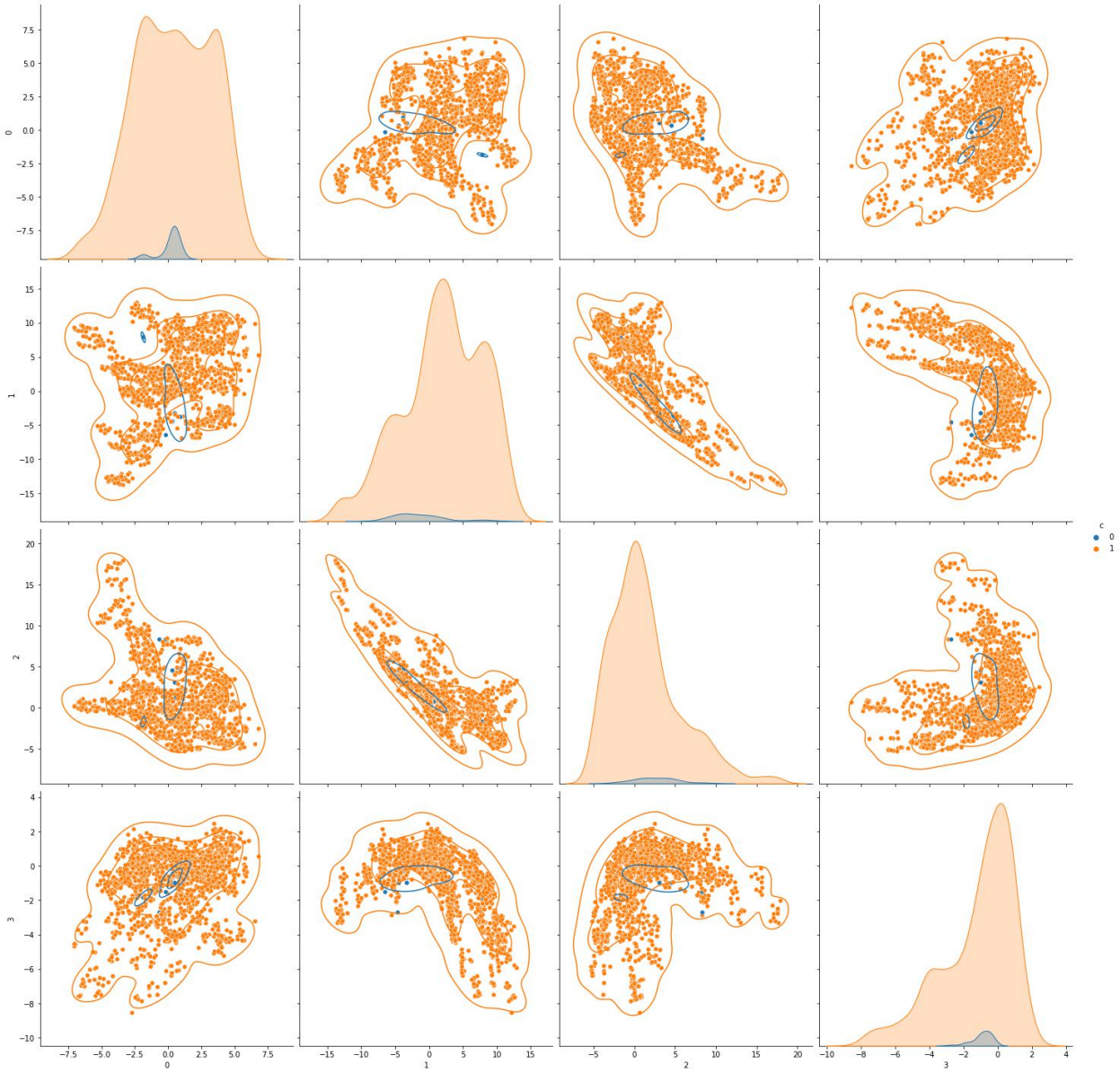
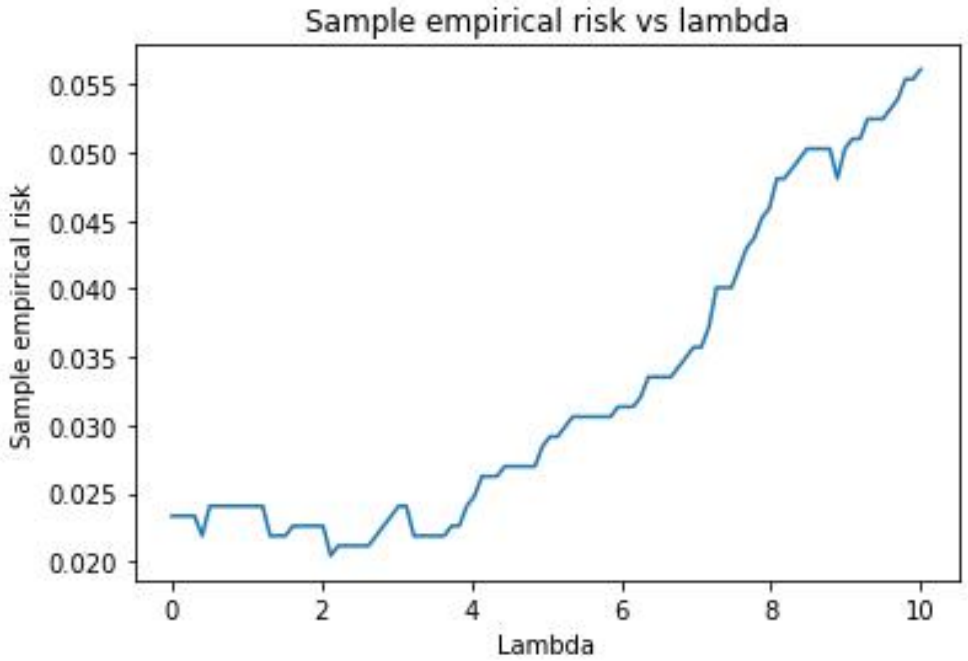


Figure 6: The lambda which minimizes risk in this range is 2.1212



3. e)

The linear classifier appears to work very well given that the in-sample empirical risk is only 0.020408, corresponding to 33 misclassifications out of 1372 observations. However, it is not 0 which implies that the data is not perfectly linearly separable. Since the decision boundary for a linear classifier is a hyperplane, it cannot achieve perfect separation if the data is not linearly separable. One way to overcome this inadequacy is to perform non-linear transformations to our data before fitting in the hopes that this transformation will make the data linearly separable.

3. f)

I experimented with various transformations to the data. Among these, the transformation that gave the best empirical classification risk was simply to cube the 2nd feature (skewness), i.e raising all the x_2 to the third power for all the observations.

Applying the same linear classifier on this transformed data set yielded a sample empirical risk of 0.008017, corresponding to only 12 misclassifications out of 1372 observations. *Figure 7* shows a pairwise plot of the transformed data, and *Figure 8* visualizes the misclassifications.

Figure 7: Pink represents genuine banknotes and purple, forged.

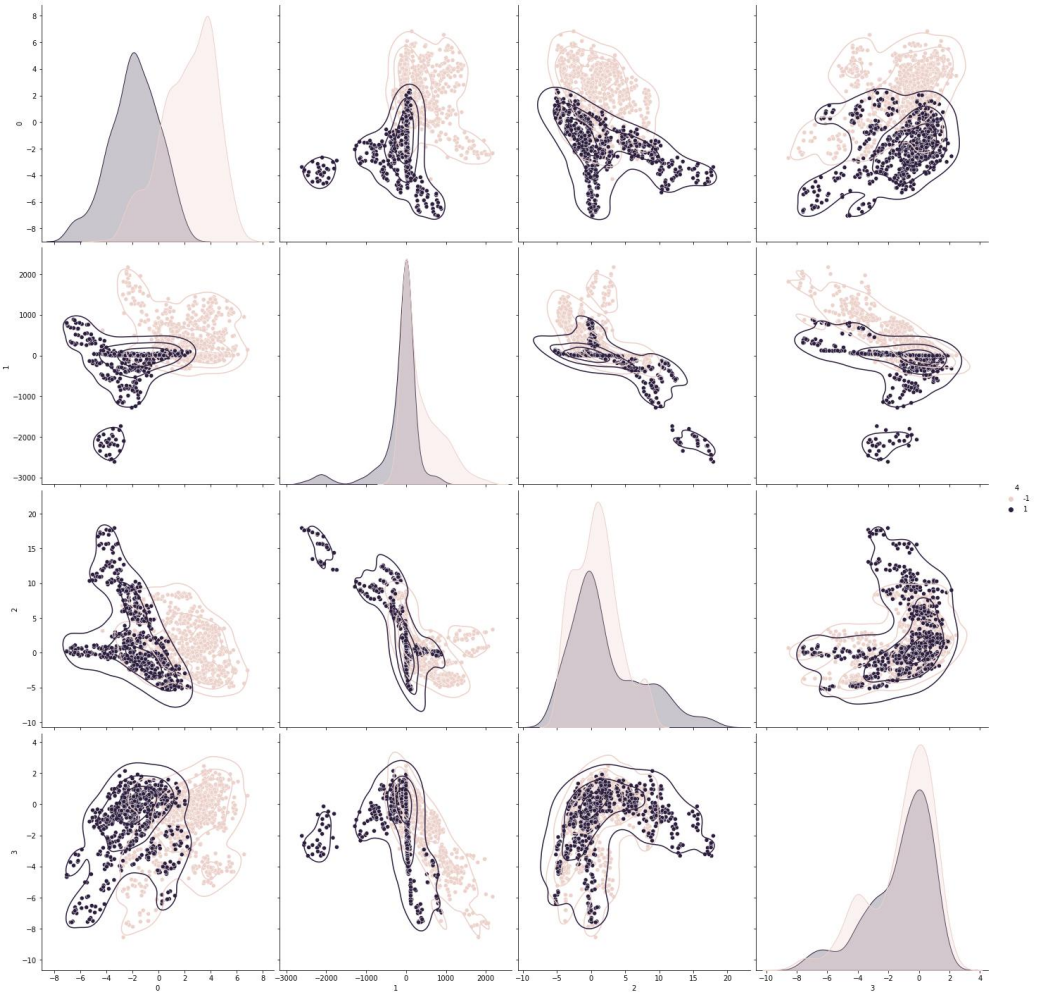


Figure 8: Orange represents correctly classified data and blue, misclassifications.

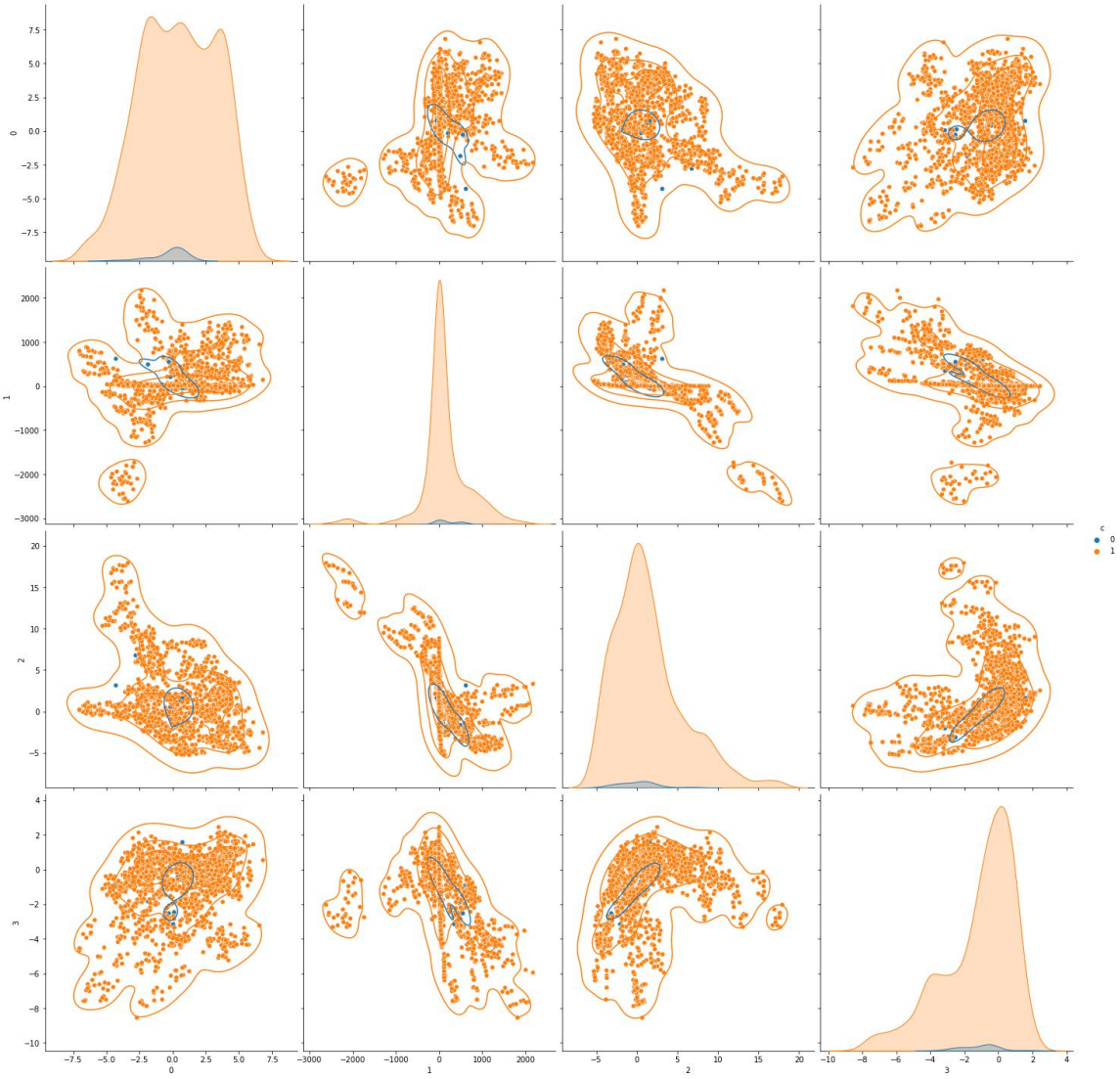


Figure 9: The lambda which minimizes risk in this range is 7.474747