Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Relatório Atividade 2
Estatística Multivariada 1

Grupo 2: Antônio M. dos Santos Jr. - 744845

Crystiane Souza - 760955

Douglas Nestlehner - 752728

Eric Sato - 729739

Setembro, 2021

Sumário

Capítulo 1

Introdução

Este trabalho tem como objetivo realizar análises multivariadas no banco de dados de Clientes de Cartão de Crédito, retirado da plataforma Kaggle. O banco de dados contém 10127 observações e 21 variáveis. A ferramenta de linguagem de programação utilizada, será o software estatístico R.

Contudo, para a realização deste estudo, as variáveis categóricas do banco de dados foram separadas e então, serão apenas analisadas 10 variáveis contínuas. Essas variáveis são:

- $\mathbf{X}_1$ **:** Idade do cliente (em anos);
- $\mathbf{X}_2$ **:** Período de relacionamento com banco (em meses);
- $\mathbf{X}_3$ **:** Limite de crédito no cartão de crédito (U$);
- $\mathbf{X}_4$ **:** Saldo rotativo total no cartão de crédito (U$);
- $\mathbf{X}_5$ **:** Linha de crédito aberta para compra (média dos últimos 12 meses);
- $\mathbf{X}_6$ **:** Mudança no valor da transação (Q4 sobre Q1);
- $\mathbf{X}_7$ **:** Valor total da transação (últimos 12 meses);
- $\mathbf{X}_8$ **:** Contagem total de transações (nos últimos 12 meses);
- $\mathbf{X}_9$ **:** Mudança na contagem de transações (Q4 sobre Q1);
- $\mathbf{X}_{10}$ **:** Taxa de utilização média do cartão.

Nesse sentido, no Capítulo 2 escolhemos um par de covariáveis e calculamos a distância estatística generalizada a distância centrada na origem e também a distância centrada no vetor de médias dessas covariáveis, além de fazer os gráficos da elipse e da rotação desses dados. Depois, calculamos a distância estatística para todo o conjunto de dados. Ainda no Capítulo 2, descrevemos a função de densidade de probabiliade da normal bivariada e plotamos o gráfico. No Capítulo 3, descrevemos sobre as conclusões do trabalho. Por fim, no Apêndice estão os códigos utilizados no trabalho.

Capítulo 2

Resultados

2.1 Distância Estatística Generalizada

2.1.1 Para duas variáveis

Para calcular a distância estatística generalizada, escolhemos as covariáveis X_1 e X_2 da base de dados, sendo essas, a idade do cliente (em anos) e o período de relacionamento com o banco (em meses). Primeiramente, vamos calcular a distância de $P = (\tilde{x}_1, \tilde{x}_2)$ à origem $O = (0, 0)$ em termos das coordenadas originais, que é representada por,

$$d(O, P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2}, \tag{2.1.1}$$

em que,

$$a_{11} = \frac{\cos^2(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) + 2\cos(\theta)\sin(\theta) + S_{22}\sin^2(\theta)} + \frac{\sin^2(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) - 2\cos(\theta)\sin(\theta) + S_{22}\sin^2(\theta)}$$

$$a_{22} = \frac{\sin^2(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) + 2\cos(\theta)\sin(\theta) + S_{22}\sin^2(\theta)} + \frac{\cos^2(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) - 2\cos(\theta)\sin(\theta) + S_{22}\sin^2(\theta)}$$

$$a_{12} = \frac{\cos(\theta)\sin(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) + 2\cos(\theta)\sin(\theta) + S_{22}\sin^2(\theta)} + \frac{\sin(\theta)\cos(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) - 2\cos(\theta)\sin(\theta) + S_{22}\sin^2(\theta)}$$

Contudo, é necessário encontrar qual o θ mais adequado para ser utilizado ao longo dos cálculos. Nesse sentido, temos uma aproximação de qual será o θ apresentado na Figura 2.1.

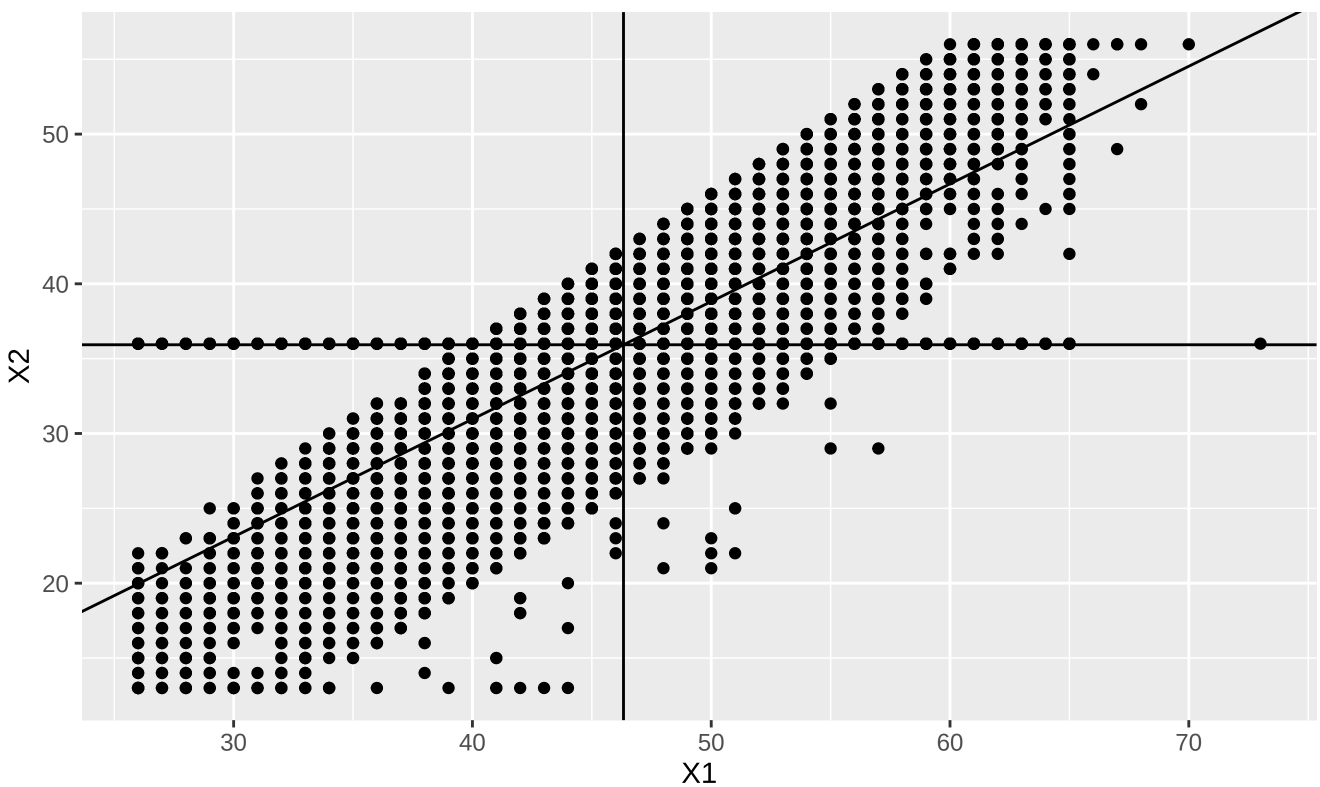

Figura 2.1: Gráfico de dispersão das covariáveis X_1 e X_2 com reta traçada

Nesse sentido, analisando a Figura 2.1, podemos escolher um θ que tenha 35º graus, logo,

$$a_{11} = 0.02597891;$$

$$a_{22} = 0.04366909;$$

$$a_{12} = -0.02430169;$$

e que a distância estatística generalizada (2.1.1) do par de covariáveis X_1 e X_2 é representada pela Figura 2.2.

```
> options(max.print = 200)
> d
  [1] 5.807672 6.490751 5.910202 5.093833 4.471855 5.468324 6.778748 4.055102
  [9] 5.236466 5.697712 4.950879 8.154883 6.330924 4.482500 7.213892 5.563440
 [17] 5.697712 5.137987 8.220990 5.609426 6.202953 7.553500 5.046367 5.634327
 [25] 6.523368 4.702174 7.136303 8.281537 5.297735 6.202953 5.960126 7.066920
 [33] 5.341078 6.176223 7.356069 6.240726 6.240726 5.379000 7.314762 8.394780
 [41] 6.028417 6.561138 6.418268 5.834418 4.476173 5.498235 6.886165 7.707052
 [49] 5.950774 6.246799 6.327110 6.375089 8.202486 5.764907 6.330924 5.764907
 [57] 5.764907 7.386010 5.379648 5.186776 7.029332 5.629954 4.951466 5.155460
 [65] 6.646431 6.036222 7.643710 4.999184 5.557195 5.971518 6.363568 6.583668
 [73] 6.606180 5.484536 5.379648 6.069157 5.468324 4.487696 5.498235 5.799996
 [81] 5.713942 5.468324 6.835405 7.707052 6.025789 6.922814 6.025789 5.327435
 [89] 5.095701 6.940298 5.910202 5.910503 5.519508 6.466523 5.060001 7.920783
 [97] 5.436140 6.025789 7.386010 5.713942 6.380612 5.137987 5.989403 5.155460
[105] 4.911684 5.159695 6.069157 5.468324 8.249745 6.153429 6.540708 5.379000
[113] 5.379000 6.153429 5.629954 5.764907 7.038930 6.049081 5.764907 6.646431
[121] 5.835781 6.240726 6.954766 6.093969 6.583668 5.634327 7.673754 6.384933
[129] 6.704102 5.835781 5.379000 5.740020 7.026466 6.504920 5.713942 5.664671
[137] 6.069157 5.915384 7.033703 4.491028 5.853292 6.363568 6.082777 6.617716
[145] 7.817692 5.853292 4.961862 7.355247 5.634327 4.582382 4.622450 8.484757
[153] 5.619434 6.093969 6.954766 5.379000 5.853292 5.910503 5.379648 5.240071
[161] 6.721390 6.617716 5.574884 6.093969 6.132787 6.459729 5.634327 7.166957
[169] 5.361631 6.954766 5.519508 6.704102 6.069157 6.634726 5.341078 5.519508
[177] 6.466523 8.438349 6.423900 6.617716 5.359627 5.910202 7.932295 6.744303
[185] 5.634327 5.910503 5.598156 8.076629 5.835781 4.884839 7.117729 4.841617
[193] 7.117729 6.446268 6.561138 5.495515 5.341078 6.553764 8.145316 6.617716
 [ reached getOption("max.print") -- omitted 9927 entries ]
```

Figura 2.2: Distância estatística generalizada do par de covariáveis X_1 e X_2.

Distância em torno da origem

Uma outra forma de se calcular a distância entre um ponto (P) e a origem (O), é utilizando a seguinte fórmula:

$$d(O,P) = \sqrt{x'S^{-1}x} \tag{2.1.2}$$

Onde:

- x': Matriz transposta do vetor das covariáveis;
- S^{-1}: Matriz inversa de S (Var-Cov);
- x: Matriz do vetor das covariáveis.

Nesse contexto, utilizando a fórmula 2.1.2 chegamos aos resultados apresentados na Figura 2.3. O mesmo resultado seria obtido se utilizássemos o comando mahalanobis(X, 0, cov(X)) disponível no R.

```
> options(max.print = 200)
> d2
  [1] 5.661820 6.213674 6.415772 5.016773 5.423816 5.496080 6.474909 4.009381
  [9] 4.825941 5.997715 5.254952 8.129688 7.173533 4.395272 7.139938 5.510562
 [17] 5.997715 5.127056 7.784177 5.623063 5.952709 7.733945 5.116703 5.865795
 [25] 6.736439 5.186127 7.359917 7.968871 5.489739 5.952709 6.842221 6.736382
 [33] 5.171841 6.652894 7.268211 7.016785 7.016785 5.274326 7.161436 8.275579
 [41] 5.729373 6.411527 6.282390 6.119894 4.755246 6.353370 6.988221 7.448425
 [49] 5.791253 6.488778 6.486423 6.754163 8.244121 6.133516 7.173533 6.133516
 [57] 6.133516 7.091092 5.489130 5.242881 6.903552 6.013092 4.888266 5.513594
 [65] 6.628367 6.375758 7.259940 6.240976 6.254443 6.242480 6.373623 6.865125
 [73] 6.736810 5.301738 5.489130 6.710729 5.496080 4.615328 6.353370 5.866668
 [81] 5.862692 5.496080 6.870101 7.448425 6.748695 6.605617 6.748695 5.369347
 [89] 5.536743 7.110176 6.415772 6.113038 5.614687 6.611481 5.810425 7.990742
 [97] 5.613699 6.748695 7.091092 5.862692 6.180243 5.127056 6.792563 5.513594
[105] 5.825298 5.366170 6.710729 5.496080 7.874450 6.862461 7.207523 5.274326
[113] 5.274326 6.862461 6.013092 6.133516 6.647172 6.237183 6.133516 6.628367
[121] 6.272948 7.016785 6.700253 5.920807 6.865125 5.865795 7.351985 6.612447
[129] 6.540785 6.272948 5.274326 5.467592 7.114277 6.500900 5.862692 5.532514
[137] 6.710729 5.692051 8.326894 4.921326 5.987680 6.373623 6.951901 7.656823
[145] 7.482520 5.987680 5.114453 6.998067 5.865795 5.381710 5.038418 8.498164
[153] 6.461913 5.920807 6.700253 5.274326 5.987680 6.113038 5.489130 5.363890
[161] 6.988359 7.656823 5.738024 5.920807 6.238541 6.389355 5.865795 7.240550
[169] 5.773781 6.700253 5.614687 6.540785 6.710729 6.344260 5.171841 5.614687
[177] 6.611481 8.385153 7.332550 7.656823 5.620104 6.415772 7.521983 6.646171
[185] 5.865795 6.113038 6.214418 7.653062 6.272948 5.120316 7.124206 5.539526
[193] 7.124206 6.742228 6.411527 5.893170 5.171841 6.616793 8.487265 7.656823
 [ reached getOption("max.print") -- omitted 9927 entries ]
```

Figura 2.3: Distância estatística do par de covariáveis X_1 e X_2 em torno da origem.

Distância em torno da média

Uma outra distância que podemos ter interesse em calcular, é a distância generalizada estatística para o par de covariáveis X_1 e X_2 em torno da média $(\bar{x}_1, \bar{x}_2)$. Sendo assim, a fórmula da distância passa a ser:

$$d(Q,P) = \sqrt{(x-\bar{x})'S^{-1}(x-\bar{x})}$$

onde,

- x: Matriz do vetor de covariáveis;

- $\bar{x}$: Vetor de médias da matriz de covariáveis;
- S^{-1}: Matriz inversa de S (Var-Cov).

assim, obtendo a distância que está representada pela Figura 2.4.

```
> options(max.print=200)
> d3
  [1] 0.85437423 1.26135923 0.93731275 1.00358842 2.17684596 0.48373537
  [7] 1.42813032 1.84901598 1.90459558 0.32842559 0.62304618 2.42992527
 [13] 1.95223488 1.53716745 1.52834771 0.65822311 0.32842559 0.80803963
 [19] 2.52540700 0.46135578 1.13234120 1.96115742 0.71203935 0.12563455
 [25] 0.95727847 1.01139627 1.58109342 2.51818645 0.29085542 1.13234120
 [31] 1.86195888 1.62104408 1.09265573 1.05415901 1.65859573 1.74924839
 [37] 1.74924839 0.88967523 1.63803619 2.62768035 1.25666681 1.06913800
 [43] 0.96792683 0.39408739 1.07621464 1.66985623 1.24326677 2.01239484
 [49] 0.88276972 0.71216378 0.71867774 1.03718240 2.51103192 0.53136773
 [55] 1.95223488 0.53136773 0.53136773 1.78701106 0.34319039 0.61731301
 [61] 1.38643326 0.50267957 1.08046774 0.51705302 1.01228499 0.66817321
 [67] 2.05507543 2.88772834 1.27315113 0.48751514 0.76072720 1.09259267
 [73] 0.98043648 1.05891529 0.34319039 1.34327756 0.48373537 1.16616691
 [79] 1.66985623 0.32522543 0.13877703 0.48373537 1.17546133 2.01239484
 [85] 1.50848352 1.52190924 1.50848352 0.53858639 0.69517878 1.33870241
 [91] 0.93731275 0.33360990 0.28080653 0.84936115 1.41287909 2.24897420
 [97] 0.16701185 1.50848352 1.78701106 0.13877703 1.06621360 0.80803963
[103] 1.68216764 0.51705302 1.87686952 0.41979214 1.34327756 0.48373537
[109] 2.51361637 1.54626252 1.71659295 0.88967523 0.88967523 1.54626252
[115] 0.50267957 0.53136773 1.70493127 0.45982420 0.53136773 1.01228499
[121] 0.73433571 1.74924839 1.44998687 0.92926535 1.09259267 0.12563455
[127] 2.02369206 0.83325707 1.17650092 0.73433571 0.88967523 1.23053848
[133] 1.37489182 0.88557951 0.13877703 0.84590360 1.34327756 1.05501879
[139] 3.37314840 1.11915099 0.25947849 0.76072720 1.88327984 2.56119664
[145] 2.14445261 0.25947849 0.66641543 1.83185976 0.12563455 1.71182219
[151] 0.99278327 2.77351333 1.65721148 0.92926535 1.44998687 0.88967523
[157] 0.25947849 0.33360990 0.34319039 0.43775750 1.21156766 2.56119664
[163] 0.07819008 0.92926535 0.51792187 0.90518629 0.12563455 1.50660792
[169] 0.54601360 1.44998687 0.28080653 1.17650092 1.34327756 1.34083158
[175] 1.09265573 0.28080653 0.84936115 2.69394425 2.15522183 2.56119664
[181] 0.24488998 0.93731275 2.28706329 1.14058336 0.12563455 0.33360990
[187] 1.07782977 2.40555756 0.73433571 0.68136371 1.43929440 1.35187858
[193] 1.43929440 0.97700868 1.06913800 0.50788431 1.09265573 0.91189302
[199] 2.71067668 2.56119664
 [ reached getOption("max.print") -- omitted 9927 entries ]
```

Figura 2.4: Distância estatística generalizada do par de covariáveis X_1 e X_2 em torno da média.

Ao compararmos os resultados das distâncias, vemos que a distância estatística generalizada das covariáveis X_1 e X_2 e a distância com centro na origem possuem resultados muito próximos. Contudo, quando comparamos os essas distâncias com a distância que possui a origem em torno da média, vemos que os valores são menores e diferentes dos outros.

Nesse sentido, após realizar os cálculos da distância estatística generalizada de duas formas diferentes, plotamos o gráfico de dispersão das covariáveis X_1 e X_2 com a elipse centrada na média e também plotamos o gráfico de dispersão com a rotação da elipse, que são representados, respectivamente, pelas Figuras 2.5 e 2.6.

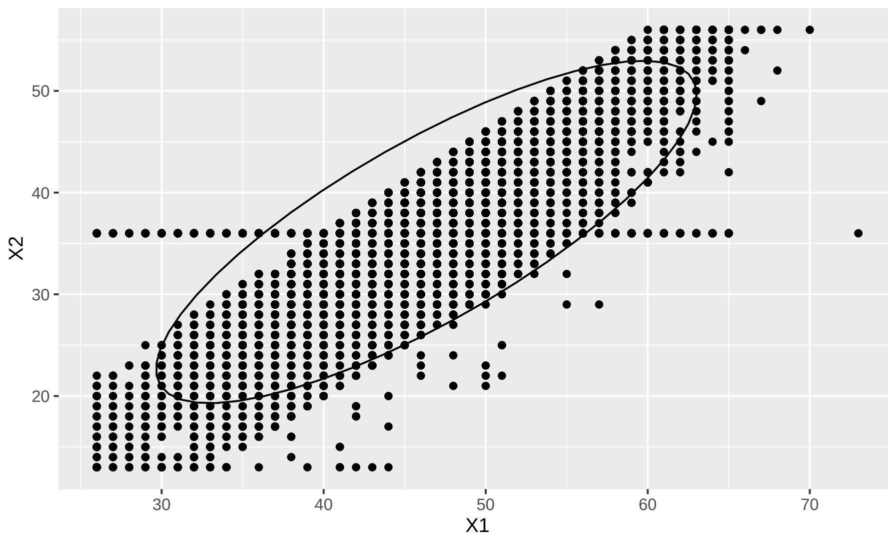

Figura 2.5: Gráfico de dispersão e elipse do par de covariáveis X_1 e X_2.

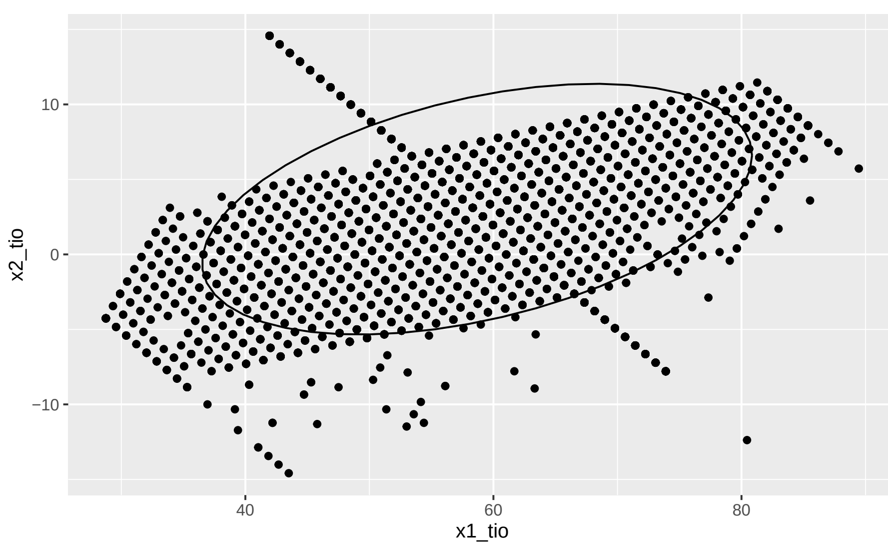

Figura 2.6: Gráfico da rotação do par de covariáveis X_1 e X_2.

2.1.2 Para todo o conjunto de dados

Para calcular a distância estatística para todo o conjunto de dados, usaremos a extensão da distância estatística para mais de duas dimensões. Nesse sentido, considerando que os pontos P e Q possuem p coordenadas, onde $P = (x_1, x_2, ...x_p)$ e $Q = (x_1, x_2, ..., x_q)$. Além disso, considerando as variâncias amostrais $s_{11}, s_{22}, ..., s_{pq}$ constrídas a partir de n medições, respectivamento. Então, a distância estatística de P para Q será de,

$$d(P,Q) = \sqrt{\frac{(x_1 - x_1)^2}{s_{11}} + \frac{(x_1 - x_2)^2}{s_{12}} + \cdots + \frac{(x_p - x_q)^2}{s_{pq}}}.$$

A distância estatística para todo o conjunto de dados está representada na Figura 2.7.

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.00	59.68	365.77	332.98	291.42
[2,]	59.68	0.00	306.09	273.30	231.74
[3,]	365.77	306.09	0.00	32.80	74.35
[4,]	332.98	273.30	32.80	0.00	41.56
[5,]	291.42	231.74	74.35	41.56	0.00
[6,]	84929.22	84869.54	84563.45	84596.24	84637.80
[7,]	470.47	410.79	104.70	137.49	179.05
[8,]	5150.85	5091.17	4785.08	4817.87	4859.43
[9,]	1991909.63	1991849.95	1991543.86	1991576.65	1991618.21
[10,]	1581380.58	1581440.26	1581746.36	1581713.56	1581672.00

	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	84929.22	470.47	5150.85	1991910	1581381
[2,]	84869.54	410.79	5091.17	1991850	1581440
[3,]	84563.45	104.70	4785.08	1991544	1581746
[4,]	84596.24	137.49	4817.87	1991577	1581714
[5,]	84637.80	179.05	4859.43	1991618	1581672
[6,]	0.00	84458.75	79778.37	1906980	1666310
[7,]	84458.75	0.00	4680.38	1991439	1581851
[8,]	79778.37	4680.38	0.00	1986759	1586531
[9,]	1906980.41	1991439.16	1986758.78	0	3573290
[10,]	1666309.80	1581851.05	1586531.43	3573290	0

Figura 2.7: Distância estatística de todo conjunto de dados.

Ao analisarmos a Figura 2.7, notamos que a menor distância que ocorre são entre as covariáveis (X_3 , X_4), (X_4 , X_5) e (X_1 , X_2). Ademais, as maiores distâncias que ocorrem é quando relacionamos as covariáveis X_9 ou X_{10} com qualquer outra covariável.

2.2 Normal Bivariada

A função de densidade de probabilidade da Normal Bivariada, é dada por,

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12})}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

Após isso, atribuímos valores para μ e Σ e plotamos o gráfico da densidade bivariada com duas visões diferentes, que está representada pela Figura 2.8.

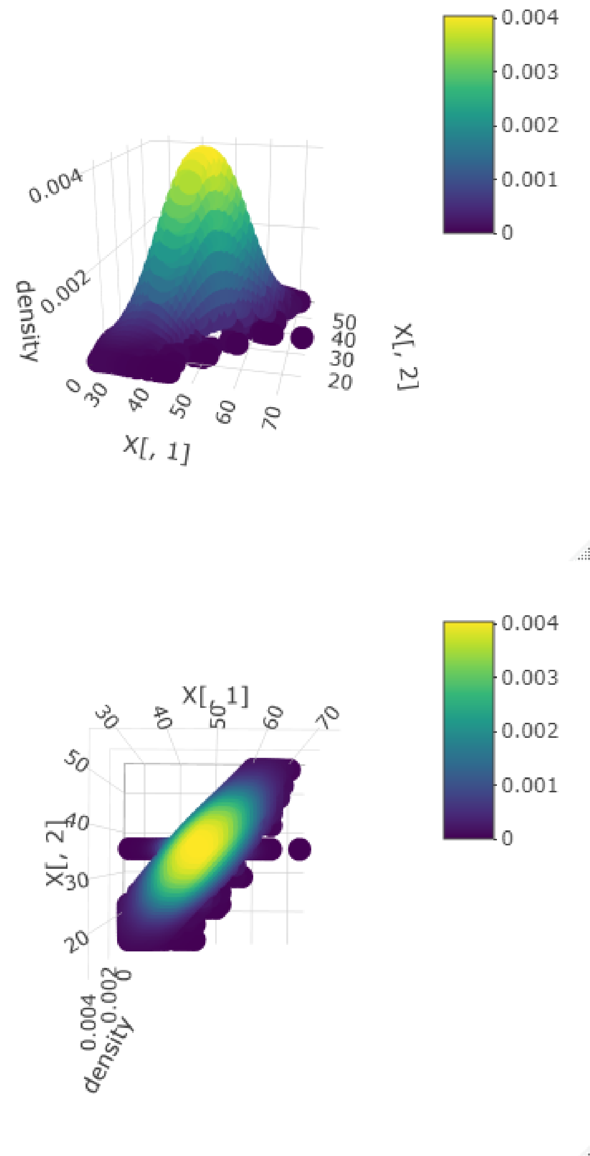

Figura 2.8: Gráficos da densidade da normal bivariada.

Capítulo 3

Conclusão

Sabe-se que são inúmeras as formas de se calcular a distância entre pontos. De certa forma, todas são usuais, variando a interpretação e o ambiente/circunstância qual estão sendo aplicadas.

A distância mais conhecida e comum é a Euclidiana, que basicamente é provada pela aplicação repetida do teorema de Pitágoras. Aplicando essa fórmula como distância, o espaço euclidiano torna-se um espaço métrico. Sua usabilidade é então limitada por conta disso, ela não leva em consideração informações estatísticas como média, variância e correlação das variáveis.

Sendo assim, para melhor analisar a distância em um ambiente variável, detém-se a distância estatística, qual se municia além dos informes usuais de distância, da variabilidade observada, atribuindo um peso maior para variáveis mais concentradas em torno da média e um peso menor para aquelas de ampla abrangência.

Quando comparamos então estas distâncias, quanto maior a variabilidade do ambiente, mais elas se destoam. Percebe-se com os resultados que a distância estatística torna-se importante não só por conta da sua forma de aplicação, mas por motivos interpretativos. Interpretação essa qual teríamos outra completamente diferente, e não personalizada, caso utilizássemos a distância euclidiana para metrificarmos as informações do nosso banco de dados.

Apêndice A

Códigos

```
### Atividade 2 - Estat stica Multivariada 1 ###

library(data.table)
library(ggplot2)

## Importando a base de dados
original <- read.csv("E:/MULTIVARIADA_1/datasets/BankChurners.csv",
                     header = TRUE, sep = ",")


dados = original[,-c(22,23)]
nomes=c("Identificador","Atividade","Idade","Sexo","Dependentes","N vel
   Educacional","Estado Civil",
        "Renda Anual", "Tipo do Cart o", "Per odo de Relacionamento", "N  de
    Produtos Mantidos",
        "Meses inativos U.A","N  de Contatos U.A","Limite de Cr dito","Saldo
   Rotativo","M dia de Cr dito Aberto U.A",
        "Mudan a no Valor Transacional","Valor Total da Transa  o U.A","N
   de Transa  es U.A","Mudan a no N  Transacional", "Taxa de Utiliza  o
   M dia")

dadosnum=dados[,c(3,10,14,15,16,17,18,19,20,21)]
dadoscat=dados[,-c(3,10,14,15,16,17,18,19,20,21)]

## Organizando base Numerica
nomes2=c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")
setnames(dadosnum, nomes2)

dadosnum
```

```
## Escolhendo as 2 covariaveis
df_ATV2 = data.frame(dadosnum$X1, dadosnum$X2)
nomes3=c("X1","X2")
setnames(df_ATV2, nomes3)
df_ATV2

## Calculando a linha de regress o
reg <- lm(X2 ~ X1, data = df_ATV2)
reg

## Encontrando o theta
c <- ggplot(df_ATV2, aes(x=dadosnum$X1, y=dadosnum$X2))  + geom_point()
c + geom_vline(xintercept = mean(dadosnum$X1)) + geom_hline(yintercept = mean(
    dadosnum$X2)) +
  labs(x="X1", y="X2") + geom_abline(intercept = -0.4801, slope = 0.7859)

## Valor do theta
t = 35*pi/180

## Gr fico de dispers o entre x1 e x2 e elipse
a <- ggplot(df_ATV2, aes(x=dadosnum$X1, y=dadosnum$X2))
a + geom_point() + labs(x="X1", y="X2") + stat_ellipse()

## Gr fico de rota  o
x1_tio <- dadosnum$X1*cos(t) + dadosnum$X2*sin(t)
x2_tio <- -dadosnum$X1*sin(t) + dadosnum$X2*cos(t)

b <- ggplot(df_ATV2, aes(x= x1_tio, y = x2_tio))
b + geom_point() + stat_ellipse()

## Calculando a matriz VarCov
s = cov(df_ATV2)
s

## Distancia Estat stica Generalizada
a11 = cos(t)^2/(s[1]*cos(t)^2 + s[4]*sin(t)^2 + 2*s[2]*cos(t)*sin(t)) +
  sin(t)^2/(s[1]*sin(t)^2 + s[4]*cos(t)^2 - 2*s[2]*cos(t)*sin(t))
a11

a22 = sin(t)^2/(s[1]*cos(t)^2 + s[4]*sin(t)^2 + 2*s[2]*cos(t)*sin(t)) +
  cos(t)^2/(s[1]*sin(t)^2 + s[4]*cos(t)^2 - 2*s[2]*cos(t)*sin(t))
```

```r
a22

a12 = cos(t)*sin(t)/(s[1]*cos(t)^2 + s[4]*sin(t)^2 + 2*s[2]*cos(t)*sin(t)) -
  sin(t)*cos(t)/(s[1]*sin(t)^2 + s[4]*cos(t)^2 - 2*s[2]*cos(t)*sin(t))
a12

d = sqrt(a11*(df_ATV2$X1^2) + a22*(df_ATV2$X2^2) + 2*a12*df_ATV2$X1*df_ATV2$X2)
d

generalizada = df_ATV2
generalizada$distancia = d


# Outra
A = data.matrix(df_ATV2)
d2=0
for(i in 1:nrow(df_ATV2)){

  d2[i] = sqrt(t(A[i,])%*%solve(s)%*%A[i,])

}

d2
generalizada$Distancia_2 = d2

View(generalizada)

#### mahalanobis
d3 = sqrt(mahalanobis(df_ATV2, 0, cov(df_ATV2)))
d3

generalizada$mahalanobis = d3




# Distancia do vetor de m dias
# Distancia d(Q,P)

# Obtendo o vetor X1 - X1_mean
novoX1 = 0
for(i in 1:nrow(df_ATV2)){
```

```
  novoX1[i] = df_ATV2$X1[i] - mean(df_ATV2$X1)
}

novoX1

# Obtendo o vetor X2 - X2_mean
novoX2 = 0
for(i in 1:nrow(df_ATV2)){
  novoX2[i] = df_ATV2$X2[i] - mean(df_ATV2$X2)
}

novoX2


novoX = data.frame(novoX1, novoX2)


# calculando a distancia
C = data.matrix(novoX)
d3 = 0
for(i in 1:nrow(df_ATV2)){

  d3[i] = sqrt(t(C[i,])%*%solve(s)%*%C[i,])

}


options(max.print=100)
d3


d4 = sqrt(mahalanobis(df_ATV2, colMeans(df_ATV2), cov(df_ATV2)))
d4


generalizada$distancia_media = d3
generalizada$mahalanobis_media = d4
View(generalizada)

######## grafico da densidade
library(mvtnorm)
library(MASS)
```

```r
library(plotly)

mu = colMeans(df_ATV2)
Sigma = cov(df_ATV2) # var-covariance matrix

X = df_ATV2

density <- dmvnorm(X, mean = mu, sigma = Sigma)

plot_ly(x=~X[,1], y=~X[,2], z=~density,
        type = "scatter3d", color=density)

# gerar uma normal
X = mvrnorm(10000, mu, Sigma)

density = dmvnorm(X, mean = mu, sigma = Sigma)

plot_ly(df_ATV2, x=~X1, y=~X2, z=~density,
        type = "scatter3d", color=density)
```