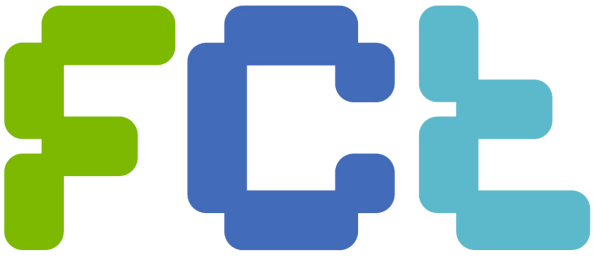


FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

World Happiness Report

Lucas Santos Fischer

dezembro, 2018

Contents

Introdução

No ambito de estudar um *dataset* foi escolhido o World Happiness Report. Este é um *dataset* com informação sobre o *ranking* de felicidade de cada país, e as variáveis que contribuem para o mesmo, durante um periodo de três anos (2015, 2016 e 2017). Para além da informação sobre o **nome** e **região** de um determinado país, estes três ficheiros possuem as seguintes variáveis que explicam a importância de um certo fator para a posição de um país no *rank* de felicidade:

- **GDP** - A importância do produto interno bruto de um país para o *rank*
- **Family** - A importância da família para o *rank*
- **Life Expectancy** - A importância da esperança média de vida de um país para o *rank*
- **Freedom** - A importância da liberdade de um país para o *rank*
- **Government Corruption** - A importância da corrupção do governo de um país para o *rank*
- **Genorosity** - A importância de genorosidade de um país para o *rank*

Com estas informações existem várias perguntas interessantes que serão respondidas com este trabalho, nomeadamente: **Serão os dados relativamente iguais entre os três anos?** ; **Quais serão as variáveis que mais contribuem para a felicidade de um país?** ; **Como se distinguem os diferentes continentes no que toca a sua felicidade?** Neste trabalho serão respondidas estas perguntas bem como demonstrado todos os passos necessários até obter as suas respostas.

Métodos

Durante o estudo deste *dataset* foram utilizados diferentes metodos que serão resumidos neste capítulo. De maneira a ser possível responder a todas as perguntas propostas foram utilizados os métodos de **Análise de Perfis**, **Teste de Mauchly**, **Análise de componentes principais** e **Análise Discriminante**.

Análise de perfis

A primeira grande pergunta com interesse em responder é: **Serão os dados significativamente iguais entre os três anos?** Neste capítulo será descrito o processo feito de maneira a obter a sua resposta. Visto que o estudo é realizado sobre amostras independentes, uma vez que não existe ligação entre um *dataset* e outro, i.e. não existe ligação entre os diferentes anos de medição, o processo estatístico para obter a resposta a esta pergunta passa pela **Análise de Perfis**. Neste processo existem três questões com particular interesse:

- Serão os perfis paralelos?
- Serão os perfis coincidentes (dado que são paralelos)?
- Serão os perfis horizontais (dado que são paralelos e coincidentes)?

Teste ao paralelismo dos perfis

A primeira questão a ser respondida deverá sempre ser quanto ao paralelismo dos perfis uma vez que esta constitui uma condição de modo aos seguintes estudos serem possíveis. Este teste ao paralelismo pode expressar-se pela hipótese:

$$H_0 : C\mu_1 - C\mu_2 = 0 \tag{1}$$

Onde em 1 **C** representa a matriz de contrastes dos perfis que neste estudo equivale à matriz:

$$C = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \tag{2}$$

Sob $H0$ a estatistica teste é:

$$T^2 = (C(\bar{x}_1 - \bar{x}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) CS_{pooled}C' \right]^{-1} (C(\bar{x}_1 - \bar{x}_2)) \sim T^2_{q-1}(n_1 + n_2 - 2) \tag{3}$$

e o quantil critico:

$$\frac{(n_1 + n_2 - 2)(q-1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q);1-\alpha} \tag{4}$$

Este teste permite-nos concluir se a variância entre os pontos observados nos *datasets* é igual independentemente dos perfis (anos dos *datasets*).

Teste à coincidência dos perfis

Caso seja provado o paralelismo entre os dois perfis pode-se estudar a sua coinciência. No teste da coincidência de perfis procura-se determinar se os valores médios de cada ponto observados nos dois perfis são os mesmo, dado que já se comprovou a igualdade na variância dos pontos. Este teste é expressado pela hipotese descrita (Bispo 2018) em 5

$$H_0 : 1'\mu_1 - 1'\mu_2 = 0 \tag{5}$$

Onde a estatistica teste sob $H0$ é igual à expressão descrita em

$$T^2 = (1(\bar{x}_1 - \bar{x}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) 1'S_{pooled}1 \right]^{-1} (1'(\bar{x}_1 - \bar{x}_2)) \tag{6}$$

e o seu quantil critico:

$$F_{(1, n_1+n_2-q);1-\alpha} \tag{7}$$

Teste à horizontalidade dos perfis

A ultima questão relevante a responder na análise de perfis relaciona-se com a horizontalidade dos perfis, i.e. testar se as médias são iguais independentemente das variáveis. Uma vez verificada o paralelismo e coincidência entre dois perfis faz sentido estudar quanto a sua horizontalidade. Este teste pode ser expressado pela hipótese descrita em 8 (Marques, n.d.).

$$H_0 : \frac{C(\mu_1 + \mu_2)}{2} = 0 \tag{8}$$

Sob H_0 advém a estatistica teste

$$T^2 = (n_1 + n_2)(C\bar{x})'(CS_{pooled}C')^{-1}C\bar{x} \tag{9}$$

e o quantil critico que é idêntico ao quantil critico descrito em 4.

Teste de Mauchly

De modo testarmos a adequabilidade de usarmos a técnica da análise de componentes principais precisamos de primeiro verificar se existem variáveis correlacionadas/redundantes de modo a fazer sentido a aplicação desta técnica. O teste de Mauchly pretende testar isto mesmo, testando se a matriz de correlações é significativamente diferente da matriz idêntidade através da hipótese 10

$$H_0 : \Sigma = \sigma^2 I \tag{10}$$

Sob H_0 temos a estatistica teste 11

$$U^* = -\left(n - 1 - \frac{2p^2 + p + 2}{6p}\right) lnU \tag{11}$$

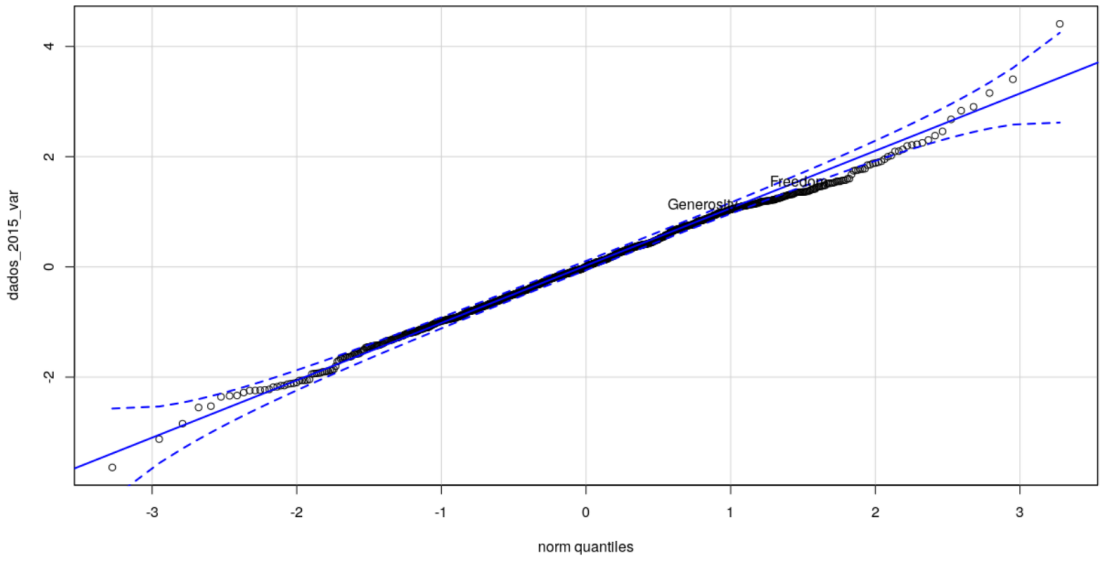
com $U = \frac{p^p \det S}{(tr(S))^p}$

e com o quantil critico descrito por 12

$$\chi^2_{\frac{1}{2}p(p+1)-1;1-\alpha} \tag{12}$$

No entanto o teste de esfericidade de Mauchly pressupõe normalidade multivariada, por isso é necessário primeiro verificarmos esta propriedade sobre os nossos dados, algo que conseguimos fazer através de um **gráfico QQ**. Este gráfico relaciona os quantis amostrais (empiricos) retirados dos nossos *datasets* e relaciona-os com os quantis teóricos que seriam obtidos caso os dados sejam provenientes de uma distribuição normal multivariada.

Figure 1: QQ *plot* do *dataset* de 2015 de modo a verificar o seu ajustamento a uma normal multivariada



Observando o gráfico 1 conseguimos observar que os pontos não se afastam das linhas tracejadas, isto indica que o *dataset* está bem ajustado a uma normal multivariada (algo que se verifica para os restas dois *datasets*) então podemos executar o teste de esfericidade de Mauchly.

Análise de componentes principais

A resposta às questões: **Quais serão as variáveis que mais contribuem para a felicidade de um país?** e **Poderemos representar os dados de uma maneira mais resumida?** são obtidas através de uma técnica conhecido como Análise de componentes principais. O objetivo desta técnica é a **redução da dimensionalidade** dos dados, mantendo a maior percentagem de variabilidade possível, e a **transformação de variáveis correlacionadas em variáveis independentes**. Esta redução da dimensionalidade consiste em encontrar uma projeção em menores dimensões que permita preservar a maior variabilidade possível dos dados. Esta projeção é encontrada sobre os vetores próprios da matriz variância-covariância **S** então as K componentes principais são definidas por:

$$y = a_{k1}x_1 + a_{k2}x_2 + ... + a_{kp}x_p \tag{13}$$

onde a_k representa o k-ésimo vetor próprio de **S**.

Uma vez obtidas as componentes principais é necessário escolher quais as que serão retidas. Esta decisão pode ser feita segundo diversos critérios :

- Percentagem de variabilidade explicada - Neste critério podemos estabelecer uma percentagem da variabilidade explicada que queremos reter e guardar as componentes principais suficientes para atingir essa percentagem
- Critério de Kaiser - Reter apenas as variáveis cujo valor próprio seja superior a 1 (caso as variáveis estejam padronizadas)
- Scree plot - Metodo mais gráfico que determina o número de componentes principais pelo ponto onde a linha desenhada diminui acentuadamente de inclinação tornando-se horizontal

Análise Discriminante

De maneira a se responder à ultima questão: **Como se distinguem os diferentes continentes no que toca a sua felicidade?** é necessário identificar um discriminante entre os vários continentes. Este discriminante é obtido através da análise discriminante. O objetivo desta técnica é discriminar grupos a partir de informação recolhida sobre os elmentos que constituem esses grupos, em termos práticos esta técnica coincide com determinar as combinações lineares de variáveis que maximizam a diferença entre os grupos (*discriminação*). Com base nestas combinações lineares é possível posteriormente prever a pertença de um ponto não agrupado a um certo grupo (*classificação*).

Visto neste caso existirem mais que dois grupos (6 continentes) os coeficientes das combinações lineares são obtidas maximizando a razão 14

$$\frac{a'Ba}{a'Wa} \tag{14}$$

Então os coeficientes das combinações lineares que melhor descriminam os grupos equivalem assim aos coeficientes dos vetores próprios da matriz $W^{-1}B$

$$y_i = a_i'x(i = 1, ..., s) \tag{15}$$

Resultados

Neste capítulo serão apresentados os resultados obtidos para os diferentes metodos de estudo a cima descritos, bem como uma interpretação aos mesmos de modo a conseguir-se responder às questões propostas para este *dataset*. De maneira a ser possível estudar o *dataset* é necessário primeiro processar os dados de maneira a facilitar a sua utilização, este pre-processamento encontra-se no ficheiro `world_happiness.R`.

Análise preliminar dos dados

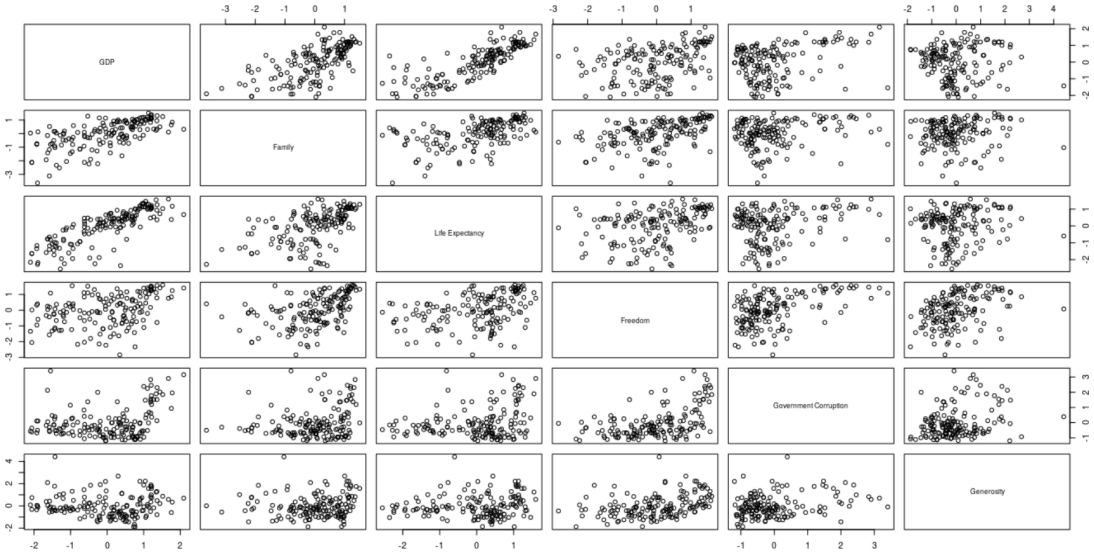
A análise preliminar dos dados (ou *Exploratory data analysis*) constitui uma boa prática para iniciar o estudo de qualquer *dataset*. Esta análise tem como objetivo obter uma familiarização dos dados de maneira a ser possível obter uma maior sensibilidade quanto ao estudo dos mesmos.

Nesta pequena análise introdutória são normalmente verificadas as médias e variâncias dos dados bem como alguns gráficos que permitam visualizar estes resultados e também outros como a correlação entre os dados.

```
pairs(dados_2015_var)
```

```
#pairs(dados_2016_var)
#pairs(dados_2017_var)
```

Figure 2: Representação gráfica da correlação entre pares de variáveis para o *dataset* de 2015



Fazendo o *plot* das correlações entre pares de variáveis através da função `pairs` conseguimos obter uma representação gráfica da correlação entre as variáveis presente na figura 2, especialmente entre as variáveis **GDP** e as variáveis **Family** e **Life Expectancy**.

Com esta análise introdutória ficamos assim sensibilizados à correlação entre as variáveis e à possível existência de variáveis redundantes, algo que será estudado com mais detalhe no capítulo **Teste de Mauchly** e **Análise de componentes principais**.

Análise de perfis

Durante o estudo da análise de perfis foi primeiro executado os testes de paralelismo, coincidência e horizontalidade sobre os perfis de 2015 e 2016. O primeiro teste executado foi o teste ao paralelismo sobre estes dois perfis visto que o paralelismo entre perfis é uma condição para a sua coincidência.

```r
#Retorna TRUE caso H0 não seja rejeitada
teste_paralelismo(dados_2015_var, dados_2016_var, contrastes)
```

```
## [1] TRUE
```

Não existindo evidências estatisticas para rejeitar H_0 conclui-se então que os dois perfis são paralelos, i.e. a variância entre os pontos presentes em ambos os *datasets* é independente dos perfis. Tendo verificado o paralelismo entre estes dois perfis faz sentido estudar quanto a sua coincidência.

```r
#Retorna TRUE caso H0 não seja rejeitada
teste_coincidencia(dados_2015_var, dados_2016_var)
```

```
## [1] TRUE
```

Da execução deste teste podemos observar que não existem evidências estatisticas para rejeitar H_0 então conclui-se que estes dois perfis são coincidentes, logo com esta conclusão já podemos admitir que os perfis são exatamente iguais. O restante teste para executar sendo ele o teste à horizontalidade dos perfis pretende determinar se as médias são iguais independentemente das variãveis dos *datasets*

```r
#Retorna TRUE caso H0 não seja rejeitada
teste_horizontalidade(dados_2015_var, dados_2016_var, contrastes)
```

```
## [1] TRUE
```

Mais uma vez através da execução das instruções a cima descritas obtemos a conclusão de que não existem evidências estatisticas para rejeitar H_0, logo podemos admitir que os dois perfis são **paralelos**, **coincidentes** e **horizontais**.

Uma vez comprovado o paralelismo, coincidência e horizontalidade dos dois perfis podemos utilizar um dos dois perfis para executarmos os mesmos testes mas agora contrastando com o perfil do *dataset* de 2017, onde é possível obter as mesmas conclusões que as a cima descritas, sendo assim possível concluir que os três *datasets* (2015, 2016 e 2017) são **paralelos**, **coincidentes** e **horizontais**, i.e. os dados são significativamente iguais entre os três *datasets*.

```r
teste_paralelismo(dados_2016_var, dados_2017_var, contrastes) &&
  teste_coincidencia(dados_2016_var, dados_2017_var) &&
  teste_horizontalidade(dados_2016_var, dados_2017_var, contrastes)
```

```
## [1] TRUE
```

Teste de Mauchly

Para a execução do teste de esfericidade de Mauchly, de modo a determinar a adequabilidade do uso da análise de componentes principais, foi implementada uma função no ficheiro `world_happiness.R` que executa a o teste de esfericidade de Mauchly para o *dataset* enviado como argumento da função.

```r
mauchly_test(dados_2015_var) && mauchly_test(dados_2016_var) &&
  mauchly_test(dados_2017_var)
```

```
## [1] FALSE
```

Verificando que para os três *datasets* existem evidências estatisticas para rejeitar H_0 conclui-se então que para os três *datasets* a matriz de correlações é significativamente diferente da matriz idêntidade o que significa que existem variâveis cuja sua correlação é significativamente diferente de 0 justificando assim a adequabilidade do uso da análise de componentes principais

O R disponibiliza também uma função que permite executar o teste de esfericidade de mauchly

```
mauchly.test(lm(as.matrix(dados_2015_var)~1))
mauchly.test(lm(as.matrix(dados_2016_var)~1))
mauchly.test(lm(as.matrix(dados_2017_var)~1))
```

Onde, observando os *p-values* obtidos, se tiram as mesmas conclusões que a cima descritas.

Análise de componentes principais

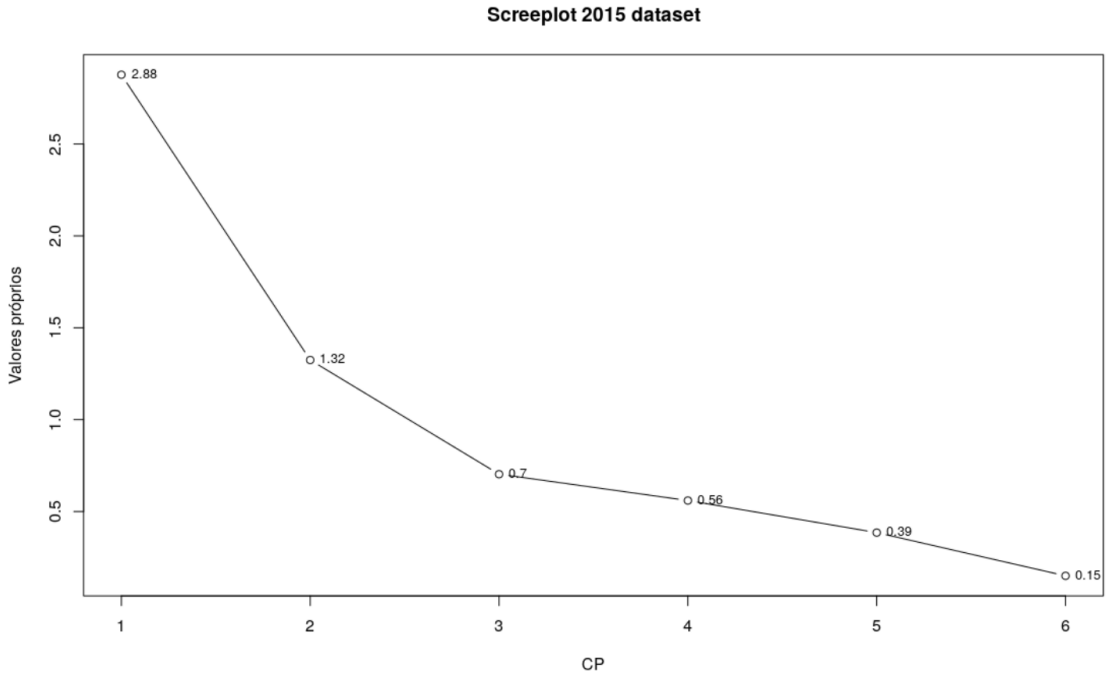
Tendo sido justificada a adequabilidade do uso do método da análise de componentes principais, seguiu-se então à implementação de uma função no ficheiro `world_happiness.R` de modo a executar este método.

```
componentes_principais_2015 = pca(s_2015, "Screeplot 2015 dataset")
```

ev	std	prop	cprop
2.877	1.696	0.479	0.479
1.325	1.151	0.221	0.700
0.703	0.839	0.117	0.817
0.560	0.748	0.093	0.911
0.385	0.621	0.064	0.975
0.150	0.387	0.025	1.000

Table 1: Resultado da análise de componentes principais para o *dataset* de 2015

Figure 3: Scree plot obtido através das componentes principais do *dataset* de 2015



A execução desta função escreve para a consola uma tabela com informação sobre os valores próprios da matriz de variância-covariância fornecida, bem como a variabilidade explicada por cada componente principal e também a vairabilidade cumulativa até a componente principal K. Esta função também esboça o *Scree plot* 3 e finalmente retorna todas as componentes principais obtidas. Analisando a tabela escrita por esta função e seguindo o critério da percentagem de variabilidade explicada, querendo manter uma variabilidade de 80% seria necessário reter as três

primeiras componentes principais, visto que só com a partir da terceira componente principal se consegue obter uma variabilidade cumulativa de 81.7%. Seguindo agora o método da análise do *Scree plot* pode-se observar que é um metodo muito mais subjetivo devido à dificuldade em analisar com precisão o ponto onde o declive do gráfico passa a ser horizontal, mas é possível identificar que seria em torno da terceira componente principal, escolhendo assim três componentes principais. Por fim temos o critério de Kaise, um critério mais objetivo, que foi o critério utilizado durante este estudo. Observando a tabela escrita pela execução da função `pca` verifica-se que seriam retidas apenas as primeiras duas componentes principais visto serem as únicas cujo valor próprio é superior a 1. Este critério baseia-se no facto de que se as variáveis não possuem a capacidade de explicar a sua própria variabilidade (o que acontece quando o seu valor próprio é inferior a 1) então não têm grande interesse em serem retidas no *dataset*.

O R possui também uma função que permite estudar a análise de componentes principais de uma maneira mais simples através da função `prcomp` e também uma função `screeplot` que premite esboçar o *Scree plot*

```
pca_2015 = prcomp(dados_2015_var)
summary(pca_2015)
screeplot(pca_2015, type = "l", main = "Screeplot 2015 dataset")
```

Onde, usando os mesmos critérios que acima descritos, obtemos exatamente as mesmas conclusões.

O mesmo procedimento foi feito para os restantes dois *datasets* obtendo resultados semelhantes, então é possível identificar que a primeira e segunda componente principal de cada *dataset* é suficiente para representar cerca de 70% da variabilidade total dos *datasets* podendo-se utilizar as mesmas como substitutas à utilização de todas as variáveis do *dataset*. No entanto, tendo em conta que a dimensionalidade dos *datasets* em estudo não é muito elevada, o resto do estudo será feito sobre todas as variáveis visto que não é necessário descartar cerca de 30% de variabilidade quando a dimensionalidade não é um problema para estes *datasets*.

Análise Discriminante

De modo a responder à última questão de interesse proposta para este *dataset* recorreu-se à biblioteca `MASS` que disponibiliza a função `lda` que permite realizar o método da análise discriminante de uma maneira simples, para tal primeiro foi necessário trabalhar os dados de maneira a obter um novo *data-frame* em que as regiões dos países fosse o continente em que se situam, este processamento encontra-se no ficheiro `world_happiness.R`

Obtendo um *data-frame* trabalhado é possível utilizar a função `lda` da biblioteca `MASS` e obter os coeficientes das combinações lineares

```
library("MASS")

resultado_lda = lda(Region ~ GDP + Family + `Life Expectancy` + Freedom +
                      `Government Corruption` + Generosity, data = todos_grupos)

# Obtenção dos coeficientes das combinações lineares
resultado_lda$scaling
```

Tendo as funções que melhor discriminam os diferentes grupos do *dataset* sabe-se como se destinguem os diferentes continentes. Com estas funções é interessante também implementar um classificador que consiga classificar países que não estejam agrupados a nenhum continente, caso por exemplo existisse algum erro na construção da tabela de dados dos países. Este processo de classificação é também possível de implementar através da função `predict` onde é passado o objeto que guarda a informação sobre as funções descriminantes (`resultado_lda`) e os dados que se pretende classificar.

```r
#Construção do classificador
predictions = predict(resultado_lda, newdata = todos_grupos[, -7])

#Obtenção das previsões
predictions$class
```

É possível agora comparar as previsões feitas pelo classificador com a *ground truth* inserindo ambos os dados numa tabela conhecida por *confusion matrix*.

```r
#Criação da confusion matrix com a ground truth e as previsões
table(todos_grupos[, 7], predictions$class)
```

	Africa	Asia	Europe	North America	Oceania	South America
Africa	47	2	9	0	0	0
Asia	4	8	5	0	1	3
Europe	3	3	41	0	0	3
North America	0	0	2	0	0	0
Oceania	0	0	1	0	1	0
South America	2	1	10	0	0	9

Table 2: *Confusion Matrix* que relaciona a *ground truth* com as previsões obtidas do classificador

Da *confusion matrix* é possível tirar informação sobre a taxa de acerto ($\frac{\#acertos}{\#exemplos}$) e a taxa de erro ($\frac{\#erros}{\#exemplos}$).

```r
#A taxa de acerto é então a soma da diagonal sobre o número total de observações
taxa_acerto = sum(diag(table(todos_grupos[, 7], predictions$class))) / nrow(todos_grupos)

#E a taxa de erro é 1 - taxa de acerto
taxa_erro = 1 - taxa_acerto

cat("Taxa de acerto:",taxa_acerto,"\nTaxa de erro:",taxa_erro)
```

```
## Taxa de acerto: 0.683871
## Taxa de erro: 0.316129
```

Conclusões

Dado como concluido o estudo a estes três *datasets* foi possível responder as perguntas **Serão os dados relativamente iguais entre os três anos?** ; **Quais serão as variáveis que mais contribuem para a felicidade de um país?** ; **Como se distinguem os diferentes continentes no que toca a sua felicidade?** As respostas a estas perguntas foram obtidas utilizando diversas técnicas de análise estatistica aprendidas durante a unidade curricular de Estatistica Multivariada lecionada pela professora Regina Bispo.

Quanto a igualdade entres os três anos de medição do indice de felicidade dos países observou-se que as variáveis que o compões mantêm-se relativamente iguais. Neste estudo observou-se também que as variáveis que mais contribuem para a felicidade de um país são o seu PIB (produto interno bruto) e a contribuição da família. Por fim foi também estudado quanto à diferença entre os continentes no que toca a sua felicidade obtendo as funções que melhor os descriminam, foi também construido um pequeno classificador com base nestas funções que obteve uma taxa de acerto de 70% o que implica que com base nas variáveis e países existentes no *dataset* não é possível classificar corretamente a que continente cada país pertence.

Referências

Bispo, Regina. 2018. “Estatistica Multivariada.”

Marques, Filipe. n.d. “Sebenta de Estatistica Multivariada.”