Statistische modellen en data-analyse: project 1

Toon Nolten, r0258654
<toon.nolten@student.kuleuven.be>

April 2014

Inleiding

De toegewezen dataset die ik moest analyseren was "Turnhout.xlsx". Dit is een kleine dataset met een aantal gegevens over de gemeenten rond Turnhout, bvb.: aantal mannelijke inwoners, gemiddelde inkomens. Om de resultaten te verkrijgen waar dit verslag op gebaseerd is, moet de code uitgevoerd worden, wegens tijdsgebrek heb ik die niet in het verslag kunnen verwerken.

Opgave 1: Principaal component analyse

1 Voer een PCA analyse uit op jullie gegevens. Argumenteer waarom je de analyse baseert op de correlatie- of de covariantiematrix van de gegevens. Bepaal het aantal componenten dat je verkiest te behouden en verklaar je keuze.

Ik heb gekozen voor de analyse op basis van de correlatie matrix omdat de schalen van de verschillende variabelen veel verschillen (aantallen, euros, procenten). Op basis van de summary en screeplot van de pca met correlatie matrix, behoudt ik 7 principale componenten, zo behouden we 98.6%, 99% van de informatie met ongeveer de helft van de variabelen.

2 Bekijk de loadings van de weerhouden componenten en tracht de PCs te interpreteren aan de hand van de loadings. Ga na welke originele variabelen het belangrijkst zijn voor elk van de weerhouden PCs.

(afgeleid uit de loadings)

De eerste PC is ongeveer een gemiddelde waarbij de variabelen waarbij geteld wordt een groter gewicht hebben, deze variabelen komen in de andere

componenten nagenoeg niet meer voor. Belangrijkste variabele: Aantal overlijdens (meeste met zeer weinig marge)

De tweede component bevat immigratie, de verhouding jonge mensen, de kostprijs van huizen en de verhouding 80+/60+: Immigranten zijn meestal jongere mensen of ze hebben veel kinderen; jongere mensen en de prijs van huizen zijn ook sterk verbonden, zij kunnen zich immers meestal geen dure villa's veroorloven. Belangrijkste variabele: Gemiddelde verkoopprijs van woonhuizen

PC3 bevat voornamelijk gemiddeld inkomen, werkzaamheidsgraad, werkloosheid, grote belastingaangiften en immigratie (ook 80+/60+). Een verband tussen werkzaamheid, werkloosheid, aangiften en inkomens is te verwachten. Immigratie gebeurt misschien vooral daar waar de lonen gunstig zijn. Belangrijkste variabele: Gemiddeld inkomen per aangifte

PC4 bestaat uit immigratie, verhouding jonge mensen, inkomen, werkloosheid en aantal geboorten. Nogmaals het verband tussen immigranten en jongeren, nu ook aantal geboorten. En ook, werkloosheid en inkomen. Belangrijkste variabele: Immigratie saldo

PC5 immigratie, 80+/60+, werkzaamheid, verkoopprijs, inkomen. Een verband tussen werkzaamheid, inkomen en verkoopprijs van huizen is te verwachten. Deze zijn ook gerelateerd aan de verhouding bejaarden en een beetje aan immigratie. Belangrijkste variabele: Werkzaamheidsgraad

PC6 immigratie, jonge mensen, inkomen, werkzaamheid, werkloosheid en verkoopprijs. Opnieuw het verband jonge mensen, immigratie. Ook, inkomen, verkoopprijs en immigratie, werkzaamheid, werkloosheid. Belangrijkste variabele: Gemiddelde verkoopprijs van woonhuizen

PC7 werkloosheid, werkzaamheid, inkomen, jonge mensen. Opnieuw jonge mensen, werkloosheid, werkzaamheid en inkomen. Belangrijkste variabele: Werkloosheidsgraad

3 Onderzoek de impact van de centrumstad op het resultaat van de PCA analyse.

(afgeleid uit biplots tussen verschillende PCs)

De eerste principale component wordt gedicteerd door de centrumstad, Turnhout. De andere componenten worden nagenoeg niet beinvloed door de centrumstad. Dit is te verklaren aan de hand van de samenstelling van de PCs: De eerste PC bevat aantallen en die zijn voor een grote centrumstad groter, de andere PCs bestaan vooral uit verhoudingen en gemiddelden waarop de absolute hoeveelheid inwoners weinig invloed heeft.

Opgave 2: Testen van de multivariate normaliteit

1 Onderzoek of de veronderstelling van multivariate normaliteit aanne- melijk zou zijn voor jullie gegevens. Gebruik hiervoor ook de scores van de PCA analyse.

Univariate marginalen: het aantal mannen, vrouwen, geboorten, overlijdens en kleine belastingaangiften is licht rechtsscheef. De andere marginalen komen redelijk goed overeen met de normale verdeling.

Bivariate marginalen: een aantal variabelen vertoont een sterke correlatie en bij een aantal variabelen zijn er mogelijk uitschieters maar over het algemeen zien de bivariate verdelingen er redelijk normaal verdeeld uit.

Mahalanobis afstanden lijken redelijk chi-kwadraat verdeeld. Wijken niet genoeg af van de rechte om te veronderstellen dat ze niet chi-kwadraat verdeeld zijn. De horizontale bij $\chi^2_{0.025}$ ligt onder alle punten, dit is niet wat ik verwachte te zien en ik ben niet zeker waarom dat zo is.

2 Als de multivariate normaliteit niet aannemelijk zou blijken te zijn, on- derzoek dan of je de data kan aanpassen om beter aan deze veronderstelling te voldoen.

De multivariate normaliteit lijkt wel aannemelijk.

Appendix: Code

```
# Voor deze opgave analyseren jullie gegevens afkomstig van de Vereniging
# van Vlaamse Steden en Gemeenten (http://www.vvsg.be). Deze vereniging
# brengt gegevens uit verschillende databanken samen en levert uitgebreide
# informatie over demografische en economische indicatoren per gemeente,
# regio, provincie, etc. Binnen het Vlaamse gewest worden 13 centrumsteden
# herkend, nl. Aalst, Antwerpen, Brugge, Genk, Gent, Hasselt,
# Kortrijk, Leuven, Mechelen, Oostende, Roeselare, Sint-Niklaas en Turnhout.
# Elke groep analyseert de gegevens van een regio bestaande uit een
# centrumstad en zijn omliggende gemeenten.

# De datasets bevatten de meest recente gegevens (afkomstig van 2011 of 2012)
# voor de volgende variabelen:

#+----------------+-------------------------------------------------------+
#|                | Totaal aantal mannen                                  |
#|                | Totaal aantal vrouwen                                 |
#|                | Aantal geboorten                                      |
#| Demografisch   | Aantal overlijdens                                    |
#|                | Immigratie saldo (Inwijkelingen-uitwijkelingen)       |
#|                | 80+/60+ verhouding (in%)                              |
#|                | (0-19)/totaal verhouding (in %)                       |
#+----------------+-------------------------------------------------------+
#|                | Aantal belastingsaangiften < 10.000 euro              |
#|                | Aantal belastingsaangiften > 50.000 euro              |
#| Economisch     | Gemiddeld inkomen per aangifte                        |
#|                | werkzaamheidsgraad (% werkenden van (18-64+) bevolking)|
#|                | werkloosheidsgraad (% werkzoekenden van (18-64+) bevolking)|
#|                | Gemiddelde verkoopprijs van woonhuizen                |
#+----------------+-------------------------------------------------------+

# De groepen worden samengesteld tijdens het practicum.
# De datasets worden als volgt verdeeld:

#        +-------+-----------+
#        | Groep | Dataset   |
#        +-------+-----------+
#        |   1   | Aalst     |
#        |   2   | Gent      |
#        |   3   | Hasselt   |
#        |   4   | Kortrijk  |
#        |   5   | Leuven    |
#        |   6   | Mechelen  |
#        |   7   | Sint-     |
#        |       |  Niklaas  |
#    ==> |   8   | Turnhout  |
#        +-------+-----------+
```

```
48 library(gdata)
49 turnhout = read.xls(xls='Turnhout.xlsx', sheet=1, pattern='Mannen')
50 rownames(turnhout) = turnhout[,1]
51 turnhout = turnhout[,-1]
52 colnames(turnhout) = c('Mannen',
53                        'Vrouwen',
54                        'n_geboorten',
55                        'n_overlijdens',
56                        'Immigratie_saldo',
57                        '80+/60+_(in%)',
58                        '(0-19)/totaal_(in_%)',
59                        'Aantal_aangiften_<_10.000_euro',
60                        'Aantal_aangiften_>_50.000_euro',
61                        'Gemiddeld_inkomen_per_aangifte',
62                        'Werkzaamheidsgraad',
63                        'Werkloosheidsgraad',
64                        'Gemiddelde_verkoopprijs_van_woonhuizen')
65 for(column in 1:ncol(turnhout)) {
66     turnhout[,column] = as.numeric(as.character(turnhout[,column]))
67 }
68
69 cat('\n_', 'Summary_turnhout.xlsx:', '\n',
70     '-----------------------------', '\n')
71 print(summary(turnhout))
72
73 # Opgave 1: PCA
74 #    1. Voer een PCA analyse uit op jullie gegevens. Argumenteer waarom je
75 #       de analyse baseert op de correlatie- of de covariantiematrix van de
76 #       gegevens. Bepaal het aantal componenten dat je verkiest te behouden
77 #       en verklaar je keuze.
78
79 pca_cov_turnhout = prcomp(turnhout, scale=FALSE)
80 pca_cor_turnhout = prcomp(turnhout, scale=TRUE)
81
82 cat('\n_', 'Pca_on_covariance_matrix:', '\n',
83     '-------------------------------', '\n')
84 print(summary(pca_cov_turnhout))
85 #screeplot(pca_cov_turnhout, type='lines')
86
87 cat('\n_', 'Pca_on_correlation_matrix:', '\n',
88     '-------------------------------', '\n')
89 print(summary(pca_cor_turnhout))
90 #screeplot(pca_cor_turnhout, type='lines')
91
92 ### Kies de correlatie matrix omdat de schalen van de verschillende
93 ### variabelen veel verschillen (aantallen, euros, procenten)
94 ### Behoud 7 principale componenten, zo behouden we 98.6, ~99% van de
95 ### informatie met ~1/2 van de variabelen.
```

```
97  #    2. Bekijk de loadings van de weerhouden componenten en tracht de PCs
98  #       te interpreteren aan de hand van de loadings. Ga na welke originele
99  #       variabelen het belangrijkst zijn voor elk van de weerhouden PCs.
100
101 cat('\n ', 'Loadings pca on correlation matrix:', '\n',
102     '-------------------------------------', '\n')
103 print(pca_cor_turnhout$rotation[,1:7])
104
105 ###   De eerste PC is ongeveer een gemiddelde waarbij de variabelen waarbij
106 ### getelt wordt een groter gewicht hebben.
107 ###      Aantal overlijdens (meeste met zeer weinig marge)
108 ###   De tweede component bevat immigratie, de verhouding jonge mensen,
109 ### de kostprijs van huizen en de verhouding 80+/60+:
110 ### Immigranten zijn meestal jongere mensen of ze hebben veel kinderen;
111 ### jongere mensen en de prijs van huizen zijn ook sterk verbonden, zij
112 ### kunnen zich immers meestal geen dure villa's veroorloven.
113 ###      Gemiddelde verkoopprijs van woonhuizen
114 ###   PC3 bevat voornamelijk gemiddeld inkomen, werkzaamheidsgraad,
115 ### werkloosheid, grote belastingaangiften en immigratie (ook 80+/60+).
116 ### Een verband tussen werkzaamheid, werkloosheid, aangiften en inkomens is
117 ### te verwachten. Immigratie gebeurt misschien vooral daar waar de lonen
118 ### gunstig zijn.
119 ###      Gemiddeld inkomen per aangifte
120 ###   PC4 bestaat uit immigratie, verhouding jonge mensen, inkomen,
121 ### werkloosheid en aantal geboorten.
122 ### Nogmaals het verband tussen immigranten en jongeren, nu ook aantal
123 ### geboorten. En ook, werkloosheid en inkomen.
124 ###      Immigratie saldo
125 ###   PC5 immigratie, 80+/60+, werkzaamheid, verkoopprijs, inkomen.
126 ### Een verband tussen werkzaamheid, inkomen en verkoopprijs van huizen is te
127 ### verwachten. Deze zijn ook gerelateerd aan de verhouding bejaarden en
128 ### een beetje aan immigratie.
129 ###      Werkzaamheidsgraad
130 ###   PC6 immigratie, jonge mensen, inkomen, werkzaamheid, werkloosheid en
131 ### verkoopprijs. Opnieuw het verband jonge mensen, immigratie. Ook, inkomen,
132 ### verkoopprijs en immigratie, werkzaamheid, werkloosheid.
133 ###      Gemiddelde verkoopprijs van woonhuizen
134 ###   PC7 werkloosheid, werkzaamheid, inkomen, jonge mensen.
135 ### Opnieuw jonge mensen, werkloosheid, werkzaamheid en inkomen.
136 ###      Werkloosheidsgraad
```

```
#    3. Onderzoek de impact van de centrumstad op het resultaat van de PCA
#       analyse.

biplotmatrix = function (x, y, ...){
    par(new=TRUE)
    biplot(pca_cor_turnhout, c(x,y), pc.biplot=TRUE)
}

#pairs(t(seq(7)), panel=biplotmatrix)

### De eerste principaal component wordt gedicteerd door de centrumstad,
### Turnhout. De andere componenten worden nagenoeg niet beinvloed door de
### centrumstad. Dit is te verklaren aan de samenstelling van de PCs:
### De eerste PC bevat aantallen en die zijn voor een grote centrumstad
### groter, de andere PCs bestaan vooral uit verhoudingen en gemiddelden
### waarop de absolute hoeveelheid inwoners weinig invloed heeft.

# Opgave 2: Testen multivariate normaliteit
#    1. Onderzoek of de veronderstelling van multivariate normaliteit
#       aannemelijk zou zijn voor jullie gegevens. Gebruik hiervoor ook de
#       scores van de PCA analyse.

#par(mfrow=c(3,5))
#qqnorm(turnhout[,1])
#qqline(turnhout[,1])
#qqnorm(turnhout[,2])
#qqline(turnhout[,2])
#qqnorm(turnhout[,3])
#qqline(turnhout[,3])
#qqnorm(turnhout[,4])
#qqline(turnhout[,4])
#qqnorm(turnhout[,5])
#qqline(turnhout[,5])
#qqnorm(turnhout[,6])
#qqline(turnhout[,6])
#qqnorm(turnhout[,7])
#qqline(turnhout[,7])
#qqnorm(turnhout[,8])
#qqline(turnhout[,8])
#qqnorm(turnhout[,9])
#qqline(turnhout[,9])
#qqnorm(turnhout[,10])
#qqline(turnhout[,10])
#qqnorm(turnhout[,11])
#qqline(turnhout[,11])
#qqnorm(turnhout[,12])
#qqline(turnhout[,12])
#qqnorm(turnhout[,13])
#qqline(turnhout[,13])
```

```
### Univariate marginalen: het aantal mannen, vrouwen, geboorten, overlijdens
### en kleine belastingaangiften is licht rechtsscheef. De andere marginalen
### komen redelijk goed overeen met de normale verdeling.

#pairs(turnhout)

### Bivariate marginalen: een aantal variabelen vertoont een sterke correlatie
### en bij een aantal variabelen zijn er mogelijk uitschieters maar over het
### algemeen zien de bivariate verdelingen er redelijk normaal verdeeld uit.

#y = mahalanobis(turnhout, center=sapply(turnhout, mean), cov=cov(turnhout))^2
#dof=13
#qqplot(qchisq(ppoints(500), df= dof), y,
#                main = expression("Q-Q plot for" ~~ {chi^2}[nu == dof]))
#qqline(y, distribution = function(p) qchisq(p, df = dof),
#                prob = c(0.25, 0.75), col = 2)
#abline(h=qchisq(p=0.975, df=dof))
#mtext("qqline(*, dist = qchisq(., df=13), prob = c(0.1, 0.6))")

### Mahalanobis afstanden lijken redelijk normaal verdeeld. Wijken niet
### genoeg af om anormaliteit te veronderstellen.
### De horizontale by chi^2_0.025 ligt onder alle punten, niet zekere wat
### dat betekent.

#    2. Als de multivariate normaliteit niet aannemelijk zou blijken te zijn,
#       onderzoek dan of je de data kan aanpassen om beter aan deze
#       veronderstelling te voldoen.




# Opgave 3: Clustering
#    1. Voer een cluster analyse uit op de gegevens. Kies op basis van de
#       resultaten volgens verschillende clustermethodes een ``optimaal''
#       aantal clusters.


#    2. Stel de clusterresultaten grafisch voor en bespreek de kwaliteit van
#       de gevonden clustering. Bespreek gelijkenissen en/of verschillen tussen
#       de methoden.


#    3. Je kan ook een clustering uitvoeren vertrekkende van de scores van de
#       PCA analyse. Zou dit zinvol kunnen zijn of heeft een voorafgaande PCA
#       analyse weining zin? Levert dit een ander resultaat voor jullie
#       gegevens?
```