

DRIE-WEG METHODEN VOOR HET ANALYSEREN VAN KWALITATIEVE EN KWANTITATIEVE TWEE-WEG GEGEVENS

Voor de exploratieve analyse van resultaten van (sociaal-wetenschappelijk) onderzoek is Principale Componenten Analyse (PCA) een veel gebruikte en nuttige techniek. Het doel van deze methode is een aantal variabelen efficiënt samen te vatten door middel van een aantal zogenaamde componenten.

PCA is alleen bruikbaar voor kwantitatieve variabelen. Voor het analyseren van niet-kwantitatieve variabelen, met name variabelen van nominaal meetnivo, zijn verschillende methoden voorgesteld, die elk slechts voor een deel hetzelfde doel bereiken als PCA. Enerzijds zijn er technieken ontwikkeld die gebaseerd zijn op gegeneraliseerde korrelatie-koëfficiënten. Dergelijke koëfficiënten drukken het verband uit tussen twee nominale variabelen of tussen één nominale variabele en één kwantitatieve variabele. Ze kunnen op dezelfde manier als gewone korrelatie-koëfficiënten gebruikt worden voor PCA van een stel variabelen. In tegenstelling tot PCA voor kwantitatieve variabelen, kan een dergelijke vorm van PCA voor nominale variabelen echter niet direct een weergave van de observatie-eenheden (objecten) leveren.

Anderzijds zijn er technieken voorgesteld waarin de categorieën van nominale variabelen worden opgevat als (binaire) variabelen en op deze variabelen een soort PCA wordt uitgevoerd. De bekendste representant van dit type methoden is Multipale Correspondentie Analyse (MCA). Deze methode heeft als voordeel boven de gegeneraliseerde korrelatie-koëfficiënten methode dat zij wel een weergave levert voor de objecten. Daar staat echter tegenover dat de techniek van het oorspronkelijke doel van PCA van de nominale variabelen afwijkt doordat het de categorieën van de variabelen centraal stelt in plaats van de variabelen zelf.

In dit onderzoek worden methoden voor PCA van nominale variabelen ontwikkeld die een compromis vormen tussen de bovengenoemde twee typen methoden. Het gaat namelijk om methoden die een weergave leveren voor de objecten (net als MCA) en tegelijkertijd een goede weergave van de variabelen geven (net als methoden voor PCA gebaseerd op gegeneraliseerde korrelatie-koëfficiënten). Omdat nominale variabelen vaak gezamenlijk met kwantitatieve

variabelen geanalyseerd moeten worden (we spreken dan van gemengde gegevens) zijn de bovenbeschreven compromis-methoden gegeneraliseerd zodat ze toepasbaar zijn voor het analyseren van dergelijke gemengde gegevens.

Het onderzoek bestaat uit drie delen. In het eerste deel wordt een skala van methoden beschreven voor het analyseren van nominale of gemengde gegevens. Deze methoden berusten op toepassingen van drie-weg methoden op zogenaamde kwantifikatie-matrices. In hoofdstuk 2 wordt een aantal drie-weg methoden beschreven. Het gaat hierbij om een speciaal type drie-weg methoden dat bedoeld is voor het analyseren van een aantal gelijkenissen-matrices die gelijkenissen geven tussen steeds dezelfde objecten, ten aanzien van verschillende aspecten. Voor de in dit onderzoek genoemde methoden wordt aangetoond dat ze onderling een hiërarchie vormen. De eerste methode is de meest algemene. De tweede methode kan opgevat worden als de eerste methode als hieraan bepaalde randvoorwaarden worden opgelegd. De derde methode is op te vatten als de tweede na toevoeging van bepaalde extra randvoorwaarden, etc.

Methoden voor het analyseren van nominale variabelen worden gekonstrueerd door drie-weg methoden toe te passen op kwantifikatie-matrices. Kwantifikatie-matrices zijn matrices die gelijkenissen tussen objecten geven zoals die kunnen worden bepaald op grond van een nominale variabele. Een heel eenvoudige vorm van zo'n kwantifikatie-matrix is die die de gelijkens tussen twee objecten die in dezelfde categorie van een (nominale) variabele vallen als 1 aangeeft, en de gelijkens tussen twee objecten die in verschillende categorieën van een variabele vallen als 0. In hoofdstuk 3 worden deze en meer gekompliceerde kwantifikatie-matrices behandeld. Naast kwantifikatie-matrices voor nominale variabelen worden kwantifikatie-matrices voor kwantitatieve variabelen besproken, en wordt aangegeven hoe kwantifikatie-matrices voor ordinale variabelen kunnen worden gekonstrueerd.

In het laatste hoofdstuk van het eerste deel, hoofdstuk 4, wordt aangetoond wat voor methoden er ontstaan als de in hoofdstuk 2 genoemde drie-weg methoden worden toegepast op de in hoofdstuk 3 genoemde kwantifikatie-matrices. Hier wordt met name aangetoond dat deze combinatie leidt tot een aantal bestaande methoden, waaronder diverse PCA technieken die gebaseerd zijn op gegeneraliseerde korrelatie-koëfficiënten, alsmede MCA. Het levert ook een groot aantal nieuwe methoden op die elk een compromis vormen

tussen gegeneraliseerde korrelatie–koefficienten methoden en varianten van MCA.

Nadat in het eerste deel van het onderzoek een breed scala van voornamelijk nieuwe technieken voor het analyseren van nominale en gemengde gegevens is beschreven, wordt in het tweede deel één techniek in het bijzonder behandeld. Het gaat om de toepassing van INDSCAL met orthonormaliteits–voorwaarde (INDORT genoemd) op de kwantifikatie–matrices die ook (impliciet) in MCA gebruikt worden. Deze methode wordt INDOQUAL genoemd. In hoofdstuk 5 wordt deze methode uitvoerig besproken. Er wordt aangetoond op welke manier deze methode een compromis vormt tussen één bepaalde gegeneraliseerde korrelatie–koefficienten methode en MCA. Het blijkt dat deze methode zeer goed kan worden opgevat als PCA van nominale variabelen onder de randvoorwaarde dat de componenten voor de variabelen direkt gerelateerd zijn aan componenten voor de objecten. Voorts wordt uitgelegd hoe de resultaten van een dergelijke analyse kunnen worden geïnterpreteerd.

In hoofdstuk 6 wordt INDOQUAL vergeleken met enkele andere methoden. Impliciet wordt hiermee een aantal nieuwe interpretaties van INDOQUAL gegeven.

INDOQUAL is alleen bruikbaar voor het analyseren van nominale variabelen. Gemengde variabelen kunnen worden geanalyseerd met behulp van de in hoofdstuk 7 beschreven generalisatie van INDOQUAL, die INDOMIX genoemd is. Er wordt beschreven hoe deze methode, net als INDOQUAL, te zien is als compromis tussen bestaande methoden. Voorts wordt beschreven hoe de resultaten van deze methode geïnterpreteerd kunnen worden. Tenslotte worden enkele speciale gevallen behandeld.

In hoofdstuk 8 worden relaties van INDOMIX (en dus impliciet van INDOQUAL) met simple structure rotatie–technieken beschreven. Dergelijke technieken dienen ertoe om bij gewone PCA de componenten zodanig te roteren dat de ladingen (korrelaties van componenten met variabelen) extreem zijn, dat wil zeggen, hetzij groot (dicht bij 1), hetzij klein (dicht bij 0). Er wordt aangetoond dat INDOMIX gezien kan worden als een methode die simple structure maximaliseert volgens het quartimax criterium. Deze methode verschilt hierin van simple structure rotatie–technieken voor PCA dat bij INDOMIX de simple structure niet alleen gemaximaliseerd wordt over mogelijke *rotaties* van componenten, maar over *alle* mogelijke componenten.

Daarnaast zijn, om een zinvolle vergelijking van INDOMIX met MCA mogelijk te maken, technieken ontwikkeld voor simple structure rotatie van MCA-oplossingen.

In hoofdstuk 9 worden algorithmen beschreven voor INDOQUAL en INDOMIX die speciaal ontworpen zijn om het mogelijk te maken gegevensbestanden met grote aantallen objecten te analyseren. Tevens worden in dit hoofdstuk nog enige technische aspecten van de oplossing van INDOQUAL en INDOMIX besproken.

Het laatste deel van het onderzoek betreft toepassingen en ervaringen met INDOQUAL en INDOMIX. Er worden zeven toepassingen beschreven, waarin steeds andere aspecten van de methoden worden belicht. Er wordt uitvoerig ingegaan op de interpretatie van INDOQUAL en INDOMIX oplossingen voor empirische gegevens. Bovendien wordt in twee gevallen de stabiliteit van de oplossingen onderzocht. In het eerste geval wordt hierbij gebruik gemaakt van de zogenaamde jackknife methode. Daarbij wordt steeds één ander object uit het gegevensbestand weggelaten, en worden de verkregen oplossingen voor deze deel-steekproeven onderling vergeleken om na te gaan hoeveel invloed het weglaten van één object heeft op de oplossing. In het tweede geval wordt de stabiliteit bestudeerd door middel van kruis-validatie. Hierbij wordt de totale steekproef in tweeën gesplitst, vervolgens worden de gewichten die resulteren uit de oplossing van de ene deel-steekproef toegepast op de andere, en tenslotte wordt de aldus verkregen weergave voor de objecten uit de tweede deel-steekproef vergeleken met de weergave van de objecten die gevonden wordt door een gewone INDOMIX analyse van deze deel-steekproef. Door middel van deze procedure wordt nagegaan in welke mate de oplossingen afhangen van specifieke kenmerken van elk van de deel-steekproeven. In de twee voorbeelden waar de stabiliteit is bestudeerd bleek deze zeer groot te zijn.

AUTHOR INDEX

Bäckström	113, 129–130, 135, 164
Baumerder	22, 41, 157
Benzécri	2, 58, 157
Bonnefous	22, 41, 157
Carroll, J.B.	74, 157
Carroll, J.D.	9, 13, 35, 54, 100, 109, 157
Cazes	22, 28, 41, 50, 106, 157
Chang	13, 54, 109, 157
Clarkson	73, 80, 157
Coppi	3, 22, 157
Crawford	73, 158
Cuadras	34, 70, 158
D'Alessio	50, 158
D'Ambra	3, 22, 30, 35, 158, 161
Daniels	20, 26, 158
De Leeuw	11, 33, 35, 52, 61–62, 68, 75, 78, 115, 158, 161, 165
De Soete	100, 109, 157
Di Ciaccio	35, 158
Domenges	35, 159
Eckart	58, 159
Escofier	33, 35, 55–56, 62, 159
Escoufier	21–22, 29–31, 43–44, 51, 159
Ferguson	73–74, 158–159
Fichet	31, 159
Gbegan	31, 159
Gifi	2, 108, 113, 135–136, 159
Gower	14, 26, 31, 33–34, 58, 70, 140–141, 159–160
Grasse	119, 121, 160
Green	144, 160
Greenacre	35, 106, 160
Guttman	2, 160

Harman	71, 160
Hartigan	90, 160
Hayashi	2, 160
Heiser	58, 94, 152–153, 160, 164
Horan	54–57, 160
Hubert	23, 160
Jaffrennou	14, 160
Janson	21, 23, 25, 28–29, 32, 35, 47, 70, 160–161, 165
Jennrich	73, 80, 157, 161
Jones	94, 161
Kaiser	71, 73–75, 79, 82, 155, 161
Kendall	21, 26
Kiers	12, 16, 30–31, 33–34, 41, 43–44, 61, 75, 83, 95–96, 102, 161, 163
Knol	43, 75, 95–96, 163
Kroonenberg	9, 11–12, 14, 95, 109, 161
Lauro	35, 161
Lebart	58, 162
Levin	14, 162
L'Hermier des Plantes	10, 162
Marchetti	3, 22, 30–31, 55, 109, 114, 148–152, 158, 162
Marcotorchino	22, 162
Messick	11, 163
Meulman	35, 58, 60, 108, 115, 117–119, 122, 135, 158, 160, 162
Miller	115, 162
Morineau	58, 162
Muthén	35, 162
Neuhaus	74, 162
Nishisato	2, 28, 33, 35, 62, 162
Pagès, J.	55–56, 159
Pagès, J.P.	22, 41, 157
Pruzansky	75, 78, 100, 109, 157–158
Sabatier	35, 162
Saporta	3, 21–22, 25, 28, 30, 32, 41–43, 48–49, 63, 70, 162

Saunders	74, 163
Sibson	94, 161
Spearman	21, 26
Sugiyama	114, 152, 163
Tabard	58, 162
Takane	35, 52, 61, 165
Ten Berge	28, 43, 75, 77–78, 82–83, 86, 95–100, 108, 163, 165
Tenenhaus	2, 61, 88, 163
Ter Braak	35, 163
Torgerson	13–14, 58, 163
Tschuprow	23, 25, 28–31, 43–44, 129, 163
Tucker	10–14, 17, 55, 109, 144, 163
Van Buuren	94, 164
Van den Burg	113, 142–144, 164
Van der Burg	35, 87, 119, 164
Van der Heijden	35, 164
Van Rijkevorsel	33, 35, 107, 158, 164
Van Zomeren	113, 142–144, 164
Vegelius	21–23, 25, 28–29, 32–35, 47, 70, 113, 129–130, 135, 160–161, 164–165
Vescia	117, 119, 164
Volle	35, 159
Wish	9, 35, 157
Wrigley	74, 162
Yanai	35, 164
Young, F.W.	2, 35, 52, 61, 88, 163, 165
Young, G.	58, 159
Zegers	20, 22, 26–28, 165

SUBJECT INDEX

A

- adequate (representation) 18, 38, 41, 49, 53, 67
- AFM 55–56
- aggregate 5, 96, 102–103
- algorithm 4–5, 43, 75–76, 78, 82–83, 95–104, 108–109
- approximate 11, 53–56, 58–60, 66
- association coefficient 1, 20, 33
- asymmetrical
 - ...three-way analysis 50
 - ...treatment of variables 35

B

- binary variables 31, 68, 113–114, 136–138, 142, 146, 152
 - (see also dichotomous)
- bivariate frequencies 5, 101–102, 106, 109
- block-diagonal 105
- Burt-matrix 95–96, 101–102, 104–106, 109

C

- category centroids 69, 102, 105, 109, 115, 133
- Cauchy-Schwarz 85
- center(ing) 21, 25, 30, 46, 88
- centroids of object 45, 65
 - coordinates
- cetacea data 117–120, 122, 126, 129, 141
- chance (correcting for...) 24
- classification
 - forced... 28, 35
 - ...of cetacea 113, 119, 121
- cluster(ing) 71, 83, 87, 90–92, 94, 113–114, 117, 119, 121–122, 129, 147
- column-scaling 59
- column-space 24, 55, 81, 103
- column-wise
 - ...centering 25, 30, 88
 - ...orthonormal 15, 49, 55, 57, 59, 100
- comparing (comparison of)
 - ... IAF_I and IAF_S 67
 - ...INDOQUAL and MCA 4, 44, 46, 51, 54–55, 57, 60, 87, 118, 120
 - ...INDOQUAL and PCA of quantification matrices 43–44
 - ...INDOQUAL and other methods 49, 94, 148, 151
 - ...qualitative and quantitative variables 20

...simple structure	
criteria	84
...solutions	115, 126, 141
component-weights	104
compromise	3, 10–11, 35, 42, 44, 46, 51, 61–63, 65, 69, 81, 89, 155
computation(al)	95–96, 98–99, 102, 104–105, 108–109, 116, 142
computer-efficiency	105
congruence (coefficient of...)	144
contingency table	5, 30, 58, 95, 101, 158
converge(nce)	82–83, 95, 97–98, 100, 102–103, 105
copies	52
core matrix	11–12, 14, 152
correlation coefficient	20–23, 26, 28, 32, 64, 70, 129
correlation-matrix	34
correspondence analysis	2, 31, 41, 58, 61, 84, 87, 95–96, 106, 113
CP-indices	32–33
cross-classification	4, 29–30, 32–35
CROSSMIN	74–76, 79–80, 83
cross-validation	104, 113–114, 116–117, 140–141

D

deviation scores	20, 26
diagonalization	75, 77–79
dichotomous variables	44, 67–69, 142
(see also binary)	
dimensionality	103, 130–131, 135, 138, 142, 146, 149, 154
discriminant analysis	34, 94
discriminate	47, 88–90
discrimination measure	47, 52, 72–73
discriminatory capability	87, 90, 92, 94
dissimilarity	13–14, 26
distance	11, 13, 122
χ^2 -...	57–60
distributional equivalence	96, 105–106

E

eclectic	17
E-coefficients	21
E-correlation	69
eigendecomposition	57–58, 85, 97
eigenvalue	11, 46, 48–49, 51, 58, 66–67, 77–78, 86, 129
eigenvector	28, 46, 54–55, 57–58, 68, 70, 72, 77, 86
empirical	5, 18, 36–37, 90
eta squared (η^2)	34, 47, 63–66, 69–70, 73
Euclidean	21
exploratory analysis	1–2, 61

F

fictitious data	144–145
-----------------	---------

Figure	9, 92–93, 119–121, 126, 147–148, 151
frequencies	5, 21, 24, 59, 68–69, 95, 101–102, 104, 106, 109
fuzzy coding	106–107
G	
generalization	2, 33, 62, 81, 83–84
group	88–89, 119, 130, 141, 147
GROUPALS	94
H	
half (see also split)	114, 116–117, 140–141
hierarchical	9, 16, 31, 43, 65, 119
hierarchy	3, 9, 16–18, 31, 34–35, 38, 155
HOMANA–BIN	153
Homogeneity	119, 154
I	
IAF _I	66–67, 138
IAF _S	67, 138
IDIOSCAL	9, 109
indicator matrix	21, 58, 62, 68, 108
indicator variable	2, 21, 59–60, 62, 71
INDOMIX	4–5, 33–34, 61, 63–69, 71, 81–85, 87, 95–96, 100, 102–104, 107–109, 113–117, 135–142, 154–155
INDOQUAL	4–5, 30–31, 41–51, 53–57, 59–60, 71, 87–96, 102–109, 113–115, 117, 119–135, 142–144, 146–153, 155
INDORT	4, 13–14, 16, 30–31, 34, 41–46, 50, 54–57, 61–64, 66–70, 82, 95–96, 98, 100, 103–104, 108–109, 115, 138, 148
INDSCAL	4, 9, 13–14, 16–17, 30, 34, 41, 54–57, 60, 63, 69, 100, 109
inertia	4, 44, 47–48, 51, 65–67, 72–73, 80, 83, 87, 91, 93–94, 119–120, 123, 130, 138, 149–150, 152
inertia accounted for information	4, 48, 65–67, 93, 119–120, 123, 130, 149 2, 19, 36–37, 41, 49, 52–53, 61, 95–96, 102–103, 107, 116, 120, 130, 132, 150, 152
informative(ness)	37, 147
interaction	36
interaction–relations	12
interpretation	13, 17–18, 37, 47, 49, 51–53, 55, 58, 65, 69, 71, 73, 81, 88, 94, 123, 126, 132, 135, 138, 144, 150–152, 155
interval level	19–20
iteration	98–100, 105
iterative	95, 98, 116

J

J-index	23
J-indices	29-30
jackknife	115-116, 126-128, 141

K

K-means	94
Kristof's theorem	163
Kronecker	
...delta	45, 64
...product	12
KURTMAX	74-76, 79-80, 83-85, 88
kurtosis	74

L

least squares	12-13, 54-55, 60, 75, 78
level (of measurement)	19-22
Likert scale	136
limitations	34-35, 155
list-wise deletion	107
loss function	10, 12, 14, 42, 54-55

M

Mahalanobis distance	58
matching	94, 115, 144
MCA	2-4, 28, 30-31, 33, 35-36, 41-42, 44, 46-60, 62, 68, 71-72, 87-96, 103, 106, 108, 113, 117-123, 126-132, 142-144, 149-150, 152-155
missing data	96, 107-108, 117, 136
mixed variables (mixtures)	2-4, 17, 20-22, 26, 32-36, 37-38, 61-65, 67, 71, 81, 83, 93-94, 103, 113, 135
model	3, 9, 11-14, 16-17, 38, 49, 54-57, 66-67, 138, 150, 152
multidimensional scaling	11, 115

N

nested	60, 130
nominal variables	19, 22, 113-114, 129, 132, 137-139, 144, 146, 149-150, 152
normalize	20-21, 25-31, 34, 36-37, 43-44, 55, 63, 68, 77, 146
numerical variables	1, 19, 27, 140, 142

O

operators	21
ordinal variables	19-21, 26-27, 113, 136, 142

ORMAX	73–74, 76, 84–87
orthogonal	24, 49, 51, 73–76, 78–84, 144
orthomax	73, 75–80, 82–84
orthonormal	4, 14–15, 41, 49, 55–57, 59, 63, 66, 72, 75–80, 100, 103
OVERMAX	74–76, 79–80, 83–85, 88
P	
PARAFAC2	9
parsimony	49
passive variables	138
PCAMIX	2–4, 33–36, 62–81, 83–87, 94, 103, 135, 137–138, 154
phi squared (φ^2)	30–31, 34, 42–46, 51, 63, 144, 153
plot	92, 120–121, 150–151
point–biserial correlation	69, 142
polytomized	53, 149
practice	2, 19, 23, 36, 46–47, 54, 57, 69, 72, 82–83, 87, 90, 104, 114, 153
PRINCALS	33, 52, 135
PRINCIPALS	52–53, 61
product–moment correlation	30, 34, 44, 64–65, 72, 102–103, 138
program	52, 69, 94, 106, 135–136, 144
projection	44, 55–56
Projection Pursuit	94
pseudo–indicator matrix	107
Q	
QMAX	74, 76, 79–80, 83–85, 88
quantification matrix (–ces)	1–4, 9–10, 15, 17, 19–37, 41–44, 46, 48–49, 51, 53–55, 57, 62–64, 66–70, 81, 94–96, 100, 107–109, 113–114, 148–150, 155
quantified (variables)	37, 51–52, 73
quartimax	74–75, 79–82, 84, 91–92, 103, 143–144, 155
R	
rank,	
...correlation	21, 26
...of a matrix	103
...order	19, 27
rotation (rotating)	4, 11, 13, 50, 57, 71–81, 83–85, 87, 91, 93–94, 103, 120, 123, 126, 129, 132, 138, 142–144, 150–152, 154
oblique...	80
Procrustes...	144
RV–coefficients	21
S	
sample	114–116, 136, 140–141

...size	96, 104, 106
scalar product	20–21
similarity (similarities)	1, 3–4, 13, 15, 20, 23–28, 33–34
similarity matrix (-ces)	3, 9, 13–15, 20, 23, 28
simple structure	4, 11, 71–76, 80–84, 87–91, 93–94, 120, 129, 132, 150, 154
skew-symmetric matrices	27
SP-indices	33
split (...-half)	116, 140–141
stability (stable)	5, 113–116, 126, 129, 140, 154
start (...-ing configuration)	99–100, 109, 116, 153
STATIS	3, 10–12, 15–18, 29, 38
statistically independent	24–25
sub-objects	106–108
subsets of variables	49, 71, 88–90, 152–153
SUMPCA	3, 14–18, 31, 33, 46, 64, 66–67, 138
supplementary	
...objects	104
...variables	138
symmetric (matrices)	9, 11, 27–28, 78
T	
T-index	23, 30
Table	16–17, 22–23, 29–34, 43–44, 91–92, 101, 118, 122–134, 137–139, 142–143, 145–146, 150, 153
three-mode	
...principal component an. 11	
...scaling	10, 12, 109
three-way	
data	9, 19
method	3–4, 9, 16–17, 19, 29, 31–32, 34–37, 41–43, 109
trivial axis	46, 104
TUCKALS	11–14, 16–17, 30–31, 34, 55, 69, 114, 149–152
U	
uncorrelated	117, 140–141
unique (axes)	13, 50, 54, 57, 64, 115
V	
variance	27, 52, 68, 80, 83, 88–89
varimax	74–75, 79–83, 87, 91, 92, 120, 122–123, 126–129, 131–132, 138, 142–143, 154–155
W	
weight (weighting)	
...for objects	96, 106–108
...for variables	10, 15–16, 27–28, 35–37, 48–49, 113, 137–138, 146, 149

NOTATION

In this study it is attempted to use a uniform notation for most of the symbols. In general, lower case *italic* symbols denote integer numbers, lower case greek symbols denote real numbers, lower case **bold** symbols denote vectors, and upper case *italic* (or greek) symbols denote matrices. The elements of vectors and matrices are denoted by the same characters, but in lower case *italics*, with indices to denote their position in the vectors or matrices. Finally, some short-cuts for summation signs are described. It should be noted that chapter 9 has a partly deviating notation.

g	index for categories of variable j
h	index for categories of variable k
f_g	frequency of category g
i	index for objects, $i = 1, \dots, n$.
j	index for variables, $j = 1, \dots, m$.
k	index for variables, $k = 1, \dots, m$.
l	index for components (dimensions), $l = 1, \dots, r$
m	number of variables
m_j	number of categories of qualitative variable j
n	number of objects
p	number of object-components in TUCKALS-3
r	number of components (dimensionality)
α_j	weight for variable j
$\delta_{ll'}$	Kronecker δ ; $\delta_{ll'} = 1$ if $l = l'$; $\delta_{ll'} = 0$ if $l \neq l'$
λ_j	j^{th} diagonal element of Λ , where Λ is a diagonal matrix
\mathbf{c}	m -vector with loadings for variables in SUMPCA
\mathbf{h}_j	n -vector of raw scores on variable j

\mathbf{s}_j	n^2 -vector with elements of S_j strung-out row-wise, $\text{Vec}(S_j)$
\mathbf{w}_j	vector with diagonal elements of W_j
\mathbf{x}_l	n -vector of scores on component l , normalized to unit sum of squares
\mathbf{z}_j	n -vector of standardized scores on variable j
$\mathbf{1}$	n -vector with unit elements only
A	$m \times r$ matrix of (nonsquared) loadings of variables on components
C	$m \times r$ matrix of (squared) loadings of variables on components
D_j	$r \times r$ diagonal matrix with category frequencies of variable j
E_j	$m \times m$ symmetric matrix which is diagonalized by ordinary orthomax
\tilde{E}_j	$m \times m$ symmetric matrix which is diagonalized by orthomax for PCAMIX
F	matrix of component scores ($n^2 \times r$ in chapter 2, $n \times r$ in chapter 8)
F_l	$n \times n$ matrix expressing "component" l for matrices S_1, \dots, S_m
G_j	$n \times m_j$ indicator matrix for variable j
H	$p^2 \times r$ matrix containing $\text{Vec}(H_1), \dots, \text{Vec}(H_r)$
H_l	$p \times p$ matrix containing the l^{th} frontal plane of the TUCKALS-3 core
I_r	$r \times r$ identity matrix
J	$n \times n$ centering operator, given by $(I_n - n^{-1}\mathbf{1}\mathbf{1}')$
K	$n \times n$ (orthonormal) matrix of eigenvectors of $\sum_j S_j$
K_r	$n \times r$ matrix with first r eigenvectors of $\sum_j S_j$
P_j	$n \times n$ quantification matrix $JG_jD_j^{-1}G_j'J$ for qualitative variable j
Q_j	$n \times n$ quantification matrix $n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variable j
S_j	$n \times n$ quantification matrix for variable j
S	$n^2 \times m$ matrix containing vectors $\mathbf{s}_1, \dots, \mathbf{s}_m$
T	$r \times r$ (orthonormal) rotation matrix
W	$r \times r$ diagonal matrix
W_j	$r \times r$ diagonal matrix with loadings for variable j on the diagonal

X	matrix of object coordinates (mostly $n \times r$, and $X'X = I_r$)
Y_j	$m_j \times r$ matrix with category means of object coordinates
Z	$n \times m$ matrix with standardized scores of objects on variables
Λ	diagonal matrix (often with eigenvalues of $\sum_j S_j$)
Λ_r	$r \times r$ diagonal matrix with first r eigenvalues of $\sum_j S_j$
\sum_j	summation over $j = 1, \dots, m$
\sum_l	summation over $l = 1, \dots, r$
$\text{Vec}(\cdot)$	column-vector with the elements of a matrix strung-out row-wise
\otimes	Kronecker product

Apart from this list of symbols it seems useful to provide a list of abbreviations (or acronyms) of methods that are mentioned frequently in this dissertation. The names are given with, between brackets, the number of the page where they are introduced in this study and, a short description of their objectives.

INDOMIX (p. 63)

INDORT applied to a set of quantification matrices; for qualitative variables $S_j = JG_jD_j^{-1}G_j'J$, for quantitative variables $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$

INDOQUAL (p. 41)

INDORT applied to a set of quantification matrices given by $S_j = JG_jD_j^{-1}G_j'J$, $j = 1, \dots, m$

INDORT (p. 14)

INDscal with ORThonormality constraint on object coordinates
 Minimizes $\sum_j \|S_j - XW_jX'\|^2$ over X , subject to $X'X = I_r$, and over diagonal W_j
 Maximizes $\sum_j \sum_l (\mathbf{x}_l'S_j\mathbf{x}_l)^2$ over X , subject to $X'X = I_r$

INDSCAL (p. 13)

INDividual differences SCALing

Minimizes $\sum_j \| S_j - XW_jX' \|^2$ over arbitrary X ($n \times r$) and diagonal W_j
($r \times r$)

MCA (p. 2)

Multiple Correspondence Analysis

Maximizes $\text{tr } X'J\sum_j G_j D_j^{-1} G_j' JX$ over X , subject to $X'X = I_r$

PCA

Principal Components Analysis

Maximizes $\text{tr } X'\sum_j \mathbf{z}_j \mathbf{z}_j' X$ over X , subject to $X'X = I_r$

PCAMIX (p. 62)

Generalization of MCA and PCA, for the analysis of mixed variables

Maximizes $\text{tr } X'S_j X$ over X , subject to $X'X = I_r$, where $S_j = JG_j D_j^{-1} G_j' J$
for qualitative variables and $S_j = n^{-1} \mathbf{z}_j \mathbf{z}_j'$ for quantitative
variables

PCA of quantification matrices (p. 1)

PCA of “correlation”-coefficients for qualitative variables

STATIS-1 applied to matrices S_j

Maximizes sum of squared loadings (“correlations” between variables and
components)

PCA of ϕ^2 -coefficients (p. 30)

STATIS-1 applied to matrices $S_j = JG_j D_j^{-1} G_j' J$

Maximizes sum of squared loadings (η^2 -coefficients between variables
and components)

PCA of η^2 -coefficients (p. 63)

STATIS-1 applied to matrices $S_j = JG_jD_j^{-1}G_j'J$ for qualitative variables
and $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variables
Maximizes sum of squared loadings (η^2 -coefficients or squared
product-moment correlations between variables and components)

STATIS-1 (p. 10)

First step of STATIS method
PCA of $\text{Vec}(S_1), \dots, \text{Vec}(S_m)$
Minimizes $\sum_j \|S_j - \sum_l c_{jl}F_l\|^2$ over F_1, \dots, F_r and C

SUMPCA (p. 15)

Minimizes $\|\sum_j S_j - XAX'\|^2$ over diagonal matrices A , and X , subject to
 $X'X = I_r$
Maximizes $\text{tr } X'\sum_j S_j X$ over X , subject to $X'X = I_r$

TUCKALS-3 (p. 11)

Method for least-squares fitting of Tucker's three-mode component
analysis model
Minimizes $\sum_j \|S_j - X\sum_l c_{jl}H_lX'\|^2$ over arbitrary X ($n \times p$), C ($m \times r$),
and H_l , $l = 1, \dots, r$

THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

A problem often occurring in exploratory data analysis is how to summarize large numbers of variables in terms of a smaller number of dimensions. When the variables are quantitative, one may resort to Principal Components Analysis (PCA). When qualitative (categorical) variables are involved, one may choose from a variety of methods which are adaptations of PCA, designed for this purpose.

This book is about such adapted PCA methods for qualitative variables, and for mixtures of qualitative and quantitative variables.

Part I treats adapted PCA methods, like Multiple Correspondence Analysis (MCA), in retrospect. The methods are systematized such that the choice between the different methods is facilitated. In order to fill certain gaps in this system a number of new methods is developed. One of these new methods is discussed in detail in Part II, which contains various innovatory contributions. One of the main new developments is the construction of techniques for rotating the MCA solution to what is called simple structure. These techniques are generalizations of the well-known VARIMAX and QUARTIMAX techniques for rotating PCA solutions.

The former rotation techniques are implicitly constrained in the sense that the new components are rotations of the MCA solution. In order to optimize VARIMAX or QUARTIMAX from the start a different technique is offered in which this constraint is dropped. This new technique, which is based on INDSCAL, finds solutions in which clusters of variables are identified more clearly than is done in MCA, even after simple structure rotation. An important additional consequence of this is that the observation units (objects, individuals) are also represented more clearly in clusters than is the case with MCA, which implies that the method is useful as a cluster technique.

Finally, Part III shows how certain selected methods work in practice. For this purpose seven example data sets have been analyzed. The results are discussed in detail, including some stability analyses. Particular attention is paid to the clustering of variables and objects.

DSWO PRESS

ISBN 90 6695 037 4