Nonparametric inference in nonlinear principal components analysis: Exploration and beyond

Linting, M.

Summary in Dutch (Samenvatting)

Samenvatting

Dit proefschrift laat zien dat niet-lineaire principale componentenanalyse als analysemethode meer te bieden heeft dan zijn exploratieve status doet vermoeden. Principale componentenanalyse (PCA) wordt gebruikt in exploratief onderzoek om structuur te herkennen in datasets. De methode reduceert een (groot) aantal variabelen tot een aantal onderliggende variabelen, de zogenaamde principale componenten, die de informatie in de geobserveerde variabelen zo goed mogelijk weergeven. De principale componenten geven dus eigenlijk een samenvatting van de geobserveerde variabelen die bijvoorbeeld gebruikt kan worden om subgroepjes in variabelen te onderscheiden, zoals vaak gedaan wordt bij schaalconstructie. De effectiviteit van deze samenvatting wordt weergegeven door de verklaarde variantie die per principale component wordt uitgedrukt in een zogenaamde eigenwaarde. De correlatie tussen de variabelen en een principale component wordt weergegeven als een componentlading. Een gekwadrateerde componentlading geeft de proportie van de variantie in een variabele die verklaard wordt door een principale component.

PCA heeft twee belangrijke beperkingen: (1) de variabelen moeten lineair aan elkaar gerelateerd zijn, en (2) de interpretatie is alleen zinvol als de variabelen gemeten zijn op numeriek (interval- of ratio-) niveau. In de sociale en gedragwetenschappen wordt vaak niet aan deze voorwaarden voldaan. In dat kader is een niet-linear alternatief voor PCA ontwikkeld dat niet-lineaire of categorische PCA (NLPCA of CATPCA) genoemd wordt. Deze methode kan structuur ontdekken in datasets die variabelen bevatten met verschillende meetniveaus (nominaal, ordinaal, interval en/of ratio) en waartussen mogelijk niet-lineaire verbanden bestaan.

In dit proefschrift wordt eerst een praktische handleiding voor NLPCA gegeven, waarbij de overeenkomsten en verschillen met lineaire PCA aan bod komen en waarin een uitgebreide toepassing van de methode op empirische data getoond wordt. Vervolgens worden verschillende manieren beschreven om inferentiële maten te bepalen binnen de voornamelijk exploratieve context van PCA. Ten eerste wordt de nonparametrische bootstrap toegepast om de stabiliteit (robuustheid) van de oplossing te bepalen. Ten tweede worden permutatietoetsen gebruikt om de significantie van de lineaire PCA-oplossing te bepalen, waarbij twee verschillende strategieën worden toegepast. De meest effectieve van deze permutatiestrategieën wordt vervolgens gebruikt om de significantie van de bijdrage van variabelen aan de NLPCA-oplossing vast te stellen. Uiteindelijk wordt de relatie tussen de resultaten van de bootstrap en permutatietoetsen toegelicht.

NLPCA als exploratieve methode

Hoofdstuk 2 van dit proefschrift is een didactische handleiding voor NLPCA die een uitgebreide en gebalanceerde beschrijving van de methode biedt, gevolgd door een stap voor stap beschreven toepassing op een empirische dataset.

NLPCA vertoont veel overeenkomsten met lineaire PCA, zowel in datatheoretisch opzicht, als wat betreft uitvoer en interpretatie. Als alle variabelen in NLPCA als numeriek worden behandeld is de uitkomst zelfs exact gelijk aan een lineaire PCA-oplossing. Echter, NLPCA is te verkiezen boven PCA wanneer: (1) de te analyseren dataset variabelen bevat die niet numeriek (categorisch) zijn, en/of (2) de variabelen in de dataset (mogelijk) niet-lineair aan elkaar gerelateerd zijn. Als variabelen inderdaad niet-lineaire relaties vertonen, zal NLPCA meer variantie in de data kunnen verklaren en kan de interpretatie van de oplossing verbeterd worden ten opzichte van de lineaire PCA-oplossing. Bovendien kan NLPCA-software inzichtelijke uitvoer opleveren, ook als alle variabelen numeriek en lineair gerelateerd zijn. Het programma CATPCA (onderdeel van de SPSS-module Categories) (Meulman et al., 2004) levert bijvoorbeeld grafische weergaven van de componentladingen en componentscores (objectscores) afzonderlijk, zowel als zogenaamde biplots en triplots, waarin ladingen, scores en eventueel categoriepunten gezamenlijk worden afgebeeld in een diagram met de principale componenten als assen.[2]

Het voornaamste verschil tussen lineaire en niet-lineaire PCA is dat NLPCA de categorielabels van nominale en ordinale variabelen kan omzetten in waarden waarmee gerekend kan worden (kwantificeren) op een zodanige wijze dat de relatie tussen de gekwantificeerde variabelen geoptimaliseerd wordt (d.w.z. de verklaarde variantie in de gekwantificeerde variabelen wordt zo hoog mogelijk gemaakt). Ook variabelen die op numeriek niveau zijn gemeten kunnen niet-lineair getransformeerd worden, als de onderzoeker vermoedt dat deze variabelen niet-lineaire relaties hebben met andere variabelen in de dataset. De onderzoeker kan voor elke variabele zelf een analyseniveau specificeren. Bij het toepassen van NLPCA is het nuttig om verschillende analyseniveaus voor de variabelen uit te proberen, bij voorkeur uitgaand van het nominale niveau (het minst restrictieve niveau) en oplopend via het ordinale niveau naar het numerieke niveau (het meest restrictieve niveau). De resulterende oplossingen kunnen dan met elkaar vergeleken worden wat betreft fitmaten en de vorm van de transformaties. Als de oplossing met numerieke analyseniveaus niet veel verschilt van de oplossing met ordinale of nominale analyseniveaus, is niet-lineaire transformatie niet nodig. Als de verschillen aanzienlijk zijn kan niet-lineare transformatie tot nieuwe inzichten leiden. Er kan gekozen worden

[2]De Universiteit Leiden beschikt over het copyright van de procedures in het SPSS Categories pakket en de Datatheorie Groep ontvangt de royalties.

voor een mix van analyseniveaus, waarbij elke variabele het best passende analyseniveau wordt toegekend. Het is belangrijk om bij het vergelijken van de oplossingen zowel aandacht te besteden aan de fit – bijvoorbeeld de verklaarde variantie of andere fitmaten, zoals de multivariate voorspelbaarheid van de data uit de NLPCA-oplossing (Gower & Blasius, 2005) – als aan de kwaliteit van de interpretatie.

In het algemeen blijkt NLPCA een zeer waardevolle analysemethode die de relaties tussen variabelen op een inzichtelijke manier visueel kan weergeven. NLPCA is een nuttig alternatief voor lineaire PCA dat om kan gaan met meer verschillende soorten variabelen, zonder van tevoren aannamen te doen over de verdeling van en relaties tussen deze variabelen in de populatie.

Stabiliteit van NLPCA

NLPCA wordt gezien als een exploratieve methode en levert in dat kader geen inferentiële maten. In de praktijk hoeft het onderscheid tussen exploratieve en confirmatieve methoden niet zo strikt gehanteerd te worden als soms gesuggereerd wordt (zie ook De Leeuw, 1988). Elke confirmatieve methode bevat immers beschrijvende elementen en een exploratieve methode kan een handvat bieden voor inferentie. Een aspect van inferentie is de stabiliteit van een bepaalde analysetechniek. Stabiliteit wordt hier gedefinieerd als de mate waarin de oplossing gelijk blijft, als de analyse herhaald wordt voor een nieuwe steekproef.

In Hoofdstuk 3 van dit proefschrift wordt de stabiliteit van zowel de lineaire als de niet-lineaire PCA-oplossing onderzocht met behulp van de nonparametrische bootstrap procedure. Deze procedure omvat het aselect en met teruglegging trekken van een groot aantal (bijvoorbeeld 1000) nieuwe steekproeven uit de geobserveerde steekproef. Elke nieuwe steekproef (i.e., bootstrapsteekproef) is van dezelfde grootte als de geobserveerde, maar sommige personen uit de geobserveerde steekproef komen niet voor en andere komen meerdere keren voor. Elke bootstrapsteekproef wordt aan dezelfde analyse onderworpen als de geobserveerde steekproef en de waarden waarin de onderzoeker geïnteresseerd is worden bepaald. De gezamenlijke resultaten voor de bootstrapsteekproeven worden gebruikt om betrouwbaarheidsintervallen te bepalen. Binnen de beschreven studie naar stabiliteit van lineaire en niet-lineaire PCA wordt aandacht besteed aan betrouwbaarheidsintervallen voor de eigenwaarden, de componentladingen, de objectscores en (voor NLPCA) de categoriekwantificaties.

Na afloop van de bootstrapstudie konden we concluderen dat de NLPCA-oplossing even stabiel kan zijn als de lineaire PCA-oplossing, ondanks het feit dat het aantal waarden dat geschat moet worden in NLPCA veel groter is

(aangezien voor elke afzonderlijke categorie een kwantificatie wordt berekend). Echter, categorieën met kleine marginale frequenties vertonen veel instabiliteit in de kwantificaties, wat vervolgens instabiliteit in de maten voor de corresponderende variabelen en objecten veroorzaakt. Dit probleem kan verklaard worden doordat personen die in zeldzame categorieën scoren niet voorkomen in sommige bootstrapsteekproeven en meerdere keren in andere bootstrapsteekproeven, waardoor de oplossing de ene keer heel anders kan zijn dan de andere keer. NLPCA is hiervoor gevoeliger dan lineaire PCA, aangezien de mogelijkheid tot niet-lineaire kwantificatie de techniek meer vrijheid verschaft. Het samenvoegen van categorieën met kleine marginale frequenties biedt een oplossing voor dit probleem.

Ondanks het feit dat de resultaten van Hoofdstuk 3 gebaseerd zijn op een specifieke dataset, zijn er veel overeenkomsten met eerder onderzoek wat betreft de stabiliteit van andere multivariate categorische analysetechnieken, zoals homogeniteitsanalyse of multipele correspondentie-analyse (zie Markus, 1994). Hiervan uitgaande, kunnen voorzichtig de volgende algemene adviezen geformuleerd worden:

1. In het algemeen (zowel in NLPCA als andere methoden) zullen kleinere steekproeven tot minder nauwkeurige resultaten leiden dan grotere steekproeven (mits het analyseniveau en het aantal categorieën van de geobserveerde variabelen gelijk blijft). Betrouwbaarheidsintervallen zullen breed zijn voor kleine steekproeven, aangezien er voor kleine steekproeven veel variatie in oplossingen is. Dus, ook voor kleine steekproeven geeft de bootstrap een correcte indicatie van de stabiliteit van de oplossing, ondanks het feit dat de analyseresultaten voor deze steekproeven veel kunnen verschillen van de corresponderende populatiekenmerken (Markus, 1994). De resultaten uit dit proefschrift en de aanbevelingen van Markus wijzen erop dat de oplossingen voor datasets met een paar honderd objecten redelijk stabiel zijn.

2. Trek een groot aantal bootstrapsteekproeven. Zowel volgens dit proefschrift als volgens Markus is een aantal van 1000 of meer afdoende. Wanneer gebruik gemaakt wordt van moderne computers zal het uitvoeren van een bootstrapstudie met 1000 steekproeven niet veel tijd kosten: De code die gebruikt is in dit proefschrift werkt met een SPSS-macrofile in combinatie met CATPCA en kost ongeveer een half uur rekentijd voor een dataset met ongeveer 600 objecten (Pentium 4, 3.00 GHz).

3. Gebruik een orthogonale Procrustes procedure om de componentladingen uit een bootstrapanalyse in de richting van de ladingen uit de oorspronkelijke analyse te roteren. Als dit achterwege wordt gelaten, kan

de stabiliteit van de oplossing sterk onderschat worden, vooral als de eigenwaarden van de principale componenten ongeveer even groot zijn.

4. Voeg categorieën met kleine marginale frequenties samen. Zulke categorieën kunnen zorgen voor instabiliteit in de NLPCA-oplossing. Het aantal benodigde observaties binnen een categorie kan verschillen per dataset: in grotere datasets zal de absolute minimale frequentie per categorie hoger moeten zijn dan in kleinere datasets. In Hoofdstuk 3 stellen we een minimum van 2.5% van het totaal aantal observaties, in overeenstemming met het feit dat in een normaalverdeling de 2.5% observaties in de staarten van de verdeling als extreem worden beschouwd. In het bijzonder voor kleine datasets is het beperken van het aantal categorieën per variable raadzaam.

5. Behalve aan de bootstrapellipsen, die een grafische weergave van betrouwbaarheidsintervallen bieden, moet ook aandacht besteed worden aan de verdeling van de bootstrappunten binnen die ellipsen. Deze verdelingen kunnen flink afwijken van een normaalverdeling en daardoor is het mogelijk dat een ellips niet de beste representatie geeft van de spreiding in de bootstrapresultaten.

Statistische significantie van de bijdrage van de variabelen aan de lineaire PCA-oplossing

Behalve de stabiliteit van een analysemethode kan ook de statistische significantie van elementen van de oplossing van belang zijn. Om statistische significantie te bepalen op een nonparametrische manier kunnen permutatietoetsen worden toegepast. Een permutatietoets wordt uitgevoerd door een groot aantal (bijvoorbeeld 999; i.e., 1000, inclusief de geobserveerde dataset) nieuwe datasets te genereren door per variabele de volgorde van de observaties aselect te permuteren. Hierdoor wordt de correlatiestructuur van de dataset verwoest en ontstaan datasets zonder specifieke structuur, maar wel met dezelfde univariate verdelingen als in de geobserveerde dataset. De resultaten van de gezamenlijke gepermuteerde datasets worden gebruikt om nulverdelingen te construeren voor de waarden waarin de onderzoeker geïnteresseerd is. Als een geobserveerde waarde voldoende extreem is ten opzichte van zijn nulverdeling, wordt deze aangemerkt als significant. Om te bepalen of de geobserveerde waarde voldoende extreem is worden p-waarden berekend op de volgende manier: $p = (q+1)/(P+1)$, waarbij q het aantal permutatieresultaten gelijk aan of extremer dan de geobserveerde waarde weergeeft, en P het aantal permutaties. Verschillende permutatiestrategieën kunnen worden aangewend om de structuur van de geobserveerde dataset in meer of mindere mate te verstoren.

In Hoofdstuk 4 worden twee verschillende permutatiestrategieën om de significantie van de bijdrage van de variabelen aan de lineaire PCA-oplossing te bepalen toegepast in een simulatiestudie: (1) het tegelijkertijd en onafhankelijk permuteren van alle variabelen in de dataset (zie ook Buja & Eyuboglu, 1992), en (2) het afzonderlijk en onafhankelijk permuteren van de variabelen, met andere woorden, één variabele wordt gepermuteerd, terwijl de andere constant gehouden worden. In de simulatiestudie worden de omvang en structuur van de datasets gevarieerd. Bovendien worden verschillende significantieniveaus toegepast: een ongecorrigeerde alpha van 5% alsmede significantieniveaus die gecorrigeerd zijn voor meervoudig toetsen op dezelfde dataset, met behulp van twee verschillende methoden. De eerste correctiemethode, de Bonferroni-correctie, deelt het significantieniveau door het aantal toetsen dat wordt uitgevoerd op de dataset (in Hoofdstuk 4 het aantal variabelen). De tweede correctiemethode is bedoeld om de *False Discovery Rate* (FDR) onder controle te houden (Benjamini & Hochberg, 1995) en houdt in dat de gevonden p-waarden op grootte worden gesorteerd en vervolgens kleinere p-waarden aan strengere significantieniveaus worden onderworpen dan grotere (met als maximum het ongecorrigeerde significantieniveau).

De resultaten van deze simulatiestudie vormen de basis voor de volgende adviezen:

1. Permuteer voor het bepalen van de significantie van de bijdrage van de variabelen altijd afzonderlijke variabelen, terwijl de andere constant worden gehouden. Op die manier wordt de correlatie van de gepermuteerde variabele met de andere variabelen verstoord, maar blijft de structuur tussen de andere variabelen intact. Theoretisch gezien is dit de meest zinvolle strategie, omdat hij gericht is op het vaststellen van de bijdrage van een specifieke variabele bovenop de bijdrage van de andere variabelen, in plaats van op de bijdrage van een variabele vergeleken met de bijdrage van een soortgelijke variabele binnen een compleet aselecte structuur. In de praktijk levert deze strategie ook betere resultaten op wat betreft Type I en Type II fouten dan het permuteren van alle variabelen tegelijk.

2. Voer geen Bonferroni-correctie uit op het significantieniveau. Vooral voor kleine datasets (met 100 personen of minder) leidt deze vorm van correctie tot enorm verlies van power (onderscheidend vermogen) van de permutatietoets.

3. Voor grote datasets (met meer dan 100 personen) kan het best de FDR-correctie van het significantieniveau worden toegepast. Voor kleinere datasets moet een afweging gemaakt worden tussen de ernst van het

maken van een Type I fout en het maken van een Type II fout: Als het risico op een Type I fout als ernstiger wordt beschouwd kan het beste FDR-correctie worden gebruikt; anders kan correctie van het significantieniveau achterwege gelaten worden.

4. In het algemeen werken permutatietoetsen het best voor datasets met meer dan 100 personen. Voor kleinere datasets is de power enigszins laag.

5. Voer permutatietoetsen niet lukraak uit op datasets die mogelijk geen specifieke structuur vertonen. Met andere woorden, verzamel data op een theoretisch zinvolle manier. Voor ongestructureerde datasets zijn de proporties Type I fouten namelijk sterk verhoogd ten opzichte van gestructureerde datasets.

6. Gebruik tenminste 999 permutaties. De p-waarden die kunnen worden berekend op basis van de permutatieresultaten hebben namelijk een ondergrens die bepaald wordt door het totaal aantal permutaties P, namelijk $1/(P+1)$. Bijvoorbeeld, als er 999 permutaties worden uitgevoerd is de kleinste p-waarde die kan worden gevonden gelijk aan $1/1000 = 0.001$, terwijl de kleinst mogelijke p-waarde bij 99 permutaties gelijk is aan $1/100 = 0.01$. Als in het laatste geval een significantieniveau van 0.01 gebruikt wordt, kan een waarde nooit significant bevonden worden. Als het significantieniveau op enige wijze wordt gecorrigeerd moet de kleinst mogelijke p-waarde kleiner zijn dan het kleinste gecorrigeerde significantieniveau.

Statistische significantie van de bijdrage van de variabelen aan de NLPCA-oplossing

In Hoofdstuk 5 wordt de strategie van het afzonderlijk en onafhankelijk permuteren van de variabelen toegepast binnen NLPCA, waarbij p-waarden worden bepaald voor de verklaarde variantie, zowel van de variabelen in totaal als afzonderlijk per principale component. Vooral in de context van permutatietoetsen is het evalueren van p-waarden alleen niet voldoende, aangezien de ondergrens van de p-waarden die kunnen worden berekend afhangt van het aantal permutaties dat is uitgevoerd. Met andere woorden, de p-waarden van geobserveerde resultaten die buiten hun permutatieverdeling vallen zijn allemaal gelijk aan $1/(P+1)$, met P het aantal permutaties. Behalve naar p-waarden moet daarom ook gekeken worden naar maten voor effectgrootte (effect size).

Een voorbeeld van een maat voor effectgrootte is de bijdrage van een variabele aan de totale verklaarde variantie (variance accounted for; VAF) van de oplossing, die wordt weergegeven door de som van de gekwadrateerde componentladingen over de principale componenten. Deze maat is een r^2 en kan daarom gebruikt worden als effectmaat (zie ook Cohen, 1988). Deze effectmaat houdt echter geen rekening met het feit dat sommige geobserveerde waarden verder van hun permutatieverdeling afliggen dan andere (aangezien de permutatieverdelingen niet voor alle variabelen even veel spreiding vertonen). Als alternatief voor de VAF op zichzelf kan daarom gekeken worden naar de afstand tussen de geobserveerde VAF-waarde en de mediaan van de permutatieverdeling. Voor de dataset die geanalyseerd wordt in Hoofdstuk 5 vertoont deze effectmaat echter weinig verschil met de VAF zelf, omdat de mediaan van de permutatieverdeling zeer klein is voor alle variabelen. Dit zal waarschijnlijk opgaan voor de meeste datasets, aangezien de permutatieverdeling een verdeling is van de bijdrage van een gepermuteerde variabele aan een structuur die bepaald wordt door de andere variabelen in de dataset. Deze bijdrage zal naar verwachting klein zijn, zodat ook de mediaan van de verdeling klein zal zijn. Met andere woorden, de afstand van de VAF-waarde tot de mediaan van de permutatieverdeling is een slechts enigszins meer genuanceerde effectmaat dan de VAF-waarde op zichzelf.

In een dataset met n observaties is het totaal aantal mogelijke permutaties van een variabele in principe gelijk aan $n!$. Echter, wanneer meerdere personen dezelfde waarde hebben gescoord op de variabele is een aantal van deze permutaties aan elkaar gelijk. Het aantal mogelijke *verschillende* permutaties van een variabele j is gelijk aan $n!/\prod_{k=1}^{k_j}(f_k!)$, met k_j gelijk aan het aantal verschillende waarden (categorieën) van j en f_k gelijk aan het aantal personen dat een specifieke waarde k scoort. Als het mogelijke aantal verschillende permutaties klein is, zal de permutatieverdeling weinig spreiding vertonen. In NLPCA kunnen behalve numerieke variabelen ook nominale en ordinale variabelen worden geanalyseerd. Vooral bij deze nominale en ordinale variabelen is het aantal verschillende categorieën vaak aanmerkelijk kleiner dan het aantal observaties (n). Daardoor zal het aantal mogelijke verschillende permutaties in NLPCA met nominale of ordinale variabelen vaak kleiner zijn dan in lineaire PCA met numerieke variabelen. Echter, gezien de bovenstaande formule wordt het mogelijke aantal verschillende permutaties alleen heel klein, als het aantal observaties klein is en bovendien de observaties binnen een variabele zeer ongelijk verdeeld zijn over een klein aantal categorieën. In het algemeen zijn zulke variabelen niet erg informatief en leveren ze een probleem op voor elke analysemethode.

Tenslotte zijn de resultaten van de bootstrapstudie uit Hoofdstuk 3 vergeleken met de resultaten van de permutatiestudie uit Hoofdstuk 5 (uitgevoerd op dezelfde dataset). Deze vergelijking leidt tot de conclusie dat beide methoden elkaar aanvullen. In de traditionele context van hypothesetoetsen waarin normaalverdelingen worden verondersteld, leiden betrouwbaarheidsintervallen en p-waarden tot exact dezelfde conclusies. Dit gaat niet op voor de bootstrap betrouwbaarheidsintervallen en de p-waarden uit de permutatietoets voor afzonderlijke variabelen. Hieruit wordt het advies afgeleid om beide methoden naast elkaar te gebruiken: de bootstrap voor het vaststellen van stabiliteit en permutatietoetsen voor het bepalen van statistische significantie.

Ideeën voor toekomstig onderzoek

De studies die zijn uitgevoerd in het kader van dit onderzoek hebben niet alleen een aantal vragen beantwoord, maar ook een aantal nieuwe vragen opgeroepen die een vruchtbare bodem kunnen vormen voor toekomstig onderzoek. In het algemeen kunnen we zeggen dat zowel de bootstrapstudie in Hoofdstuk 3 als de permutatiestudie in Hoofdstuk 5 zijn uitgevoerd op slechts één empirische dataset. Hoewel de resultaten van deze studies overeenstemmen met eerder onderzoek op vergelijkbaar gebied kan het raadzaam zijn aanvullende simulatiestudies uit te voeren.

De bootstrap

Uit de bootstrapstudie werd duidelijk dat categorieën met kleine marginale frequenties problemen opleveren voor de stabiliteit. Een oplossing hiervoor is het samenvoegen van categorieën, maar in bepaalde gevallen zou men ook kunnen denken aan het regulariseren van de optimale transformaties. Op het moment zijn zogenaamde *spline* transformaties (Ramsay, 1988; Winsberg & Ramsay, 1983) beschikbaar in CATPCA (zie Hoofdstuk 2), maar ook andere vormen kunnen in ogenschouw genomen worden.

De bootstrapellipsen bleken enigszins conservatief in het weergeven van de stabiliteit van de NLPCA-resultaten wanneer de bootstrappunten niet evenredig verdeeld waren over de ellips. Alternatieven zijn mogelijk, zoals de onregelmatig gevormde *convex hulls*, *minimum volume ellipses* (Rousseeuw, 1984), of tweedimensionale boxplots (*bagplots*) (Gardner & le Roux, 2003; Rousseeuw et al., 1999).

Behalve de bootstrapprocedure die in dit proefschrift beschreven is kunnen andere methoden worden toegepast voor het bepalen van met name de externe stabiliteit (generaliseerbaarheid) van de NLPCA-resultaten. Hierbij

kan gedacht worden aan bijvoorbeeld de multivariate voorspelbaarheid van de geobserveerde data uit de NLPCA-oplossing (Gower & Blasius, 2005), of aan de .632 bootstrap (Efron, 1983) die gezien kan worden als een vorm van kruisvalidatie.

Permutatietoetsen

Behalve voor het bepalen van de significantie van de bijdrage van variabelen aan de NLPCA-oplossing kunnen permutatietoetsen voor andere doeleinden worden ingezet, bijvoorbeeld voor het bepalen van het optimale aantal te interpreteren principale componenten, waarbij een variant kan worden gebruikt op de methode van Buja en Eyuboglu (1992). Een andere mogelijkheid is om permutatietoetsen te gebruiken om de significantie van de optimale kwantificatie van de variabelen in NLPCA te bepalen. Hierbij kunnen de verschillen tussen een oplossing met en een oplossing zonder optimale kwantificatie in de permutatiestudie betrokken worden.

Evenals de lineaire PCA-oplossing kan de NLPCA-oplossing geroteerd worden in de richting van een bepaalde (simpele) structuur (zoals bij VARIMAX-rotatie). Als de oplossing voor de geobserveerde data geroteerd wordt, lijkt het vanzelfsprekend dat dat ook bij de oplossing voor de gepermuteerde data moet gebeuren. Echter, het roteren van de permutatieresultaten zal waarschijnlijk weinig verschil maken, aangezien de gepermuteerde variabele geen specifieke correlatie (meer) heeft met de andere variabelen en dus naar verwachting alleen kleine ladingen zal krijgen die in verschillende richtingen wijzen ten opzichte van de andere variabelen.

In het onderzoek in Hoofdstuk 4, hadden de gesimuleerde datasets een bepaalde structuur, waarin ervan uitgegaan werd dat de variabelen die een bijdrage leverden aan de oplossing (signaalvariabelen) en de variabelen die dat niet deden (ruisvariabelen) ongecorreleerd waren in de populatie. In volgend onderzoek zou ook aandacht besteed kunnen worden aan andere structuren waarin bijvoorbeeld specifieke correlaties tussen de signaal- en ruisvariabelen toegestaan worden, waardoor een completer beeld ontstaat van de effectiviteit van permutatietoetsen.

In de permutatiestudie hebben we p-waarden berekend die theoretisch vergelijkbaar zijn met de p-waarden uit de traditionele toetstheorie. In de literatuur is echter kritiek geuit op het gebruik van traditionele p-waarden (zie bijvoorbeeld Cohen, 1994; Killeen, 2005, 2006). Het belangrijkste argument van de critici is dat de vraag die we eigenlijk willen beantwoorden met een hypothesetoets – namelijk hoe groot de kans is dat de nulhypothese waar is gegeven de geobserveerde data – verkeerd wordt beantwoord door de kans te berekenen dat de data waar zijn gegeven de nulhypothese. Als alternatief

wordt wel de p_{rep} voorgesteld (Killeen, 2005), een maat die aangeeft hoe groot de kans is dat de richting van een resultaat gerepliceerd kan worden in volgend onderzoek. Deze maat kan gemakkelijk toegepast worden in de context van nonparametrische toetsen, bijvoorbeeld door de proportie bootstrapresultaten die in dezelfde richting wijzen als het geobserveerde resultaat te bepalen. Andere auteurs uiten echter op hun beurt kritiek op p_{rep} (Macdonald, 2005; Wagenmakers & Grünwald, 2006), vooral omdat deze maat uitgaat van een vooraf veronderstelde verdeling van de te meten effectgrootte die moeilijk te bepalen is.

Behalve de maten voor effectgrootte die voorgesteld worden in Hoofdstuk 5 zouden nog andere maten kunnen worden gebruikt, zoals de afstand tussen het gemiddelde van de bootstrapresultaten en het centrum van de populatieverdeling, waarbij het gemiddelde bootstrapresultaat wordt gezien als een betere schatter van de populatiewaarde dan de geobserveerde steekproefwaarde. Echter, aangezien het centrum van de permutatieverdelingen bij het permuteren van afzonderlijke variabelen heel klein is, zal ook deze maat wellicht niet veel verschillen van de geobserveerde VAF-maat op zichzelf.

Implementatie

Dit proefschrift is mede bedoeld om onderzoekers inzicht te geven in de situaties waarin zij NLPCA in ogenschouw zouden moeten nemen als alternatief voor lineaire PCA en in de voordelen die hen dat kan opleveren. De methode is breed beschikbaar, bijvoorbeeld als CATPCA in de SPSS Categories module (Meulman et al., 2004) en als PRINQUAL in SAS (SAS, 1992).

Met behulp van de nonparametrische inferentiemethoden die in dit proefschrift gebruikt worden om de stabiliteit en de statistische significantie van de NLPCA-oplossing te bepalen is aangetoond dat NLPCA niet strikt als exploratieve methode gezien hoeft te worden. Voorlopig zijn deze methoden beschikbaar als SPSS-macrofiles (de bootstrap) en Matlabcode (zowel de bootstrap als permutatietoetsen). In de nabije toekomst zullen zowel de bootstrap- als de permutatieprocedure worden geïmplementeerd in CATPCA, zodat ze beschikbaar worden voor een groot publiek.