SUBTYPE-DETECTIE OP BASIS VAN VARIABELENSTRUCTUUR

Een combinatie van clustering en netwerk-methoden

Netwerkmethodologie biedt de unieke mogelijkheid om de innerlijke werking van complexe vraagstukken beter te begrijpen. In de klinische psychologie wordt deze methodologie steeds meer gebruikt om interacties tussen klinische symptomen in kaart te brengen. Echter, verschillende (klinische) subpopulaties lijken verschillende netwerkstructuren te hebben. Met andere woorden, netwerkmethodologie kan worden gebruikt om subtypes te onderscheiden. Voor mijn masterthese ontwikkelde ik een methode die automatisch subtypes identificeert aan de hand van netwerkstructuren. Deze methode combineert netwerkmethodologie met clustering. De methode wordt vervolgens toegepast voor het identificeren van subtypen in de publieke opinie over immigranten en vluchtelingen.

Graffiti in Gent: Foto: Matteo Paganelli via Unsplash

Vladimir Hazeleger

Netwerkmethodologie is in een razend tempo populair aan het worden in de sociale wetenschappen. Met goed recht: netwerken hebben een solide wiskundige basis en bieden een rijkdom aan methodes om data te bestuderen. Met name *psychologische netwerken* (Epskamp, Borsboom & Fried, 2018) hebben in de afgelopen vijf jaar veel toepassingen gezien. Deze ongerichte netwerken hebben als doel klinisch psychologische stoornissen, zoals bijvoorbeeld depressie of posttraumatische stressstoornis, te helpen verklaren, door de complexe relaties tussen de individuele symptomen in kaart te brengen.

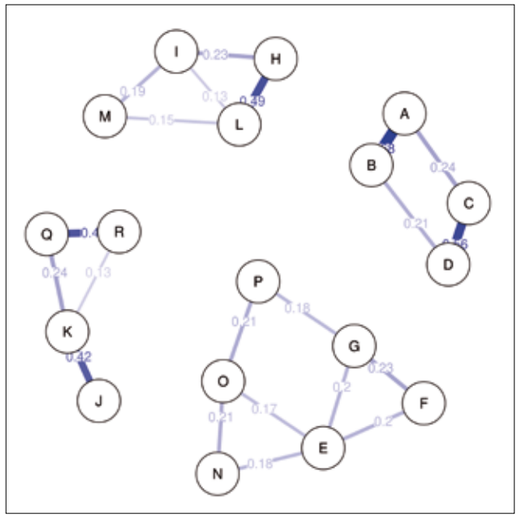
Psychologische netwerken

Een voorbeeld is weergegeven in afbeelding 1. De punten in dit netwerk staan voor individuele symptomen, en de verbindende lijnen staan voor partiële correlaties tussen de symptomen. Deze structuur staat ook wel bekend als een voorwaardelijke onafhankelijkheidsstructuur: twee punten die niet met elkaar verbonden zijn (en dus een partiële correlatie van exact 0 delen) zijn onafhankelijk van elkaar, voorwaardelijk op de rest van de punten in het netwerk. Een dergelijke structuur kan, onder andere, worden gebruikt om symptomen te identificeren die de meeste impact hebben binnen de stoornis, en die dus als doelwit kunnen dienen voor klinische interventies (voor een uitgebreid verslag, zie Epskamp, Borsboom & Fried, 2018).

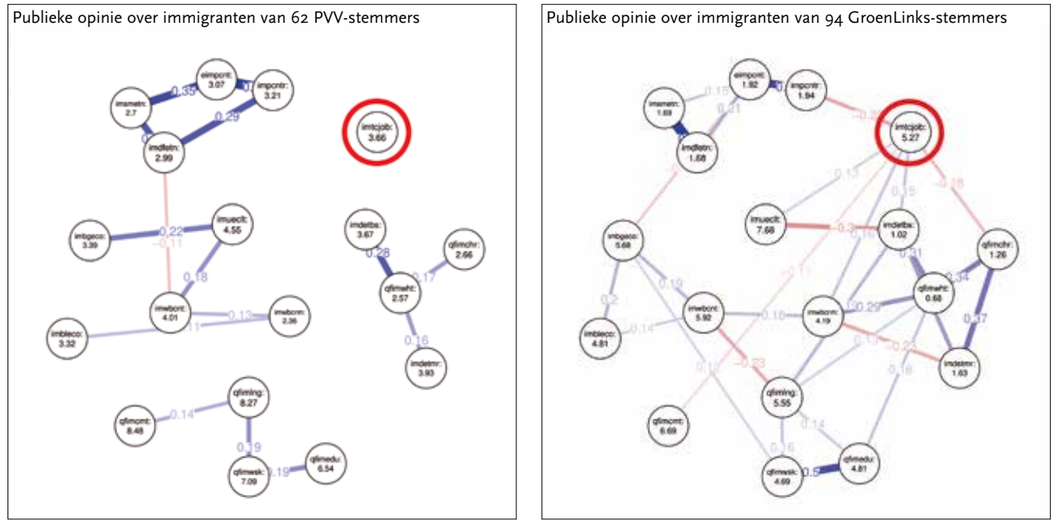
Dat deze netwerkmethodologie unieke inzichten kan bieden, wordt bevestigd door het vele gebruik van de methodologie (bijvoorbeeld Costantini, Epskamp, & Borsboom, 2015; Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016; Dalege et al., 2016). De nieuwe methode roept echter ook een aantal vragen op, waaronder de vraag of verschillende (klinische) subpopulaties ook verschillende netwerkstructuren hebben. Deze vraag is dan ook de inspiratie geweest voor het huidige werk.

Netwerken van publieke opinie

Om een antwoord te geven op deze vraag, wenden we ons tot een publieke dataset: de European Social Survey (ESS round 7, 2014). Deze tweejaarlijkse vragenlijst wordt afgenomen in een groot aantal Europese landen, en bevat vragen over allerhande sociale, politieke en economische vraagstukken. Hieronder vallen ook vragen met betrekking tot de publieke opinie over immigranten en vluchtelingen. De publieke opinie over dit onderwerp is sterk verdeeld. Deze verdeling is onder andere te zien in standpunten van politieke partijen. Als voorbeeld gebruiken we de Nederlandse politieke partijen GroenLinks en



Afbeelding 1. Voorbeeld van een psychologisch netwerk; de punten staan voor variabelen, de blauwe (rode) verbindingen staan voor positieve (negatieve) partiële correlaties



Afbeelding 2. Netwerken van de publieke opinie van GroenLinks-stemmers (links) en PVV-stemmers (rechts); variabele namen en gemiddelde scores staan vermeld in de punten, variabele `imctjob` is omcirkeld in rood

PVV, aan de respectievelijk linkse en rechtse uiteinden van het politieke spectrum. Afbeelding 2 geeft de psychologische netwerken weer van de publieke opinie over immigranten en vluchtelingen van GroenLinks-stemmers en PVV-stemmers. De netwerken laten sterk verschillende structuren zien. In het bijzonder valt op de variabele `imtcjob`. De vraag bij deze variabele luidt: 'Bent u van mening dat immigranten banen wegnemen (1) of creëren (10)?' De netwerkstructuur van de PVV-stemmers laat zien dat de score op deze variabele niet samenhangt met enige andere vraag voor de PVV-stemmers. Het tegendeel is waar voor de GroenLinks-stemmers: voor deze groep hangt de score op deze variabele samen met een groot aantal andere vragen.

Ontdekken van subtypes

Dit simpele voorbeeld leidt tot een tweetal conclusies. Ten eerste lijkt het waarschijnlijk dat verschillende subpopulaties verschillende netwerkstructuren hebben. Ten tweede lijkt het mogelijk om verschillen tussen de subpopulaties te interpreteren aan de hand van de verschillende netwerkstructuren. Gegeven deze aannames, is een vervolgvraag of nieuwe subpopulaties of subtypes kunnen worden ontdekt door het ontdekken van verschillende netwerkstructuren. Dit vereist een combinatie van de netwerkmethode en bestaande methodes voor het ontdekken van groepen in data, ook wel bekend als clustering. Een dergelijke exploratieve methode biedt de mogelijkheid om nieuwe subtypes te ontdekken èn te typeren aan de hand van hun unieke netwerkstructuur.

Publieke opinie over immigranten en vluchtelingen

Voor een demonstratie van de voorgestelde methode gebruiken we opnieuw de data van de European Social Survey, met name de items over immigratie en vluchtelingen. De vragen die worden meegenomen zijn terug te vinden in tabel 1. In tegenstelling tot het voorgaande voorbeeld passen we de methode nu toe op de volledige dataset (N = 31.385) en zoeken we naar subtypes van de publieke opinie over dit onderwerp.

Methode

Om de methode te introduceren worden hier kort de bestaande methodes beschreven voor het maken van netwerken en voor clustering. Vervolgens wordt beschreven hoe deze methodes gecombineerd kunnen worden om zo het gewenste resultaat te behalen. Voor een meer uitgebreide, wiskundige beschrijving van de methode en gedetailleerde beschrijving van mogelijke toepassingen, zie Hazeleger (2020).

Netwerkmethode

Voor het berekenen van psychologische netwerken wordt gebruik gemaakt van het *Gaussian Graphical Model*

KWALIFICATIE VOOR IMMIGRATIE	
`qfimedu`	Op een schaal van 1 tot 10, hoe belangrijk vindt u het opleidingsniveau van een potentiële immigrant om te kwalificeren voor immigratie in uw land?
`qfimcmt`	Op een schaal van 1 tot 10, hoe belangrijk vindt u toewijding aan de lokale manier van leven van een potentiële immigrant om te kwalificeren voor immigratie in uw land?
`qfimlng`	Op een schaal van 1 tot 10, hoe belangrijk vindt u dat een potentiële immigrant de lokale taal beheerst om te kwalificeren voor immigratie in uw land?
`qfimwsk`	Op een schaal van 1 tot 10, hoe belangrijk vindt u het dat een potentiële immigrant relevante arbeidsvaardigheden heeft om te kwalificeren voor immigratie in uw land?
`qfimwht`	Op een schaal van 1 tot 10, hoe belangrijk vindt u het dat een potentiële immigrant blank is om te kwalificeren voor immigratie in uw land?
`qfimchr`	Op een schaal van 1 tot 10, hoe belangrijk vindt u het dat een potentiële immigrant Christelijk is om te kwalificeren voor immigratie in uw land?
TOEGEVOEGDE WAARDE VAN IMMIGRANTEN	
`imtcjob`	Op een schaal van 1 tot 10, denkt u dat immigranten in uw land banen wegnemen (1) of creëren (10)?
`imbgeco`	Op een schaal van 1 tot 10, denkt u dat immigranten de economie in uw land verslechteren (1) of verbeteren (10)?
`imueclt`	Op een schaal van 1 tot 10, denkt u dat immigranten de cultuur in uw land ondermijnen (1) of verrijken (10)?
`imwbcnt`	Op een schaal van 1 tot 10, denkt u dat immigranten uw land slechter (1) of beter (10) maken?
`imbleco`	Op een schaal van 1 tot 10, denkt u dat immigranten over het algemeen meer geld uit de economie in uw land wegnemen (1) of toevoegen (10)?
`imwbcrm`	Op een schaal van 1 tot 10, denkt u dat immigranten criminaliteitsproblemen in uw land verergeren (1) of verbeteren (10)?
BEREIDHEID OM TOE TE LATEN	
`impcntr`	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit arme landen buiten Europa?
`eimpcnt`	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit arme landen binnen Europa?
`imdfetn`	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit landen met een andere etniciteit dan in uw land?
`imsmetn`	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit landen met dezelfde etniciteit als in uw land?
IMMIGRANTEN IN HET EIGEN LEVEN	
`imdetbs`	Op een schaal van 1 tot 10, hoe erg zou u het vinden als een immigrant met een andere etniciteit dan in uw land uw baas zou zijn?
`imdetmr`	Op een schaal van 1 tot 10, hoe erg zou u het vinden als een immigrant met een andere etniciteit dan in uw land met een familielid zou trouwen?

Tabel 1. Afkortingen en beschrijvingen van de 18 items uit de European Social Survey die zijn gebruikt in het huidige werk

(GGM). Het doel van de GGM is om een accurate schatting te maken van de voorwaardelijke onafhankelijkheidsstructuur van variabelen, gegeven een selecte steekproef. Voor deze netwerken wordt aangenomen dat de structuur schaars is: dat wil zeggen dat er van alle mogelijke verbindingen slechts een kleine hoeveelheid waar is. Deze structuur kan vervolgens worden weergegeven als een netwerk waarin punten staan voor variabelen en verbindingen voor partiële correlaties. De partiële correlaties ρ_{ij} kunnen direct worden berekend uit de zogenaamde precisiematrix Ω, aan de hand van de volgende formule:

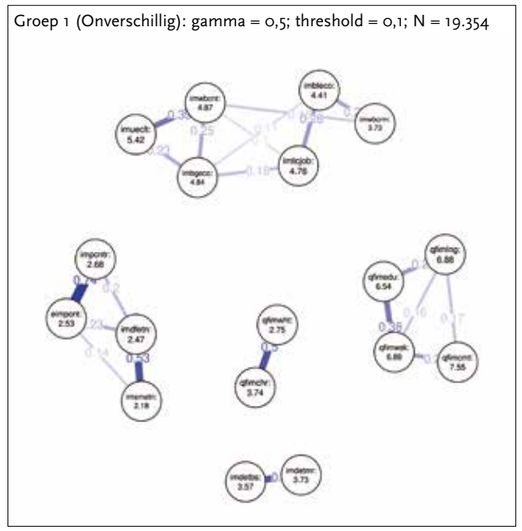
$$\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

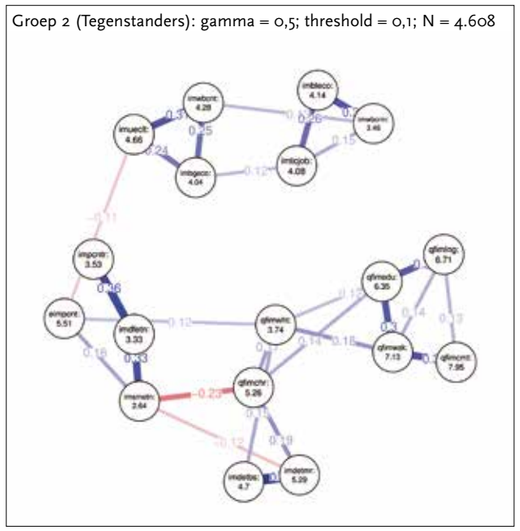
waarbij ω_{ij} staat voor waarde (i,j) van Ω. De precisiematrix is een inversie van de ware covariantiematrix Σ van de data. Echter, de ware covariantiematrix Σ is nagenoeg nooit bekend voor een steekproef. Met de beschikbare covariantiematrix van de steekproef S kan een schatting worden gemaakt van de precisiematrix Ω. Een directe inversie van S leidt echter tot een overschatting van Ω, met als gevolg een niet-schaarse structuur. Om overschatting te voorkomen en een schaarse structuur aan te sporen wordt gebruik gemaakt van regularisatie. Specifiek wordt er gebruik gemaakt van LASSO-regularisatie, waarbij de waardes van Ω worden gelimiteerd. Hoe groter de totale complexiteit van Ω, des te meer worden de waardes gekrompen. Dit proces wordt snel en efficiënt uitgevoerd door het *glasso* algoritme (Friedman, Hastie & Tibshirani, 2008), als volgt:

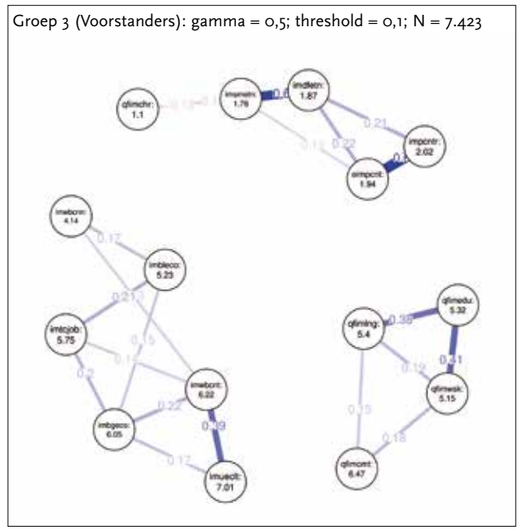
$$\Omega = \text{glasso}(S)$$

Clusteringmethode

Voor het clusteringproces wordt gebruik gemaakt van *Gaussian Mixture Models* (GMM). De GMM onderscheidt K clusters in de data door voor elk cluster een aparte Gaussiaanse (normaal)verdeling te initialiseren. Deze zogenaamde componenten bestaan uit een vector van gemiddelden μ en een covariantiematrix Σ. Vervolgens worden met behulp van *maximum likelihood estimation* (MLE) de optimale waardes van μ en Σ berekend voor elk cluster. Dit wordt gedaan door het Expectation-Maximisation (EM) algoritme. Hierbij wordt iteratief een verwachte waarde van de likelihood berekend, die vervolgens wordt gemaximaliseerd door de parameters μ en Σ te updaten. Het resultaat is een probabilistische clustering, waarbij voor elk datapunt en voor elk cluster de kans is berekend dat het datapunt bij het respectievelijke cluster hoort.







Afbeelding 3. Netwerken van 3 groepen geïdentificeerd door de gecombineerde methode; van links naar rechts: groep 1 (benoemd als 'Onverschillig), groep 2 (benoemd als 'Tegenstanders') en groep 3 (benoemd als 'Voorstanders')

Combinatie

Om de methodes te combineren, gebruiken we een gemeenschappelijke parameter: de covariantiematrix Σ. In de GMM wordt een (ware) covariantiematrix geschat voor elk cluster, welke door het EM-algoritme iteratief wordt geüpdatet. De formule voor de standaard-update is de volgende:

$$\Sigma = \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik} S$$

Om de GGM netwerkstructuur te verwerken in de GMM, maken we gebruik van het feit dat Σ ook gezien kan worden als Ω^{-1}. Zo vervangen we de standaard-update van Σ met een *glasso* schatting van Ω, om zo de schaarse structuur te verweven in het clusteringproces:

$$\Sigma = \text{glasso}(S)^{-1}$$

Beperkingen

Een aantal beperkingen dient in acht genomen te worden bij het gebruik van deze methode. Allereerst is de methode gevoelig voor steekproefgroottes. Wanneer de grootte per groep kleiner is dan 300, is de schatting van de netwerkstructuur minder accuraat en minder stabiel, met als gevolg meer onzekerheid in de interpretatie. Ten tweede is de methode strikt exploratief. De resultaten duiden op een bepaalde structuur in de data, maar dit blijft een schatting. Vervolgonderzoek – en met name statistische toetsen – zijn nodig om de ontdekte structuur te toetsen. Ten derde mist de huidige methode een manier om het aantal groepen in het model automatisch te bepalen. In de huidige iteratie is het aan de gebruiker om het aantal groepen in te voeren.

Resultaten

De methode wordt toegepast op European Social Survey data van de publieke opinie over immigranten en vluchtelingen. Afbeelding 3 geeft de resulterende netwerken weer voor het model waarbij 3 subtypes zijn berekend. De netwerkstructuren van de groepen verschillen noemenswaardig. Met name groep 3 heeft een unieke netwerkstructuur, omdat de variabelen `qfimwht`, `eimpcnt` en `imdetbs` niet opgenomen zijn in de netwerkstructuur. Dit is een gevolg van het feit dat voor alle 7.423 leden van groep 3, de scores op deze variabelen gelijk waren aan 0, met als gevolg dat deze variabelen geen variantie hebben en dus geen correlaties kunnen hebben met andere variabelen.

Interpretatie

Een intuïtieve manier om de resultaten te interpreteren, is door elke groep een benaming te geven aan de hand van eigenschappen van het bijbehorende netwerk. Voor een uitgebreid verslag van de resultaten, zie Hazeleger (2020). Als voorbeeld bekijken we hier het netwerk van groep 2. Deze groep heeft de hoogste gemiddelden op items die peilen hoe belangrijk men bepaalde eigenschappen (zoals opleiding, taalbeheersing, of werkvaardigheden) beschouwt bij het bepalen of een potentiële immigrant mag immigreren. Daarnaast heeft deze groep de laagste scores op items die peilen in hoeverre men gelooft dat immigranten een toegevoegde waarde hebben (bijvoorbeeld voor de economie of de arbeidsmarkt). Qua structurele eigenschappen is te zien dat de variabelen `qfimwht` (belang van blank zijn) en `qfimchr` (belang van Christelijk zijn) centraal verbonden zijn in het netwerk. Daarbij is er een sterke negatieve partiële correlatie tussen de variabelen `imsmetn` (hoeveel immigranten met dezelfde etniciteit zou u toelaten?) en `qfimchr` (belang van Christelijk zijn). Dit betekent dat respondenten die met een hoge score antwoorden op het ene item, waarschijnlijk met een lage score antwoorden op het andere item, en vice versa. Met andere woorden: mensen die het belangrijk vinden dat immigranten Christelijk zijn, maken zich minder zorgen om de etniciteit van immigranten, en vice versa. Dit lijkt een aanwijzing dat Christendom voor deze groep een proxy is van etniciteit.

Gezien het bovenstaande kan deze groep benoemd worden als de 'Tegenstanders' van immigratie. Het netwerk beschrijft een categorie Europeanen die sceptisch zijn over immigranten: ze plaatsen hoge eisen voor kwalificatie voor immigratie en zien minder toegevoegde waarde van immigranten. Daarnaast staan etniciteit en Christendom centraal in hun oordeel over immigranten.

Conclusie

De hier beschreven methode biedt een unieke manier voor het identificeren en interpreteren van subtypes in complexe datastructuren. De methode is een exploratieve manier om data te categoriseren en te visualiseren met behulp van de conditionele afhankelijkheidsstructuur van de data. Deze structuur, gevisualiseerd als netwerken, maakt het mogelijk om complexe structuren te beschrijven, te analyseren en te categoriseren via een andere invalshoek. Toegepast op de publieke opinie over immigranten en vluchtelingen, is de methode in staat om drie groepen te identificeren – namelijk Onverschilligen, Voorstanders, en Tegenstanders – die elk gekarakteriseerd worden door een unieke netwerkstructuur (voor een gedetailleerde beschrijving, zie Hazeleger (2020)).

Literatuur

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Môttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, *123*(1), 2–22.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212.

ESS round 7. (2014). *European social survey round 7 data*. doi: 10.21338/NSD-ESS7-2014.

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Hazeleger, V. (2020). *Identification and interpretation of heterogeneous sparse conditional independence structures using Gaussian Mixture Modelling and Gaussian Graphical Modelling*. TNO-S11150. TNO

Vladimir Hazeleger is afgestudeerd in cognitieve psychologie en kunstmatige intelligentie (KI). Voor zijn masterthese KI heeft hij, in samenwerking met TNO en de Universiteit Utrecht, bovenstaande methode ontwikkeld.
E-mail: Vladimir.hazeleger@gmail.com