Statisztikai Szemle

Közzététel: 2023. április 12.

A tanulmány címe:
Hálózatalapú modell- és adatredukciós módszer

Szerző:
KOSZTYÁN ZSOLT TIBOR
a Pannon Egyetem Gazdaságtudományi Kar Menedzsmentintézet Kvantitatív Módszerek Intézeti Tanszéke tanszékvezetője, egyetemi tanár
E-mail: kosztyan.zsolt@gtk.uni-pannon.hu
DOI: https://doi.org/10.20311/stat2023.04.hu0289

Az alábbi feltételek érvényesek minden, a Központi Statisztikai Hivatal (a továbbiakban: KSH) *Statisztikai Szemle* c. folyóiratában (a továbbiakban: Folyóirat) megjelenő tanulmányra. Felhasználó a tanulmány vagy annak részei felhasználásával egyidejűleg tudomásul veszi a jelen dokumentumban foglalt felhasználási feltételeket, és azokat magára nézve kötelezőnek fogadja el. Tudomásul veszi, hogy a jelen feltételek megszegéséből eredő valamennyi kárért felelősséggel tartozik.

1. A jogszabályi tartalom kivételével a tanulmányok a szerzői jogról szóló 1999. évi LXXVI. törvény (Szjt.) szerint szerzői műnek minősülnek. A szerzői jog jogosultja a KSH.
2. A KSH földrajzi és időbeli korlátozás nélküli, nem kizárólagos, nem átadható, térítésmentes felhasználási jogot biztosít a Felhasználó részére a tanulmány vonatkozásában.
3. A felhasználási jog keretében a Felhasználó jogosult a tanulmány:
   a) oktatási és kutatási célú felhasználására (nyilvánosságra hozatalára és továbbítására a 4. pontban foglalt kivétellel) a Folyóirat és a szerző(k) feltüntetésével;
   b) tartalmáról összefoglaló készítésére az írott és az elektronikus médiában a Folyóirat és a szerző(k) feltüntetésével;
   c) részletének idézésére – az átvevő mű jellege és célja által indokolt terjedelemben és az eredetihez híven – a forrás, valamint az ott megjelölt szerző(k) megnevezésével.
4. A Felhasználó nem jogosult a tanulmány továbbértékesítésére, haszonszerzési célú felhasználására. Ez a korlátozás nem érinti a tanulmány felhasználásával előállított, de az Szjt. szerint önálló szerzői műnek minősülő mű ilyen célú felhasználását.
5. A tanulmány átdolgozása, újra publikálása tilos.
6. A 3. a)–c) pontban foglaltak alapján a Folyóiratot és a szerző(ke)t az alábbiak szerint kell feltüntetni:
„*Forrás*: *Statisztikai Szemle* c. folyóirat 101. évfolyam 3. számában megjelent, ***Kosztyán Zsolt Tibor*** által írt, **Hálózatalapú modell- és adatredukciós módszer** című tanulmány (link csatolása)”
7. A Folyóiratban megjelenő tanulmányok kutatói véleményeket tükröznek, amelyek nem feltétlenül esnek egybe a KSH vagy a szerzők által képviselt intézmények hivatalos álláspontjával.

Kosztyán Zsolt Tibor

Hálózatalapú modell- és adatredukciós módszer*

Network-based dimensionality reduction and analysis

Kosztyán Zsolt Tibor, a Pannon Egyetem Gazdaságtudományi Kar Menedzsmentintézet Kvantitatív Módszerek Intézeti Tanszéke tanszékvezetője, egyetemi tanár
E-mail: kosztyan.zsolt@gtk.uni-pannon.hu

A hálózatelemzés új távlatokat nyit az adatelemzés területén. Az adatpontokat csomópontokként és a közöttük lévő kapcsolatokat élekként ábrázolva „adathálózatot" kapunk, amellyel megnyílik a lehetőség az exponenciálisan fejlődő hálózatos elemzés eszköztárának alkalmazására is. Tanulmányomban egy új, hálózatalapú modell- és adatredukciós módszer létrehozását javaslom, egy olyan, nem paraméteres eljárást, amely modellredukció esetében megadja a látens változók, adatredukció esetében pedig a klasztercentrumok számát. A kialakított módszer robusztus, mivel képes kevés megfigyelés alapján is meghatározni a változócsoportokat, illetve kevés változó alapján az adatcsoportokat. A javasolt módszer alkalmazható szimmetrikus és aszimmetrikus változó- és adattávolságmértékek esetén is. A módszert szimulált és valós adatokon is tesztelem. Az elkészült módszer R-programnyelvben validált csomagként is elérhető.

Kulcsszavak: hálózat, modellredukció, nem paraméteres módszerek, függvénykönyvtárak

Network analysis opens new horizons for data analysis methods, as the results of ever-developing network science can be integrated into classical data analysis techniques. This paper presents the generalized network-based dimensional analysis (GNDA) method. The main contributions of this paper are as follows: (1) The proposed GNDA method handles high dimensional low sample size datasets. In addition, compared with existing methods, we show that only the proposed GNDA method adequately estimates the number of latent variables. (2) The proposed GNDA already considers any symmetric and nonsymmetric similarity functions between indicators (i.e., variables or observations) to specify latent variables. The proposed GNDA method is compared with traditional dimensionality reduction methods on various simulated and real-world datasets. The implementation of the proposed method can be downloaded from the official CRAN site.

Keywords: network, model reduction, non-paramteric methods, packages

* A kutatás a K 142395 számú projekt keretében, a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a K22 OTKA pályázati program finanszírozásában valósult meg. Köszönetemet szeretném kifejezni Dr. Mihálykó Csabának és Dr. Mihálykóné Dr. Orbán Évának, akik még elküldés előtt átnézték és véleményezték a kéziratot, valamint prof. Dr. Görög Mihálynak, aki hasznos tanácsokkal látott el az eredmények interpretálását illetően.

A modell- és az adatredukciós módszerek a leggyakrabban használt redukciós eljárások közé tartoznak mind a társadalom-, mind a természettudományi kutatásokban, több mint egy évszázados múltra tekintenek vissza. A modellredukciós eljárások közül mind a mai napig egyik legnépszerűbb főkomponens-elemzés, a *Principal Component Analysis* (PCA) 1901-ig vezethető vissza, amikor is *Karl Pearson (1901)* kidolgozta a módszert a mechanikában használt tehetetlenségi-nyomaték-elmélet analógiájára. A modellredukcióban leggyakrabban használt másik eljárás is már több mint százéves, és a statisztika egyik nagy alakja, *Charles Spearman (1904)* nevéhez fűződik, aki az intelligencia kutatása során kétfaktoros intelligenciamodellt feltételezett. Az adatredukciós módszerek közül a klaszterezést is a társadalomtudománynak, ezen belül az antropológiának köszönhetjük *(Driver–Kroeber*, *1932)*.

Egy évszázad alatt a módszerek számos változata, fejlesztése jelent meg. Részletes történeti áttekintés helyett most inkább csak az eredeti eljárások néhány olyan hiányosságát emelem ki, amelyeket az évek során a kutatók megpróbáltak kezelni.

A főkomponens-elemzés lényege, hogy egy nagy adathalmaz – amelynek változói korrelálnak egymással – dimenzióit csökkentse, miközben a redukált modell az eredeti változók varianciáját mint információtartalmat a lehető legjobban megtartsa. Az így megfogalmazott célfüggvény miatt a modellredukciós eljárások közül a legtöbb információt a főkomponens-elemzés őrzi meg. A faktorelemzés is modellredukciós módszer, de míg a főkomponens-elemzés az eredeti változókat, addig a faktorelemzés az eredeti változók közötti korrelációs mátrixot akarja minél jobban reprodukálni. Ha a faktorelemzés során a látens változóknak főkomponenseket választunk, akkor az ún. főfaktormódszert (PFA) kapjuk. Ebben az esetben is biztosítható, hogy a legtöbb variancia megmaradjon. Mindkét módszer érzékeny a hiányos elemekre, amit pl. az ún. ritka mátrixokon alapuló PCA-módszer (*Sparse PCA*, SPCA) képes kezelni (*Croux et al*., *2013)*. A másik probléma, hogy mind a főkomponens-, mind a faktorelemzés eredeti és mind a mai napig leggyakrabban használt verziója lineáris kapcsolatokat feltételez a változók között, holott azok nem feltétlenül lineárisak. Ezt a problémát próbálja több-kevesebb sikerrel megoldani a Kernel PCA-módszertana (KPCA) *(Schölkopf et al*., *1998)*, amely bizonyos típusú nemlinearitásokat képes kezelni. Egy másik lehetőség lenne a változók közötti Pearson-féle lineáris korreláció helyett más, pl. a Spearman- vagy a Kendall-féle korreláció alkalmazása, ami monoton, nem feltétlenül lineáris kapcsolatok között is értelmezhető. Tetszőleges kapcsola-

tot pedig a *Székely–Rizzo*-féle *(2009)* különbségi korrelációval is mérhetünk, amely csak akkor ad 0 értéket, ha a két változó független egymástól *(Székely–Rizzo, 2013)*. Ez a módszer azonban rendkívül számításigényes, valamint csak nagyszámú megfigyelés esetén alkalmazható.

Mind a főkomponens-, mind a faktorelemzést a társadalomtudományban akkor alkalmazzuk, ha a megfigyelések száma jóval meghaladja a változókét. Bár *Jung–Marron (2009)* kimutatták, hogy az e két módszerben kalkulált sajátértékek és sajátvektorok ekkor is kiszámíthatók, ezáltal a látens változók is meghatározhatók, egyes kutatók – pl. *Li et al. (2017)*, *Mahmud et al. (2018)*, *Nakayama et al. (2021)* – más eljárásokat javasolnak azokban az esetekben, amelyekben a változók száma jelentősen meghaladja a megfigyelésekét. Az ilyen adattáblákat HDLSS-adattábláknak *(High Dimension Low Sample Size)* nevezik. Ennek fordítottja, a *Low Dimension High Sample Size* sok megfigyelést, de kevés változót tartalmaz, ami a klaszterelemzések esetében okozhat problémát. A problémák egyidejű kezelése egyben az átjárhatóságot is elősegítheti a modell- és az adatredukció között.

Klaszterezés során hasonló feladatot kell megoldanunk, mint a modellredukciónál, ebben az esetben azonban nem változókat, hanem megfigyeléseket kell csoportosítanunk. Választanunk kell egy hasonlósági függvényt, amely már nem feltétlenül korrelációalapú, de szimmetrikus, vagyis ugyanazt az értéket kell kapnunk, ha a hasonlóság páronkénti kiszámításánál a megfigyelések sorrendjét felcseréljük. Ezután egy megfelelő, általában heurisztikus algoritmus segítségével össze kell hasonlítanunk a megfigyeléseket vagy egymással, vagy egy számolt klaszterközéppel, más néven klasztercentrummal. A klaszterközép lehet egy fiktív elem, amely a csoportban lévő megfigyelések átlagából vagy más középértékéből (pl. mediánjából) számítható, de lehet egy konkrét megfigyelés, egy ún. reprezentáns elem is *(Aittokoski et al., 2009)*.

A modellredukció a változókat, az adatredukció a megfigyeléseket csoportosíthatja. Az alkalmazott módszerekben régóta nincs éles határ. Klaszterezési módszereket használhatunk változószelekcióra is, és ha értelmezhető a klasztercentrum (itt szándékosan nem klaszterközepet írok) látens változóként, akkor a modellredukciós módszerek a megfigyelések csoportosítására is alkalmazhatók *(Nakayama et al., 2021)*. Éppen ezért e két, ma még sokszor párhuzamosan fejlődő terület kölcsönösen elősegítheti egymás további fejlődését. Ennek ellenére mindkét redukciónál számos, ma még megoldatlan problémával találkozhatunk.

Az egyik ilyen, hogy mennyi legyen a csoportok száma. Ez modellredukció esetében a látens változók számát is jelenti, klaszterezésnél pedig a klaszterekét. Természetesen ezzel a problémával számos kutató foglalkozott már *(Henning, 2015*; *Szüle, 2019)*. Ugyanakkor e módszerek alkamazásával általában más és más eredményt kapunk. Partícionáló klaszterezésnél, ahol a teret felosztjuk *k*

klaszterre, szintén előre meg kell mondanunk a klaszterek számát *(Aguirre–Taboada, 2011)*. Hierarchikus klaszterezés esetében segítséget nyújthat az ún. dendogram, amely megmutatja, hogy a hasonlósági küszöb változtatásával hány klasztert kaphatunk. Vágni és egyúttal a klaszterek számát becsülni ott szoktuk, ahol az olyan, a dendogramban leghosszabb szakaszt kapjuk, amelyben a klaszterek száma nem változik *(Gallegos–Ritter*, *2018)*. Ugyanakkor, ha a klaszterekre kiszámítjuk a klaszterek jóságát, nem mindig az így kapott klaszterszám lesz az optimális *(Szüle, 2019)*. Ráadásul a klaszterek jóságára használt mutatók – pl. minél nagyobb klaszteren belüli hasonlóság vs. minél kisebb klaszterek közötti hasonlóság, klaszteren belüli homogenitás, kis klaszterszámok stb. – más-más klaszterszámot javasolhatnak.

A másik hiányosság, hogy mind a modellredukció, mind az adatredukció szimmetrikus hasonlósági függvényeket feltételez. Így például nem lehet a közvetett hatásokat kiszűrő szemiparciális korrelációs, vagy a kauzalitást mérő, vagy éppen egy struktúramodellben a regressziós kapcsolatokat csoportosítani.

Véleményem szerint a hálózatelmélet alkalmazása új lendületet adhat a redukciós és csoportosítási módszerek fejlődésének. Egy ilyen hálózatban, redukciós feladattól függően, csomópontok lesznek a változók vagy a megfigyelések *(Nagy–Molontay, 2022)*. A közöttük lévő hasonlóságok mértékét élek reprezentálhatják. A hálózati reprezentációnak számos előnye van, de terjedelmi korlátok miatt csak néhányat mutatunk be. Az egyik előny, hogy számos csoportosítási, ún. modulkeresési algoritmus létezik. A modulokon belül azt feltételezzük, hogy a kapcsolatok szorosabbak, mint a modulok között, így kaphatunk változó-, vagy megfigyeléseket tartalmazó csoportokat. A modulkereső algoritmusok közül némelyik alkalmazható az irányított gráfokra is, ilyenek pl. a Louvain–Leiden-módszerek *(Traag et al*., 2019*)*. Az irányított gráfok pedig az aszimmetrikus hasonlósági függvényeket, pl. a közvetett kapcsolatokat kiszűrő szemiparciális kapcsolatokat, a struktúramodellekben regressziós utakat, vagy éppen egy kauzalitási hálóban a változók közötti kauzalitás szignifikanciaszintjét is jelölhetik. Ebben a cikkben nem térek ki erre, de egy következőben majd részletesen megvizsgálom, hogy a hálózatok szervezhetők-e többrétegű hálózatokba is. Az egyes rétegek reprezentálhatnak egy-egy időszakot, ezzel lehetőség nyílik akár idősoros adatok vizsgálatára is.

A hálózaton számolt modulok meghatározása egy modulkeresési probléma optimalizációjaként fogható fel, aminek eredményeként meghatározott számú változó- vagy megfigyeléscsoportot kapunk. Szemben tehát a modellredukciós és a legtöbb klaszterezési eljárással, a klaszterek és a látens változók számát vagy a hasonlósági küszöbindexet nem a módszer alkalmazása előtt kell megadnunk, e csoportok száma és a csoportokban szereplő tagok az elemzés eredményeként adódnak.

1. A látens változók számának meghatározásától az automatikus változószelekcióig

Ebben a fejezetben a modell- és az adatredukció két legkevésbé egzakt, ugyanakkor az elemzés szempontjából nagyon kritikus lépését tárgyalom. Az első kérdés, amely a modell- és az adatredukció során felmerül, a csoportok, illetve a látens változók száma. A másik, részletesebben tárgyalandó kérdés pedig azoknak a változóknak vagy megfigyeléseknek az elhagyása, amelyek nem illeszkednek a modell- vagy az adatstruktúrába.

A modell- és az adatredukciós elemzések egyik kardinális kérdése, hogy az eredeti változók hány látens változóval jellemezhetők, illetve a megfigyelések hány klaszterbe csoportosulnak. A particionáló klaszterezési eljárások esetében általában előre meg kell határozni a klaszterek számát, majd jóságát, amit különböző mérőszámokkal mérhetünk *(Henning, 2015; Szüle, 2019)*. Ugyanígy a modellredukció esetén is meg kell mondanunk a látens változók számát az elemzés előtt. A modellredukciónál segítséget nyújthat az ún. könyökdiagram meghatározása, ami a látens változók sajátértékeit mutatja meg. Itt a mindmáig leggyakrabban alkalmazott, ugyanakkor sokat kritizált ún. Kaiser-kritérium *(Nunnally–Bernstein, 1994*; *Wasim–Brereton, 2004)* szerint addig őrizzük meg a látensváltozókat, ameddig azok sajátértéke 1-nél nagyobb. A másik, „ökölszabályként" használt kritérium szerint a megmaradó látens változók számának legalább 60%-ban meg kell őriznie az eredeti változók varianciáját *(Hair et al*., *2020)*. Ezt MCVE-módszernek *(Minimal Cumulative Variance Explain)* nevezik. *Bartlett (1950, 1951)* próbát javasolt annak eldöntésére, hogy az elhagyott látens változók szignifikánsan azonosnak tekinthetők-e. Ugyanakkor *Gorsuch (1973)* megmutatta, hogy nagy minták esetében ez a módszer nem használható, mivel nagyon sok látens változót szignifikánsan különbözőnek tekint, így a látens változók számát felülbecsüli.

Velicer (1976) egy parciális korreláción alapuló módszert alkotott, amelyet *Minimum Average Partial*nak (MAP) nevezett el, és 2000-ben kollégáival továbbfejlesztett *(Velicer et al., 2000)*. A négyzetes parciális korreláció átlagát az egyes látens változók meghatározása után számítjuk ki. Az eljárás során a látens változókat mindaddig megtartjuk, amíg a korrelációs mátrixban jelenlévő variancia szisztematikus varianciát reprezentál, ellentétben a reziduálissal vagy a hibavarianciával. A látens változó számának meghatározására leginkább javasolt *(Tran–Formann, 2009)* Horn-féle párhuzamos analízis (*Parallel Analysis*, PA *[Horn, 1965]*) egy Monte-Carlo-alapú szimulációs eljárás: összehasonlítja a megfigyelt sajátértékeket azokkal a sajátértékekkel, amelyeket a korrelálatlan normális eloszlású változókból kaphattunk volna. Egy látens változót tehát csak

abban az esetben tartunk meg, ha az ahhoz tartozó sajátérték nagyobb értéket vesz fel, mint a véletlen adatkészletből származó sajátértékek eloszlásának 95. percentilise. A kutatók a PA-eljárást ajánlják a leggyakrabban a látens változók számának meghatározására. *Tran–Formann (*2009*)* ugyanakkor elméleti és kutatási bizonyítékkal is szolgált arra vonatkozóan, hogy alkalmazása bizonyos esetekben nem javasolt, ugyanis a Horn-féle párhuzamos analízis teljesítményét olyan tényezők befolyásolhatják, mint a mintaméret, az itemdiszkrimináció vagy a korrelációs koefficiens típusa. Az eddig bemutatott eljárások akár más és más látensváltozó-számot is becsülhetnek, így a kutatóknak valamennyi javaslatot végig kell számolniuk. A végső látensváltozó-szám általában attól függ, hogy melyik faktorstruktúrát lehet leginkább interpretálni.

Míg a hagyományos főkomponens- és faktorelemzéseknél sem kapunk egyértelmű választ arra, hogy hány látens változóval dolgozzunk, addig a modern módszerek (SPCA, KPCA) esetében lényegében semmilyen támpont nem létezik erre vonatkozóan. A magyarázott varianciahányad kiszámítása segíthet *(Abonyi et al., 2022)*, de végső soron csak az interpretálhatóság dönti el a látens változók számának meghatározását.

Hasonló probléma jelentkezik a klaszterelemzés során is. Ebben az esetben további nehézségként lép fel, hogy általában nem a korreláció valamely transzformáltját használjuk a megfigyelések közötti hasonlóság mértékeként, hanem a legtöbbször egy euklideszi távolság valamely súlyozott transzformáltját, vagy általános esetben egy tetszőleges szimmetrikus hasonlóságot. Ha hierarchikus klaszterezési eljárást használunk, akkor a klaszterek becslésére egy ún. dendogramot alkalmazhatunk *(Wilkinson–Friendly, 2009)*. A dendogram egyik végén egy csoport, a másik végén az összes csoportosítandó elem áll. Elemeket, csoportokat úgy vonunk össze, hogy a minimális hasonlósági (vagy maximális távolsági) küszöböt növelve egy csoportba tartozónak vesszük azokat az elemeket, amelyek hasonlósága efelett (távolsága ez alatt) van. A küszöbértéket változtatva látjuk, hogy bizonyos küszöbértékek mellett hány klasztert kapnánk. A leghosszabb olyan szakaszt kiválasztva, ahol a klaszterek száma nem változik, becsülhetjük meg a klaszterek számát *(Wilkinson–Friendly*, *2009)*. A másik lehetőség, hogy az egyes lehetséges esetekre ún. klaszterjósági mutatókat számolunk, ami abból adódik, hogy a klaszterektől elvárjuk, hogy az egy klaszterbe tartozó egyedek lehetőleg a leghasonlóbbak legyenek, vagy másképpen fogalmazva, az átlagos távolságuk lehetőleg minél kisebb, a klaszterek közötti távolság pedig minél nagyobb legyen. Az is feltétel, hogy a klaszterek száma viszonylag alacsony, ugyanakkor a klaszterek minél homogénebbek legyenek *(Hanning, 2015)*. Ezeket a kritériumokat általában nagyon nehéz, sőt sokszor lehetetlen egyszerre teljesíteni, ezért a kutatónak el kell döntenie, hogy végül hány klaszter eredményét tudja értelmezni. Fontos megjegyezni, hogy a dendogram a változók számának

meghatározásában is segíthet, ekkor a hasonlóságfüggvény nyilván a korrelációból számítható. A módszer előnye, hogy viszonylag kevés megfigyelésnél is alkalmazható ez az eljárás *(Nakayama et al*., *2021)*.

Modellredukció esetében a látens változók, adatredukció esetében a klaszterek számának meghatározása után újra kinyílik a lehetőségek tárháza. Mind a modellredukció, mind az adatredukció területén számos módszerrel találkozhatunk *(Hu–Pei*, *2018*). A módszerek közös vonása, hogy modellredukciónál általában a változók közötti korreláció, klaszteréknél valamely szimmetrikus hasonlósági függvény alapján végezzük el a csoportosításokat. Ugyanakkor valamennyi módszernél kritikus pont a modellre vagy a klaszterekbe nem illeszkedő egyedek kezelése. Modellredukció, ezen belül is főkomponens- és faktorelemzések esetében azokat a változókat, amelyek egyetlen látens változóval sem, vagy egyszerre több látens változóval is korrelálnak hasonló mértékben, általában elhagyjuk a modellből. Ezek a lépések általában megváltoztatják a látens változók és a többi változó közötti korrelációs értékeket is, így ezeket a változószelekciókat csak lépésről lépésre lehet megtenni. *Abonyi és szerzőtársai (2022)* javasoltak egy heurisztikus változószelekciós eljárást, amelynek három lépését alkalmazhatjuk a főkomponens- és faktorelemzési módszerekre is. Az első lépés során azoktól a változóktól szabadulhatunk meg, amelyek nem korrelálnak egyetlen másikkal sem, vagy a korreláció a küszöbérték alatt volt. A második lépésben az alacsony kommunalitású (ami a legnagyobb korrelációnégyzet az eredeti változó és látens változók között) változókat dobjuk ki egyesével a modellből. Ezt követően azokat az ún. közös indikátorokat hagyjuk el, amelyekről nem dönthető el egyértelműen, hogy melyik látens változóhoz sorolhatók be. Egy indikátor akkor lesz közös, ha az eredeti és a látens változók között korrelációt számolva a két legnagyobb korreláció között a különbség egy adott küszöbérték (pl. 0,2) alatt van, illetve a korrelációs különbség legalább nem kétszeres. A közös indikátorok közül mindig a legkisebb kommunalitásút kell elhagynunk, egészen addig, ameddig már nincs olyan indikátor, amely közösnek mondható. Ezek heurisztikus lépések, nem garantálják a legtöbb változót megőrző látensváltozó-struktúrát, ugyanakkor nagyszámú változók esetén e módszerek segíthetik az elemzést. Ezek az eljárások nem használhatók, ha a hasonlóság nem korrelációalapú, ha a kommunalitást nem lehet számolni, vagy értelmezni.

A kilógó elemeket a klaszterelemzés esetén is kezelni kell. A fenti megfontolás általában akkor alkalmazható, ha a megfigyelések közötti távolság pl. négyzetes korrelációval jellemezhető, és a látens változónak – ami itt inkább klaszterközép, vagy még inkább klasztercentrum – interpretálható jelentése van. Ha a hasonlósági függvény nem korrelációs függvényből, hanem pl. euklideszi vagy más távolságból származik, ez a megközelítés adatredukció esetén nem alkalmazható. Ugyanakkor itt is kezelni kell a kilógó elemeket, mivel torzítják a klaszterköze-

pek becslését. Nagyon kilógó (más elemektől nagy távolságra lévő) elemek esetében olyan, egy vagy néhány megfigyelést tartalmazó klasztereket fogunk azonosítani, amelyek megnehezítik az értelmezhetőséget. A robusztus klaszterezéseknél *(García-Escudero et al., 2010)* a középponttól nagyobb távolságra eső megfigyeléseket lehet persze kisebb súllyal számba venni, de a megoldást általában az adja, ha ezeket az egyedeket kiszűrjük az adatokból.

Meg kell jegyezni, hogy akár a modellredukciónál, akár az adatredukciónál kihagyott elem nem érdektelen a kutató számára. Sokszor ezek hordozzák a legérdekesebb információt. Hogy miért kerültek ki a modellből vagy a klaszterből? Ennek megválaszolása külön kutatást generál. Ugyanakkor a klasszikus modelleknél mindenképpen külön, a robusztus modelleknél mindenképpen kisebb súllyal kell ezeket az egyedeket figyelembe vennünk.

A javasolt módszer a modellredukció és az adatredukció egyik legkényesebb pontján próbál javítani a hálózati modellezés bevonásával. Egyrészt hálózati modulkeresési eljárás segítségével határozom meg a látens változók vagy a megfigyelési csoportok számát. Másrészt a hálózati módszerek kihasználásával kiterjesztem az eljárást aszimmetrikus hasonlósági függvények kezelésére is. Harmadrészt olyan általános változó- és adatszelekciót javaslok, ami opcionálisan ugyan, de modell- vagy adatszelekcióra is használható. Módszeremet R- és Matlab-környezetben is implementáltam. Szimmetrikus hasonlóságok esetén hivatalos R-, illetve Matlab-csomagként, aszimmetrikus hasonlóságok esetén fejlesztői csomagként bárki számára elérhető a módszer, aki az elemzéseit R-, illetve Matlab-nyelven végzi.

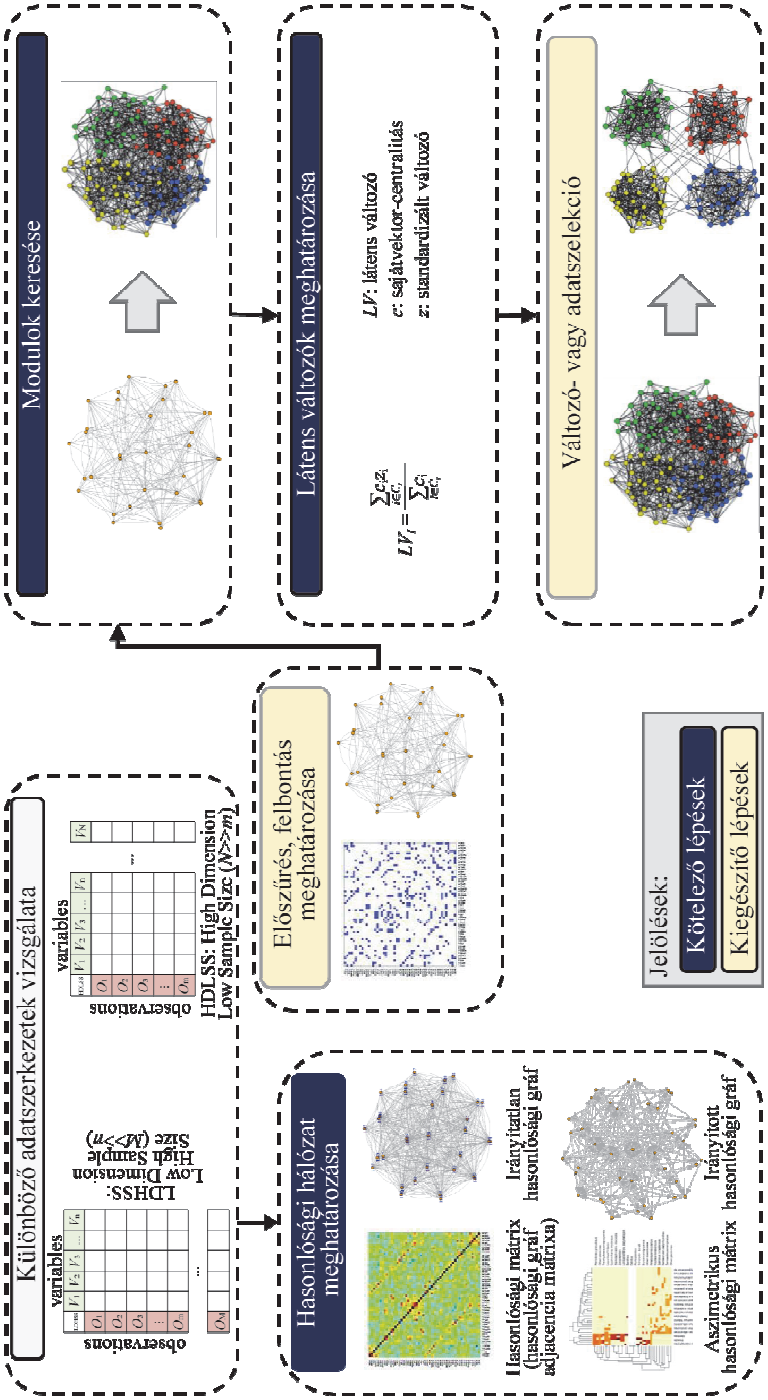
2. Hálózatalapú modell- és adatredukciós módszer

A javasolt hálózatalapú modell- és adatredukciós módszer 5+1 lépésből áll, ebből 3 kötelező, ezek a modell- és az adatredukciós lépések, 3 pedig opcionális, ezek a változó- és az adatszelekciót segítik. A módszer bemutatása során a kötelező, modell(adat-)redukciós, majd a választható, változó(adat-)szelekciós lépéseket ismertetem. A módszer működését az 1. ábra szemlélteti vázlatosan.

1. ábra

A javasolt hálózatalapú modellredukciós eljárás lépései

Steps of Generalized Network-based Dimension Analysis and reduction (GNDA) method



2.1. Modell-(adat-)redukciós lépések

A módszer alapötlete, hogy határozzuk meg – attól függően, hogy modell- vagy adatredukciót végzünk – a változók vagy az adatok egy hasonlósági gráfját. Ebben a gráfban számoljunk sajátvektor-centralitásokat, amelyek kifejezik az adott változó vagy adat relatív fontosságát, majd ezeket súlyként felhasználva alakítsuk ki a látens változókat! A módszer egyaránt használható modell- és adatredukcióra is. A jobb érthetőség kedvéért elsősorban a modellredukción keresztül mutatom be a módszer alkalmazását. Mivel a módszer hasonlósági mértéktől függően, de általában kis elemszám esetében is alkalmazható, könnyen átalakítható adatredukcióvá. A kihívást nem is a hasonlósági mérték megválasztása, sokkal inkább a látens változó értelmezése jelentheti.

1. lépés: hasonlósági gráf meghatározása

Legyen adott egy G (N, A, W) irányított vagy irányítatlan gráfstruktúra, ahol N az n darab csúcsot tartalmazó halmaz, amelyben a csúcsok lehetnek változók vagy megfigyelések is. A tartalmazza az éleket a csomópontok között. W egy $n \times n$-es súlymátrix, ahol $w_{i,j} \geq r_{\min}$ az $a_{i,j} \in A$ súlya, $r_{\min} \geq 0$ pedig a minimális hasonlóság. *Kosztyán et al. (2022a)* javaslata alapján a változók között a távolság lehet Pearson-, Spearman-, vagy Kendall-féle (négyzetes), vagy *Székely–Rizzo (2013)*-féle négyzetes különbségi korreláció is. Utóbbi előnye, hogy két vektor korrelációs értéke akkor és csak akkor 0, ha a két változó független *(Székely–Rizzo, 2013)*. Amennyiben ezek négyzetét vesszük, hasonlóságmértéket kapunk a változók között, vagyis bármely $a_{i,j} \in A$ esetén $0 \leq r_{\min} \leq r_{i,j} \leq 1$. Továbbá igaz az, hogy a hasonlóságmérték szimmetrikus, ebből adódóan a hasonlósági gráf irányítatlan, vagyis: $a_{i,j} \in A \Rightarrow r_{i,j} = r_{i,j}$, $a_{i,j} \in A$. Ugyanakkor látni fogjuk, hogy a későbbi kötelező lépések megengedik, hogy ettől eltérő hasonlóságot is megadhassunk a változók vagy az adatpontok között. Alkalmazhatjuk például a parciális vagy a szemiparciális korrelációt is, ami kiszűri a közvetett kapcsolatokat a változók között. Utóbbi esetben a szimmetria már nem feltétlenül teljesül, a hasonlósági gráf irányított lesz. A módszer alkalmazása szempontjából bármely nem negatív távolsági vagy hasonlósági függvényt használhatunk tehát, ahol teljesül, hogy $\forall_{i,j}$-re $r_{i,j} \geq 0$. A 0 hasonlóság azt jelenti, hogy a két változó vagy adatpont nem hasonló egymással, a 0 távolság pedig azt, hogy nem tudunk közöttük különbséget tenni. Fontos megjegyezni – és erre később, a 0. lépés (előszűrés) tárgyalásakor külön kitérek –, hogy a hasonlósági gráfban minimális hasonlóságot ($r_{\min}$) is előírhatunk, ami alatt két csúcsot nem kötünk össze.

2. lépés: modulok meghatározása

Egy gráf egy modulja egy részgráf, ahol a gráfon belül az összeköttetések sűrűbbek, mint a modulok között. Ezt az alapgondolatot először *Newman (2006)* vetette fel, és számos kutató továbbfejlesztette. *Newman* javaslata alapján irányítatlan gráfok esetén az (1)-es, irányított gráfok esetén a (2)-es egyenletet kell minimalizálni,

$$M = \frac{1}{2L}\sum_{i,j}\left(r_{i,j} - \gamma\hat{r}_{i,j}\right)\delta(C_i, C_j), \tag{1}$$

$$M = \frac{1}{L}\sum_{i,j}\left(r_{i,j} - \gamma\hat{r}_{i,j}\right)\delta(C_i, C_j), \tag{2}$$

ahol M az ún. modularitási érték; $r_{i,j}$ a súlyérték i, j csúcsok között, ami jelenthet pl. korreláció-, vagy parciális korrelációnégyzetet a változók, vagy éppen euklideszi távolságot az adatpontok között. $\hat{r}_{i,j}$ egy ún. nullmodell, amihez hasonlítom az összeköttetések számát, vagy a hasonlóságokat. *Newman (2006)* úgy vélekedett, hogy ez a nullmodell legyen a statisztikából ismert független eset, vagyis jelen esetben a súlyozott bemenő és a súlyozott kimenő élek szorzatának és az összes él összegének hányadosa, képlettel: $\hat{r}_{i,j} = r_{i\cdot} \cdot r_{\cdot j} / r_{\cdot\cdot}$, ahol $r_{i\cdot} = \sum_{j:=1}^{n} r_{i,j}$, $r_{\cdot j} = \sum_{i:=1}^{n} r_{i,j}$, $r_{\cdot\cdot} = L = \sum_{i:=1}^{n}\sum_{j:=1}^{n} r_{i,j}$; a $\gamma > 0$ (alapértelmezés szerint $\gamma = 1$) egy ún. felbontási paraméter, amely ha nagyobb értéket vesz fel, kisebb modulok keletkeznek, míg 1-nél kisebb értékek esetén nagyobb modulokat kaphatunk. C_i, C_j, az i-edik és a j-edik modulokat jelölik, a δ az ún. Kronecker-delta, ami akkor 1, ha i és j csúcs is ugyanabban a modulban van, különben 0. *Newman (2006)* modellje jól értelmezhető nemcsak véletlen, hanem ún. skálafüggetlen[1] hálózatokra is. Ugyanakkor számos más nullmodell is használható, amennyiben van valamilyen *a priori* információnk a csúcsokról, például azok földrajzi elhelyezkedéséről *(Barthélemy, 2011; Gadár et al., 2018).* Általános esetben *Newman (2006)* modellje használatos a modulok keresésekor.

Az (1)-es és a (2)-es egyenletben szereplő modulértéket számos kutató megpróbálta minimálni. Ebből adódóan a modulkereső algoritmusoknak egész tárháza ismert *(Zelditch–Goswami, 2021)*. A leggyakrabban az ún. Louvain-algoritmust alkalmazzák, ez egy heurisztikus eljárás, legújabb más módszerekkel összehasonlítva legjobb eredményeket szolgáltató változata az ún. Leiden-algoritmus *(Traag et al. 2019)*.

[1] A skálafüggetlen hálózatoknál a hálózat csúcsainak fokszámeloszlása a hatványeloszlást követi, így a legtöbb csúcsnak csupán néhány, míg néhány kitüntetett csúcsnak nagyon sok kapcsolata van.

A módszer eredményeként megkapjuk (1) a modulok számát és (2) az egyedek modulokba történő besorolását.

3. lépés: látens változók, klasztercentrumok meghatározása

A látens változók (modellredukció esetén), valamint a klasztercentrumok (adatredukció esetén) a sajátvektor-centralitás és a standardizált adatvektorok (egy adattáblában modellredukció esetén oszlopok, adatredukciónál sorok) lineáris kombinációjából adódnak.

$$LV_I = \frac{\sum_{i \in C_I} c_i z_i}{\sum_{i \in C_I} c_i}, \tag{3}$$

ahol LV_I az I-edik látens változó, C_I az I-edik modul. c_i az i-edik változó (megfigyelés) sajátvektor-centralitása, z_i pedig az i-edik változó (megfigyelés) standardizált értéke.

A centralitási mérőszámokat az adott csúcs fontosságának számszerűsítésére használják. A gráfelméletben a sajátvektor-centralitás (más néven sajátcentralitás vagy presztízspontszám) a hálózat egy csomópontja befolyásának a mértéke. A magas érték azt jelenti, hogy egy csomópont sok másik csomóponthoz kapcsolódik, amelyek maguk is magas pontszámmal rendelkeznek. A sajátvektor-centralitás alapötlete szerint tehát a centralitási mérőszám meghatározásánál a szomszédok nem azonos értékkel, hanem fontosságuk szerint járulnak hozzá az adott csúcs fontosságához.

$$c_k = \frac{1}{\lambda} \sum_{l \in N(k)} c_l = \frac{1}{\lambda} \sum_{l \in G} a_{k,l} c_l, \tag{4}$$

ahol $N(k)$ a k-adik csúcs szomszédjait tartalmazza, $a_{k,l}$ jelöli (k, l) él súlyát. λ pedig egy kostans.

A sajátvektor-centralitás alkalmazása több szempontból is előnyös. Egyrészt alkalmazható irányított és irányítatlan gráfok esetében is, másrészt bizonyos feltételek meglétekor egyértelmű és érzéketlen egy átlagos centralitású új elem felvételére. A sajátérték-centralitás tehát egyfajta beágyazottságot mér, a látens változó és a klasztercentrum számításánál pedig a nagyobb beágyazottságot mutató értékek nagyobb súllyal szerepelnek.

Amennyiben itt megállnék, a 2. lépésben megkapnám a modulokat, ami modellredukció esetén a változók, adatredukció esetén az adatok egy csoportja. A csoporton belül a változók/adatok hasonlóbbak egymáshoz, mint amekkora a hasonlóság a csoportok között. A 2. lépés már megadja a modulok számát, ami egyben a látens változók (klasztercentrumok) száma is lesz. Vegyük észre, hogy a minimális hasonlóság ($r_{\min}$), illetve a felbontás (γ) megválasztásával a modulok

száma növelhető, ám közvetlen ráhatás nincs a látens változók számának meghatározására.

A hálózat megjelenítésére egy ún. Force Atlas II algoritmust *(Jacomy et al., 2014)* használtam. Ez a módszer a csomópontokat tömegpontokként kezeli. A tömegpontok tömege a kapcsolataik számától és a beágyazottságától függ. Egy-egy modul középpontjában a leginkább beágyazott csomópont szerepel, a periférián pedig azok, amelyeknek nincs, vagy alig van kapcsolatuk. A megjelenítés előnye, hogy az így kialakított gráfon az elhelyezkedés további információval szolgálhat a kutató számára. Hátránya, hogy nagyon sok kapcsolat esetében a módszer rendkívül számításigényes, így érdemes a számítás gyorsításához az alacsony súlyú éleket elhagyni. Ugyanakkor meg kell jegyezni, hogy a hasonlósági gráf megjelenítése során alkalmazott élek elhagyása kizárólag a vizualizációt segíti, semmilyen hatása nincs a modulok kialakítására vagy a látens változó számosságára.

2. 2. Változó- és adatszelekciós lépések

Ezek a lépések már nem kötelezők, illetve az 5. lépés csak korrelációs gráfok esetében értelmezhető. A lépések célja, hogy azokat a változókat, amelyek nem illeszkednek a látens változókra, fokozatosan elhagyjuk. Valamennyi lépés rekurzív és heurisztikus. A rekurzív azt jelenti, hogy amennyiben egyszerre több csúcs is megfelel az elhagyási kritériumnak, mindig csak a legkevésbé megfelelőt hagyom el, és újraszámolom a látens változókat. Heurisztikus a lépés, mert abból a feltételezésből adódik, hogy a kisebb kontribúcióval rendelkező, kisebb súllyal szereplő változók kevésbé határozzák meg a klasztercentrumokat vagy a látens változókat, így kevésbé illeszkednek is rájuk.

4. lépés: periferiális csúcsok elhagyása

A 3. lépésben minden változó sajátvektor-centralitását kiszámítom. Ha ezek a centralitások egy $c_{\min} \geq 0$ érték alatt találhatók, azokat a továbbiakban periferiális csúcsoknak nevezem. Bár a sajátvektor-centralitás kevésbé érzékeny egy új elem felvételére, illetve egy periférián lévő csúcs elhagyására, a centralitásokat szükséges újraszámolni. A műveletet addig végzem, ameddig minden egyes centralitási érték e minimumérték felett van, és az előre meghatározott modulonkénti változószám felett vagyok.

A lépés nagy előnye, hogy lényegében bármely (nem negatív) hasonlósági függvénynél használható, nem csak a korrelációknál.

5. lépés: a kommunalitások vizsgálata

Szemben az előző lépéssel, ez csak akkor értelmezhető, ha létezik a változók között korreláció. Két részlépésből áll. Az elsőben az egyedi kommunalitásokat vizsgálom, a másodikban meghatározom az ún. közös indikátorok halmazát. Hasonlóan a főkomponens- vagy a faktorelemzéshez, itt is kiszámítom az eredeti változók és a látens változók közötti korrelációs négyzeteket. Ezek közül a legnagyobbat nevezem kommunalitásnak. Ha ez az adott változó esetén a $h_{\min}$-nél kisebb, akkor ezeket a változókat a legkisebb kommunalitásúval kezdve fokozatosan elhagyom, addig, ameddig a valamennyi változó kommunalitás nagyobb nem lesz, mint ez a minimális kommunalitási érték. Fontos megjegyezni, hogy ezt a részlépést is akkor lehet értelmezni, ha a hasonlósági függvény korreláció. A látens változótól mindig elvárható, hogy jól jellemezze az eredeti változókat, amit a legtöbbször korrelációval mérünk. Így tehát ezt a lépést is akkor tudom értelmezni, ha a változók közötti hasonlóság is korrelációalapú. A második részlépés is feltételezi, hogy hasonlóságnak a korrelációt választottam. Ebben az esetben azt vizsgálom, hogy a változók mely látens változókkal korrelálnak leginkább. Ha több változóval is magas a korrelációjuk, vagyis az eredeti változó és a legnagyobb és második legnagyobb korrelációjú látens változó korrelációi között nincs legalább $C_{\min} \geq 0$, vagy kétszeres különbség, akkor az adott változót *közös indikátor*nak nevezem. A közös indikátorok közül a legkisebb kommunalitásút hagyom el, mindaddig, ameddig még van közös indikátor.

0. lépés: előszűrés, felbontás változtatása

Fontos megjegyezni, hogy a látens változók számára a $c_{\min}$, $C_{\min}$, $h_{\min}$ hiperparaméterek megválasztásának nincs hatása. Az $r_{\min}$, γ paraméterek növelése azonban ritkábbá teszi a hasonlósági gráfot, amely így több modulra fog szétbomlani. Érdekes lehet e paraméterek függvényében egy dendogramot készíteni, mert ez segítheti az egyes látens változók robusztusságának vizsgálatát. Bár 0. lépésként hivatkoztam rá, e két értéket a modulok keresése előtt is meg lehet adni (lásd 1. ábra).

3. Adatforrások

A módszer eredményét két adatforráson mutatom be. Az első egy szintetikus adattábla, amelyet mesterségesen hoztam léte. A második a leideni egyetem 2020-as rangsortáblázata, amely összesen 42 indikátort tartalmaz, 1176 egyetemről, 4 különböző tudományterületen, illetve valamennyi tudományterületet ma-

gában foglalóan 7 időszakra vonatkoztatva. Ezek az időszakok 3 éves periódusokban tartalmazzák az egyetemek publikációs és kollaborációs mutatóit.

3. 1. Szintetikus adattábla generálása

A szintetikus adattáblák generálásánál több célt is megpróbáltam egyszerre megvalósítani:

1. a generálás során előre meg lehessen adni a változócsoportok és így a látens változók számát;
2. lehessen olyan adatokat is generálni, ahol jóval több változóm van, mint amennyi megfigyelésem;
3. a zaj mértékét is be lehessen állítani.

A cél elérése érdekében először b darab, n elemű független bázisvektort generáltam, ahol minden bázisvektor $\lceil n/b \rceil$ darab 1-est és a többi helyen 0-át tartalmazott. A bázisvektorok elrendezését mutatja az (5)-ös egyenlet, ahol $\lceil \cdot \rceil$ a felső egész részt jelöli.

$$\begin{aligned} \mathbf{e}_1 &= (\overbrace{1,1,\ldots,1}^{\lceil n/b \rceil},0,0,\ldots,0)^T \\ &\ldots \\ \mathbf{e}_b &= (0,0,\ldots,0,\underbrace{1,1,\ldots,1}_{\lceil n/b \rceil})^T \end{aligned} \tag{5}$$

Ezek a vektorok függetlenek egymástól. A vektorokat $\lceil m/b \rceil$-szer lemásolom, így egy bináris blokkmátrix képezhető. A következő példa egy $n = 6$ megfigyelést, $m = 5$ változót és $b = 2$, illetve $b = 3$ blokkokat tartalmazó blokkmátrixokat ad. A javasolt **B** blokkmátrix jelölése: $\mathbf{B}_b^{n\times m}$.

$$\mathbf{B}_2^{(6\times 5)} = \begin{pmatrix} \overbrace{1}^{\mathbf{e}_1}\overbrace{1}^{\mathbf{e}_1}\overbrace{1}^{\mathbf{e}_1}\overbrace{0}^{\mathbf{e}_2}\overbrace{0}^{\mathbf{e}_2} \\ 1\,1\,1\,0\,0 \\ 1\,1\,1\,0\,0 \\ 0\,0\,0\,1\,1 \\ 0\,0\,0\,1\,1 \\ 0\,0\,0\,1\,1 \end{pmatrix}, \quad \mathbf{B}_3^{(6\times 5)} = \begin{pmatrix} \overbrace{1}^{\mathbf{e}_1}\overbrace{1}^{\mathbf{e}_1}\overbrace{0}^{\mathbf{e}_2}\overbrace{0}^{\mathbf{e}_2}\overbrace{0}^{\mathbf{e}_3} \\ 1\,1\,0\,0\,0 \\ 1\,1\,0\,0\,0 \\ 0\,0\,1\,1\,0 \\ 0\,0\,1\,1\,0 \\ 0\,0\,0\,0\,1 \\ 0\,0\,0\,0\,1 \end{pmatrix}. \tag{6}$$

A következőkben a blokkmátrixban szereplő valamennyi értéket egy [0,1] intervallumon felvett egyenletes eloszlást követő véletlenszám-generátorral előállított értékkel megszorzom, majd osztom e^{λ} értékkel, amelyet kivonok az eredeti blokkmátrixból. Képlettel:

$$\mathbf{M}_{b,\lambda}^{(n\times m)} = \mathbf{B}_b^{(n\times m)} - \mathbf{B}_b^{(n\times m)} \circ \mathbf{U}^{(n\times m)} / \exp(\lambda), \tag{7}$$

$\mathbf{M}_{b,\lambda}^{(n\times m)}$ jelöli az eredményül kapott blokkmátrixot. $\mathbf{U}^{(n\times m)}$ jelöli az $n\times m$ darab, [0,1] intervallumon vett egyenletes eloszlást követő véletlenszám-generátorral generált számot. $\circ$ jelöli a pontonkénti szorzást, biztosítva ezáltal, hogy a blokkmátrix 0 elemei továbbra is 0-ák maradjanak. $1/\exp(\lambda)$ pedig az ún. csillapítási tényező, ahol $\lambda \in \mathbb{R}$ tetszőleges valós szám. Ha $\lambda \to \infty \Rightarrow \mathbf{M}_{b,\lambda}^{(n\times m)} \to \mathbf{B}_b^{(n\times m)}$, vagyis amennyiben a csillapítási tényező a végtelenhez tart, visszakapom az eredeti, bináris blokkmátrixunkat.

A kitűzött célnak megfelelően a változók száma (m), a megfigyelések száma (n), a blokkok (itt most változó-, illetve adatcsoportok) száma (b), valamint a zajt kontrolláló csillapítási tényező tetszőlegesen beállítható.

3. 2. Az egyetemi kutatás és együttműködés vizsgálata

A leideni egyetem az általa készített rangsort *CWTS Leiden Ranking* néven 2011-től teszi elérhetővé, amely több mint 1000 felsőoktatási intézmény publikációs és kollaborációs adatait tartalmazza. Bár a nevében szerepel a *ranking* szó, a lista mégse jelent egyértelmű intézményi rangsort, inkább a U-Multirankhez hasonlóan mindenki maga válogathatja ki a megfelelő indikátorokat, és azok szerint rakhatja sorba az intézményeket. A mutatók között vannak abszolútok és relatívok is. Néhány mutató esetében csak becsléseket kapunk, így azokra felső és alsó határértékek is megjelennek. Ez utóbbi mutatókat elhagytam, tehát összesen 42 mutatót kaptam (lásd a Mellékletben az M1. táblázatot). Ugyanakkor ezek a mutatók erősen korrelálnak egymással, és a tartalmuk is részben átfedő. Például mérik a folyóiratok rangsorának a felső percentilisébe (top 1%), első decilisébe (top 10%) és első két kvartilisébe (top 50%) tartozó publikációk számát, ahol az 50%-ban benne van a legjobb 10%-nyi és a legjobb 1%-nyi publikáció is.

A cikk írása közben készült el a legújabb rangsor *(CWTS Leiden Ranking 2022)*. A leideni rangsor mindig három, de legfeljebb a rangsor publikálása előtti két év publikációs és kollaborációs teljesítményét veszi számba. Így pl. a 2022-es rangsor a 2017–2020-as teljesítmények számbavételével zárul. Tanulmányomban a 2020-as rangsort vettem alapul, ebből is az utolsó, a 2015–2018-as időintervallumot, amely még a publikációk számára és sajnos minőségére is hatást gyakorló Covid19-járvány előtti időszakra vonatkozóan tartalmazta összesen 1176 felsőoktatási intézmény publikációs teljesítményét. A leideni egyetem rangsortáblázatának indikátorlistája és az egyetemek köre az évenként megjelenő rangsor kiadásával visszamenőleg is bővül, így a 2020-as listába a korábban is minden

évben mért publikációs tevékenységet és az együttműködést megjelenítő mutatókon kívül bekerült a nyílt forrású kiadványokban való publikálás száma, aránya, valamint a nemi arányok is a kutatásokban, publikációkban.

Az adatbázis használata mellett több érv is szól. Egyrészt ez olyan rangsor, amely tartalmazza az indikátorokat is (szám szerint 42-t). Fontos megjegyezni, hogy ez a legtöbb rangsor esetén egyáltalán nem evidens *(Banász et al., 2021)*, hiszen legtöbbször csak a módszertant és a végső rangsort, vagy az egy-egy részterületre adott pontértéket tartalmaznak. A *CWTS Leiden Ranking* indikátorai meglehetősen korrelálnak egymással, ami a legtöbb modellredukciós eljárásnak feltétele, így a látens változókat *(Fauzi et al., 2020)*, valamint a többdimenziós klasztereket *(Kosztyán et al., 2022b)* már számos kutató kereste. Véleményem szerint a legtovább *Abonyi és szerzőtársai (2021)* jutottak, akik aszerint minősítették az egyes indikátorokat, hogy a belőlük képzett kompozit rangsorhoz képest az eredeti indikátorok rangsortávolsága milyen messze áll. Ez alapján kimutatták, hogy a végső sorrend kialakításához az újonnan bevont gendermutatók nem járulnak hozzá, illetve pontosabban fogalmazva pont annyival járulnak hozzá, mintha egy véletlen sorrend és a végső sorrend rangtávolságát vizsgálnánk.

További fontos érv az adatbázis mellett, hogy a rangsor eredeti számított értékeket és nagyon kevés hiányzó elemet tartalmaz. Ráadásul az adatforrás bárki számára hozzáférhető. Több dimenzió is megjelenik benne, a példámban összesen 2-nek az adatait használtam. Ezek: (1) periódus: a 2015–2018-as időszak; (2) tudományterület: az 5 tudományterületet (orvostudomány, élettudomány, matematika, természettudomány, társadalomtudomány), de az összeset is vizsgáltam; (4) az intézmények közül mind az 1176 intézményt, az indikátorok közül 42 indikátort vettem számításba. Azokat az indikátorokat kihagytam, amelyek más indikátoroknak a felső vagy az alsó becsléseiként jelentek meg. Így egy 1176×42-es adattáblát kaptam, amit hagyományos főkomponens- és faktorelemzéssel is lehet vizsgálni. Eredményeimet összevetettem az általam javasolt hálózatalapú modellredukciós módszerrel.

4. Eredmények

A modellredukciók eredményét először szintetikus adatsorokon, majd a példaként választott adatbázison mutatom be.

4. 1. Látens változók számának meghatározása

Ebben a vizsgálatban arra voltam kíváncsi, hogy szintetikus adatok, valamint különböző zajhatások esetében mennyire lehet visszaállítani a faktorstruktúrát. Ehhez különböző csillapítási tényezőkkel az alább tárgyalt módon zajjal terhelt blokkmátrixokat generáltam. Ennek az az előnye, hogy pontosan lehet tudni, hogy hány látens változó (blokk) köré szerveződnek az indikátorok. Így nemcsak a látens változók számát lehet mérni, hanem azt is, hogy az indikátorok besorolása helyesen történt-e meg. A 2. ábra két szintetikus blokkmátrix korrelogramját (2 {a, b}. ábra), valamint főkomponens- (PCA) és főfaktorelemzéssel (PFA) számolt könyökdiagramot (2 {c, d}. ábra) mutat, ahol a változók száma m=50, a megfigyelések száma n=300, a blokkok száma pedig b=5. Két csillapítási tényezőt ($\lambda = 1$, $\lambda = -1$) használtam.

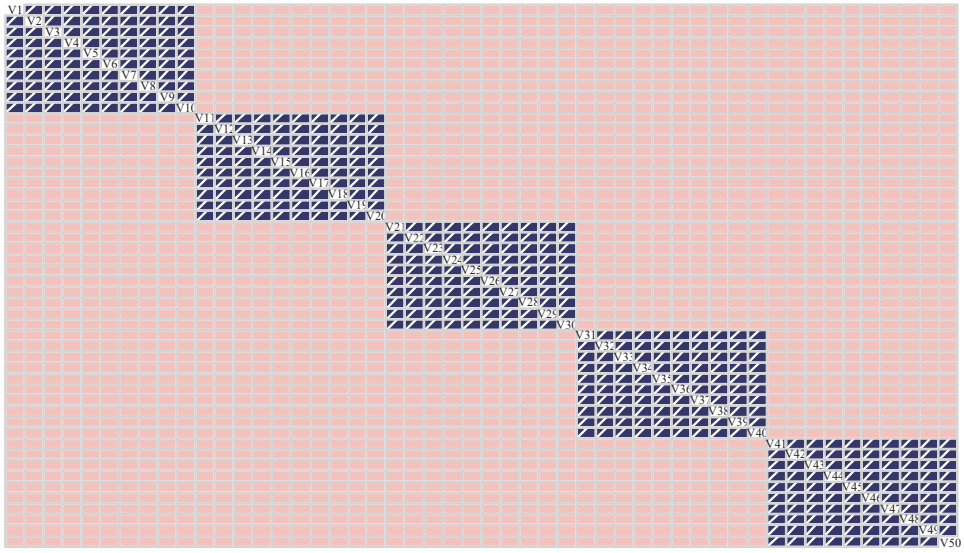
A 2 (a, b). ábrán a korrelogramokon jól azonosíthatók a blokkok. Ennek ellenére a 2 (c, d). ábra jól mutatja, hogy a Kaiser-kritériumot követve rendre alul vagy felül becsülnénk a látens változók számát.

2. ábra

A szintetikus blokkmátrix korrelogramja, valamint a főkomponens- (PCA) és főfaktorelemzéssel (PFA) számolt látens változókra vonatkozó könyökdiagramok (n=300, m=50, b=5)

Correlogram and screeplots of generated block matrices (n=300, m=50, b=5)

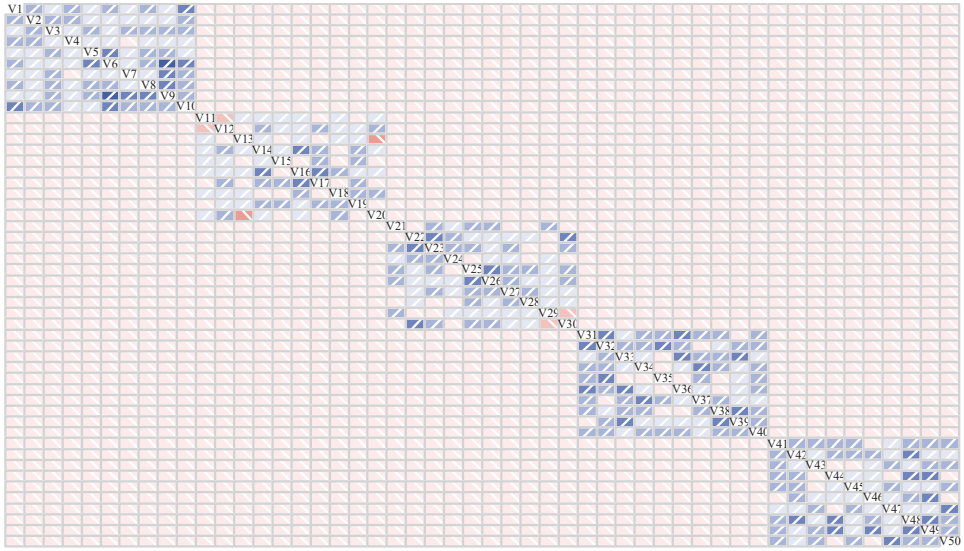
(a) Korrelogram ($\lambda = 1$)



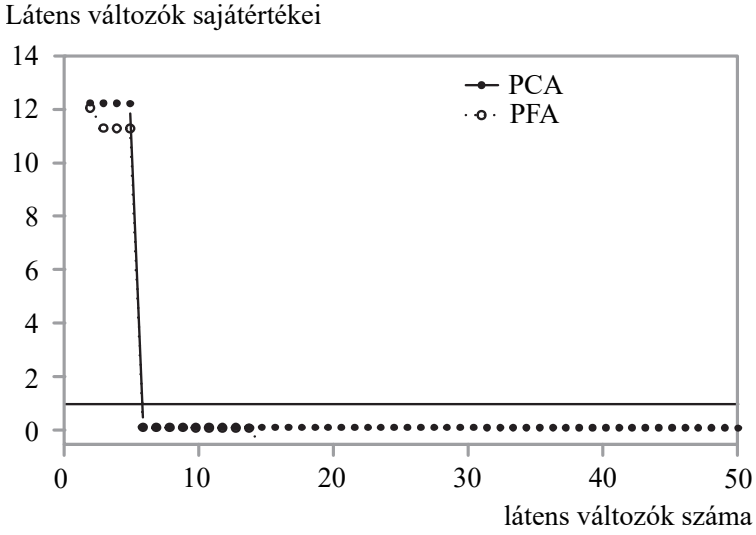
(Az ábra folytatása a következő oldalon)

(folytatás)

(b) Korrelogram ($\lambda = -1$)



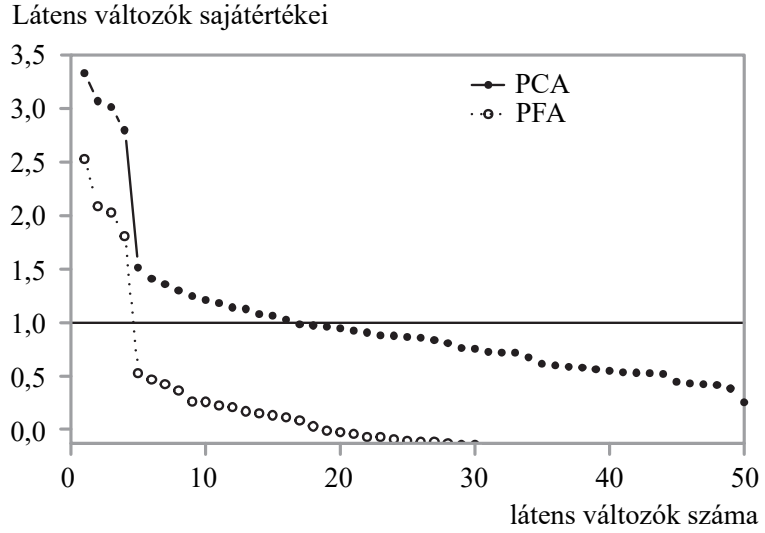
(c) Könyökdiagram ($\lambda = 1$)



(Az ábra folytatása a következő oldalon)

(folytatás)

(d) Könyökdiagram ($\lambda = -1$)



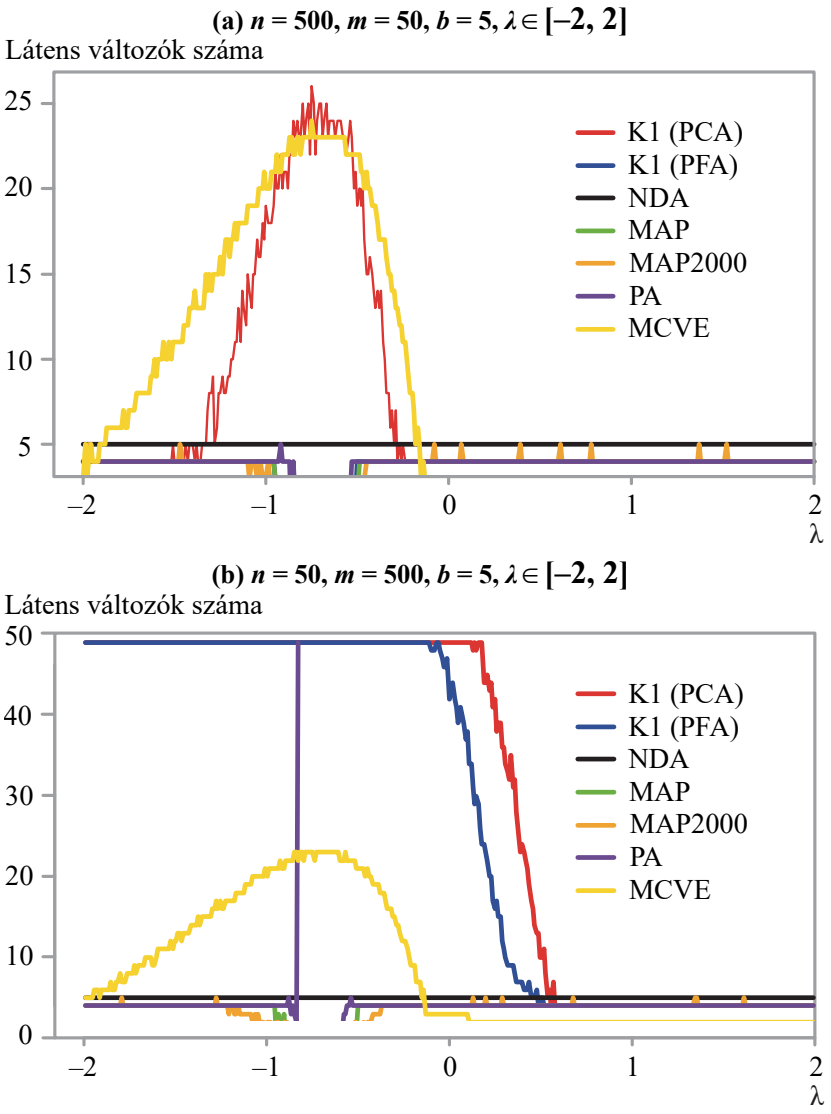
Forrás: saját szerkesztés.

A 3. ábra azt mutatja, hogy a különböző módszerek hogyan becsülnék a látens változók számát akkor, ha a megfigyelések száma (a) tízszer, illetve (b) tizedakkora lenne, mint a változóké. A csillapítási tényezőt –2 és 2 közé választottam, az alábbi függvényeket $\Delta\lambda = 0,01$-es lépésközzel számoltam.

Az ábrán látható, hogy abban az esetben, amikor a megfigyelések száma jóval kisebb, mint a változók száma, valamennyi, a látens változók számosságát becslő módszer rosszabbul teljesít. Így, nem vitatva *Jung–Marron*nak *(2009)* a főkomponens-elemzés kis számú megfigyelések esetén a főkomponens-elemzés konzisztenciájára vonatkozó eredményét, azt kell látnunk, hogy a látens változók számának becslése már kis zaj esetén is nehézkes, amennyiben a megfigyelések száma alacsony. Az is észrevehető, hogy kizárólag az általam javasolt hálózatalapú módszer (NDA) határozta meg minden esetben helyesen a látens változók számát. Ráadásul a módszer nemcsak jól határozta meg a látens változók számát, hanem helyesen is sorolta be azokat (lásd 4. ábra) a megfelelő csoportba. A 4 (a). ábra egy olyan esetet mutat, ahol a megfigyelések száma meghaladja a változókét, míg a 4 (b). ábra ennek fordítottja. Látható, hogy a modulokat ekkor is helyesen határoztam meg. Érdekes eredmény, hogy ha a távolságfüggvénynek a korreláció helyett a közvetett kapcsolatokat kiszűrő, nem szimmetrikus hasonlósági függvényt használom, akkor is helyes besorolást kapok (4 {c}. ábra).

3. ábra

Látens változók számának becslése különböző módszerekkel
Estimation of the number of latent variables



Mivel a becslő módszerek ritkán találták el a látens változók megfelelő számát, a besorolásra vonatkozó pontosságot csak akkor tudjuk meghatározni, ha a látensváltozó-számot rögzítjük. A főkomponens- és a faktorelemzés során egy változót ahhoz a látens változóhoz sorolok, amelyiknél az eredeti változó és a

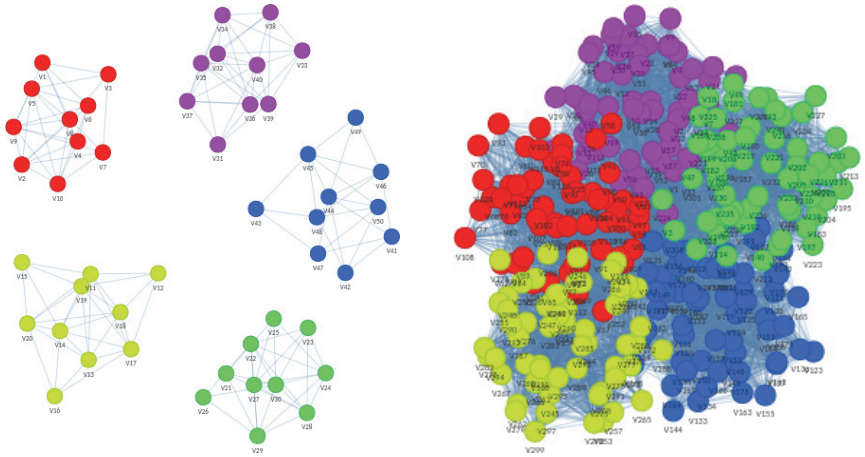
látens változó közötti korreláció a legnagyobb. A javasolt módszernél ilyen számításra nincs szükség, a modulok egyértelműen megadják, hogy mely változók (csomópontok) tartoznak az adott modulba. A 3 (a). ábrán bemutatott példában a megfigyelések száma tízszer nagyobb, mint a változóké, és a látens változók számát rögzítettem b=5-re, a zajtól függően a főkomponens-elemzés 75–83, a főfaktorelemzés 82–88%-ban adott helyes besorolást, szemben a javasolt módszerrel, amely 99,98%-ban hozott helyes besorolást.

4. ábra

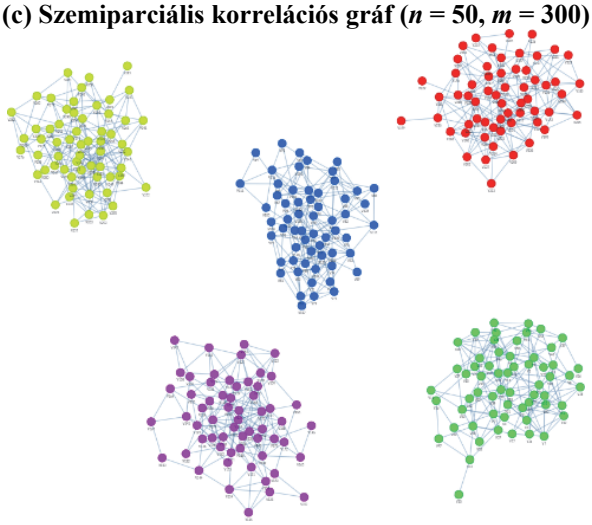
Modularizált korrelációs gráfok, B = 5, λ = –1 (GNDA)

Modules of correlation graphs

(a) Korrelációs gráf (n = 300, m = 50) **(b) Korrelációs gráf (n = 50, m = 300)**



(c) Szemiparciális korrelációs gráf (n = 50, m = 300)



Forrás: saját szerkesztés.

Az eredmények azt mutatják, hogy szintetikusan generált adatforrásoknál a javasolt hálózatalapú modellredukciós módszer valamennyi esetben helyesen határozta meg a látens változók számát, és a legtöbbször jól sorolta be az indikátorokat a megfelelő látensváltozókhoz. Alkalmazható volt továbbá nem szimmetrikus távolságfüggvényeket leíró irányított hasonlósági gráfoknál is. A módszer akkor is bevált, ha változók száma jelentősen meghaladta a megfigyelésekét. E tulajdonságok lehetővé teszik, hogy ne csak modell-, hanem adatredukció esetén is sikerrel alkalmazhassuk a javasolt módszert.

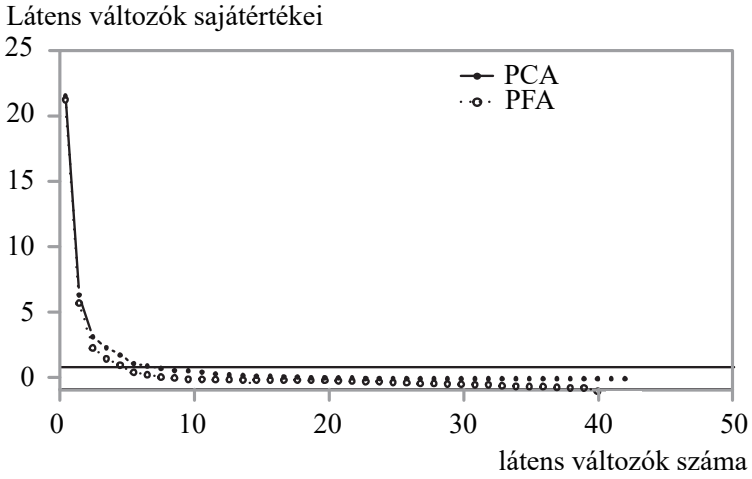
4. 2. Publikációs és kollaborációs tevékenységek komponensei

Ahogyan azt az előző fejezetben láthattuk, a főkomponens-, illetve a faktorelemzés használható olyan esetekben is, amikor a megfigyelések száma alacsonyabb, mint a változók száma (HDLSS-adatforrások), de ekkor nagyon nehéz megbecsülni mind a változók számát, mind biztosítani azt, hogy az indikátorok a megfelelő látensváltozókhoz kerüljenek. Ha pedig más, kifejezetten HDLSS-adatforrásokra kiterjesztett módszereket szeretnénk használni, még kevesebb támpontunk van a látens változók számának meghatározására. Éppen ezért a korábban említett *CWTS Leiden Ranking 2020* rangsort tekintem most példaként. Ekkor a főkomponens-elemzés 6, a főfaktorelemzés 4 látens változót azonosított a Kaiser-kritérium szerint (5. ábra). A további módszerek közül a leginkább javasolt PA- és MAP-módszerek 2 látens változót javasoltak, míg az általam javasolt NDA-módszer 3 látens változót azonosított. Az 5 (b). ábra egy ún. biplot diagramot mutat, ahol, ha 2 látens változót tekintünk, 2 dimenzióban láthatjuk, hogy az egyes indikátorok mely változóval korrelálnak leginkább. Ezek alapján kiderül, hogy az egyes látensváltozókhoz tartozó indikátorok köre meglehetősen vegyes, hiszen pl. az első látens változóhoz *(LV_1)* tartoznak kollaborációs mutatók *(PP_short_dist_collab),* nyílt hozzáférésű cikkek publikációs aránya *(PP_OA_unknown),* de található itt gendermutató *(PA_gender_unknown)* és a férfi/női arány is ***(PA_M_MF).*** A közös indikátorok e mutatók abszolút változói *(A_gender_unknown, P_OA_unknown, P_short_dist_collab).* A második látens változóhoz *(LV_2)* abszolút és relatív mutatók is tartoznak. Vegyesen találhatók itt publikációs és együttműködési mutatók is.

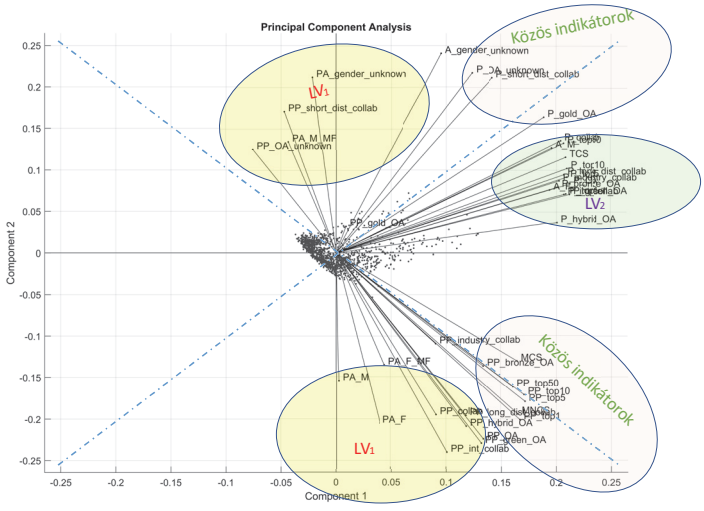
5. ábra

A *CWTS Leiden Ranking 2020* indikátorainak modellredukciója
Screen plot and PCA biplot of CWTS Leiden 2020 Ranking indicators

(a) Könyökdiagram



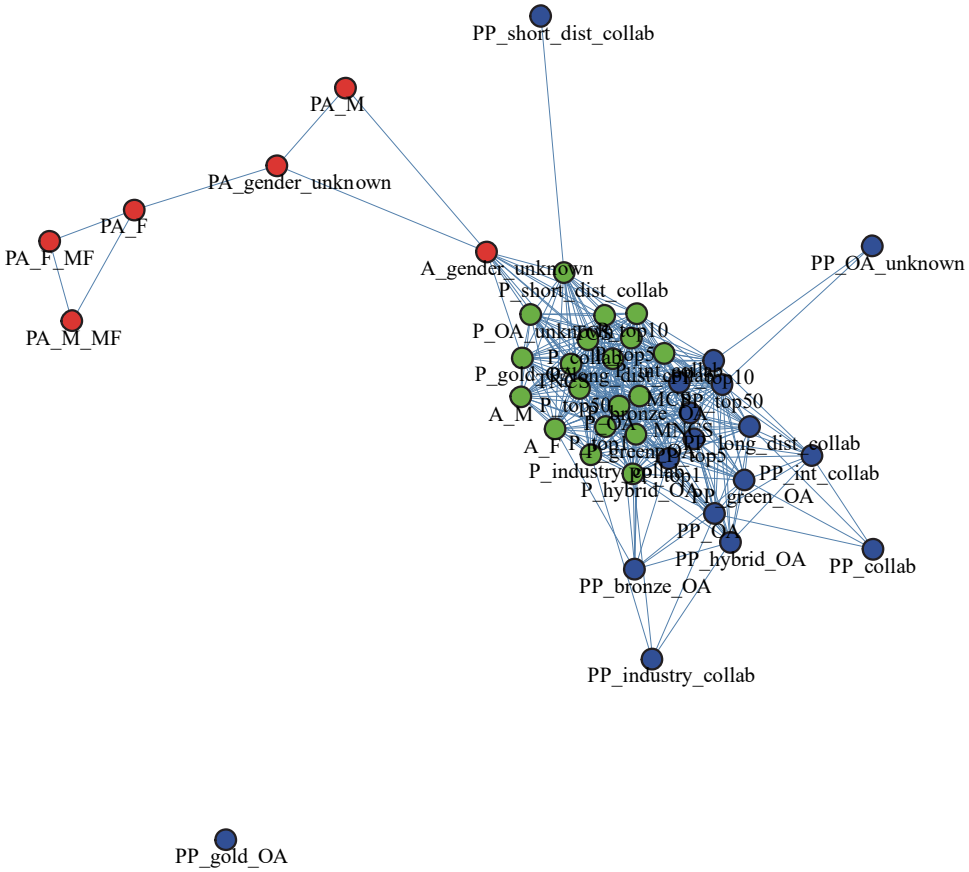
(b) Biplot diagram (PCA)



Forrás: Kosztyán et al. (2020a) alapján saját szerkesztés.

6. ábra

Hálózatalapú modellredukció alkalmazása
a *CWTS Leiden Ranking 2020* adattáblán
Modules of correlation graph of CWTS Leiden 2020 Ranking *indicators*

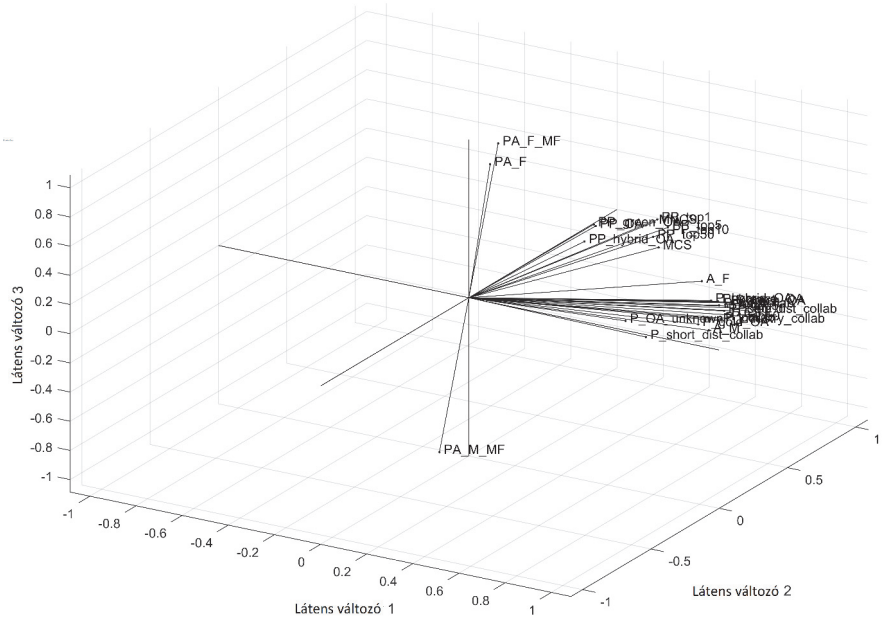


Amennyiben a javasolt NDA-módszert alkalmazom, három jól elkülöníthető csoportot kapok. Az egyik csoportba az abszolút mutatók, a másik csoportba a relatív mutatók tartoznak, és ezektől elkülönülten jelennek meg a gendermutatók. A 6. ábra mutatja az NDA-módszer alkalmazását, a korrelációs gráfstruktúrát. Az alacsony kommunalitású ($h_{min} = 0{,}2$) és a közös indikátorokat ($C_{min} = 0{,}2$) kiszűrve az alábbi struktúrát kapom (7. ábra).

7. ábra

Hálózatalapú modellredukció alkalmazása a *CWTS Leiden Ranking 2020* adattáblán, változószelekció alkalmazásával ($h_{min} = C_{mi} = 0{,}2$)

Biplot of NDA of CWTS Leiden Ranking 2020 *indicators*



Forrás: Kosztyán et al. (2020a) alapján saját szerkesztés.

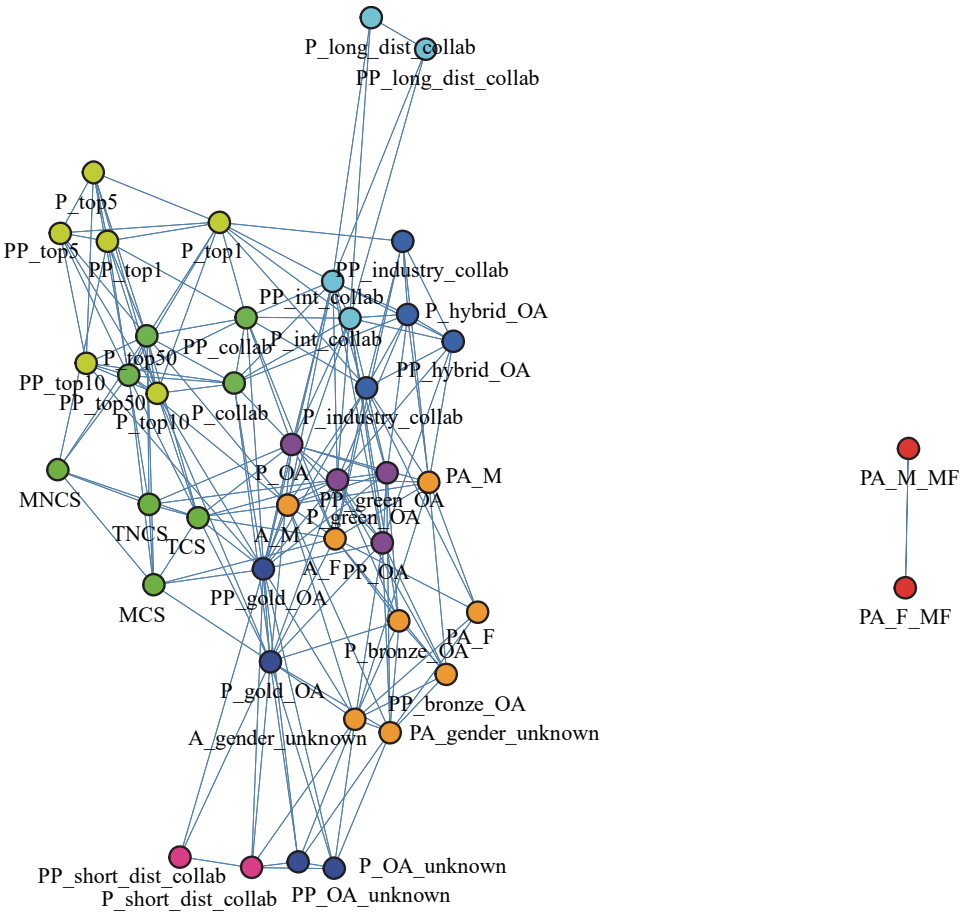
Ha a három látens változó szerint állítom sorba az intézményeket, egészen más sorrendek jönnek létre. Az első (abszolút mutatók) alapján meghatározott látens változók szerint a következő intézmények szerepelnek az első 5 helyen (zárójelben a szkórértékeket tüntettem fel): 1. Harvard University (12,76), 2. Stanford University (6,02), 3. University of Toronto (5,52), 4. University of Oxford (5,28), 5. University of Washington (4,79), vagyis egy angol kivételével amerikai, nagy presztízsű egyetemeket láthatunk. A második látens változó szerint, amely inkább a relatív mutatókat tartalmazta, az első 5 helyen angol felsőoktatási intézményeket találunk, ahol a lista első 2 helyén orvosi egyetemek szerepelnek: 1. London School of Hygiene & Tropical Medicine (3.71), 2. University of Cambridge (3.154918), 3. University of Glasgow (3.11), 4. University of Exeter (2.86), 5. University of East Anglia (2.85). A harmadik látens változó szerint – amely inkább a gendermutatókat, pl. a nők kutatásban, oktatásban, kollaborációban való arányát jelzi – az első 5 intézmény: 1. University of Stuttgart (2.59), 2. Shizuoka University (2.53), 3. University of Shahrood (2.53), 4. K.N.

Toosi University of Technology (2.49), 5. Jichi Medical University (2.48). Érdemes megjegyezni, hogy az utolsó indikátorcsoport szerinti előkelő hely a másik 2 mutatóban már nem ad magas pontértéket. A legtöbb itt felsorolt, de az első 20 intézményt is tekintve a másik 2 mutatóban a pontértékek negatívak, vagy 0-hoz közelítenek, ami egy standardizált érték esetében azt jelenti, hogy a harmadik látens változó szerint előre sorolt intézmények a másik két változócsoport tekintetében átlagosan, vagy az alatt teljesítenek.

8. ábra

Hálózatalapú modellredukció alkalmazása a Leiden-rangsoron *(CWTS Leiden Ranking 2020)* parciális korreláció alapú hasonlósági függvény alkalmazása esetén

Partial correlation graph of GNDA on CWTS Leiden Ranking 2020 *indicators*



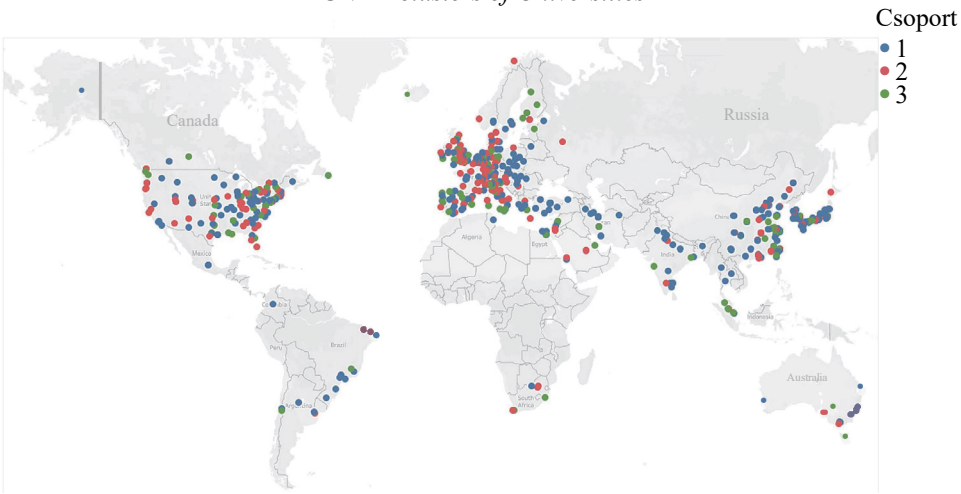
A kutatásnak nem volt célja, hogy a vállaltan nem végső sorrendet adó rangsornak én adjak végül komponensenkénti rangsorértelmezést, inkább csak demonstrálni szerettem volna a módszer alkalmazásának lehetőségeit. Amennyiben korreláció helyett parciális korrelációt használok hasonlósági függvénynek, a közvetett kapcsolatokat kiszűrve összesen 9 modult kapok. A legnagyobb modul a minőségi publikációs teljesítmény relatív és abszolút értéke, azaz kapok két, tisztán nyílt hozzáférésű publikációs teljesítményeket mérő, azaz egy viszonylag tiszta gender- és egy hivatkozásokkal kapcsolatos komponenst is (8. ábra).

A javasolt módszerem nemcsak változók, hanem adatok csoportosítására is használható. Lehet pl. a standardizált indikátorok szerinti euklideszi távolságok alapján klaszterezni, ekkor a teljesítmények alapján 3 klasztert kapok. A hasonló teljesítményértékek egy klaszterbe kerülnek (9. ábra).

9. ábra

Hálózatalapú adatredukció, standardizált indikátorok euklideszi távolságát alapul véve

GNDA clusters of Universities



A klaszterek értelmezéséhez segítséget nyújt, ha kiszámítom a klaszterek látens változókra vonatkozó átlagos pontértékét (1. táblázat).

1. táblázat

Csoportok szerinti pontértékek

GNDA scores within clusters

Csoport	NDA1	NDA2	NDA3
1	–0,25	–0,26	–0,02
2	0,59	0,51	0,05
3	0,01	0,14	–0,02

Az 1. táblázat megerősíti, hogy a gendermutatók nem játszanak szerepet a csoportképzésben, lényegében mindegyik 0, azaz átlag körül ingadozik. A másik két (abszolút és relatív) mutatót tekintve három jól elkülöníthető klaszter azonosítható. Az első klaszter átlag alatti, a második átlag feletti, a harmadik átlag közeli teljesítményt mutat. Az átlag feletti teljesítményt mutató intézmények zömében Nyugat Európában, az USA-ban és Kína keleti részén, illetve Ausztrália délkeleti részén találhatók. A többi klaszter földrajzi eloszlása meglehetősen diverz.

Fontos kiemelni, hogy ebben az esetben nem én állítottam be a klaszterek számát, hanem a modularitásvizsgálat eredményeként kaptam meg azokat.

5. A kifejlesztett programcsomag használata

E fejezet célja nem csupán annak lehetőségének bemutatása, hogy az itt ismertetett eredményeket bárki reprodukálhassa, hanem hogy ráirányítsam a figyelmet arra, hogyan lehet a javasolt alkalmazást a gyakorlatban is használni.

A javasolt hálózatelemzési módszer első változata *Kosztyán és szerzőtársai (2022a)* cikkében jelent meg. Ez alapján készült egy R- és egy Matlab-programcsomag. A Matlab-verzió még tesztelési fázisban van, de az R-csomag 0.1.6-os verziója már elérhető az R-programnyelvet kifejlesztő CRAN hivatalos oldaláról (https://cran.r-project.org/web/packages/nda/index.html). A csomag telepítése és a függvénykönyvtár beolvasása az

```
> install.packages("nda")
> library(nda)
```

parancsokkal történhet. Ez a csomag még nem tartalmazza a tetszőleges, pl. aszimmetrikus hasonlósági függvények kezelését. Ugyanakkor a kiterjesztéseket tartalmazó 0.1.9-es fejlesztői változat is elérhető már a GitHub-ról és fel is installálható a következőképpen:

```
> install.packages("devtools")
> devtools::install_github("kzst/nda")
> library(nda)
```

A 4.1. fejezetben bemutatott blokkmátrixok egyszerűen generálhatók a `data_gen` függvénnyel. Legyen példaként a megfigyelések száma: $n = 50$, a változók száma: $m = 500$, a blokkok, vagy látens változók száma: $b = 5$, a csillapítás: $\lambda = -1$.

```
> data_gen(50,500,5,-1)
```

A hálózatalapú modellredukció alkalmazásához az alábbi egyszerű parancsot használhatjuk, ahol `CWTS_2020` adattábla a függvénykönyvtár része.

```
> res<-ndr(CWTS_2020)
```

A módszer az eredményváltozóba gyűjti össze a látens változók számát (`res$factors`), a faktorszkórokat (`res$scores`), a faktorsúlyokat (`res$loadings`), a kommunalitásokat (`res$communality`) stb. Az eredményeket az alábbi függvénnyel is kiírathatjuk.

```
> summary(res)
```

A hasonlósági hálózatot a `plot` függvénnyel rajzolhatjuk ki. Ekkor egy olyan gráfot kapunk, ahol az élek a hasonlóságot, a csúcsok a változókat mutatják. A megjelenítéshez a Force Atlas II, iteratív heurisztikus eljárást alkalmazom, ami megpróbálja az ábra középpontjába helyezni a nagy centralitású csúcsokat, mintegy „tömeget" adva nekik. Ugyanakkor, ha kis hasonlóságú élek is szerepelnek az ábrában, ez a megjelenítés számításigényes, így a cuts paraméterrel az alacsony súlyú élek elhagyhatók (alapértelmezés szerint `cuts=0.3`). Az alapértelmezett megjelenítésre az alábbi példát vehetjük.

```
> plot(res)
```

Kirajzolható továbbá a biplot ábra is a `biplot` függvénnyel.

```
> biplot(res)
```

A hálózatalapú modellredukciós módszernek számos további beállítási lehetősége van, ami egyszerűen lekérdezhető:

```
> help(ndr)
```

A legújabb, fejlesztői verzióban már beépített módon, az alapértelmezett (`cor_type=1`), hagyományos (Pearson: `cor_method=1`, Spearman: `cor_method=2`, Kendall, `cor_method=3`, Distance: `cor_method=4`) korrelációk mellett lehet e korrelációk parciális (`cor_type=2`) és szemiparciális (`cor_type=3`) változatát is számolni. Ezen túlmenően tetszőleges távolságfüggvényt is használhatunk. Lássunk erre egy összetettebb példát, amelyben a módszert klaszterezésre használjuk, a `dist` függvénnyel számolunk alapértelmezés szerint euklideszi távolságot a standardizált változókon, amelyeket a `scale` függvénnyel kaphatunk meg. A hálózati adatredukciós módszernek a `covar=TRUE` paraméterrel jelezhetjük, hogy most nem beépített hasonlósággal, hanem egy hasonlósági mátrixszal számolunk, amit az `as.matrix` függvénnyel alakíthatjuk a megfelelő formára.

```
> clu<-ndr(scale(as.matrix(dist(CWTS_2020))
/max(scale(as.matrix(dist(CWTS_2020)),covar=TRUE)
```

6. Összefoglalás

A modell- és az adatredukciós módszerek már több mint száz éve a társadalomtudományi kutatások elengedhetetlen részei, alkalmazásuk a kérdőíves kutatásokban is megkerülhetetlen. Tanulmányomban kísérletet tettem arra, hogy egy új, hálózatalapú megközelítés segítségével kiterjesszem a modell- és az adatredukciós módszerek lehetőségeit. Úgy gondolom, hogy a hálózatalapú módszerek integrálása új lendületet adhat a redukciós módszerek nagy adathalmazokon való alkalmazásának. A módszer részletes tárgyalásán túl szintetikus és valós példákon is bemutattam a módszer alkalmazási lehetőségeit. A módszer alapján készült programcsomag forráskódját mindenki számára elérhetővé tettem, valamint egyszerű példákon keresztül megmutattam annak használatát is.

7. Korlátok, további fejlesztési lehetőségek

A javasolt módszer valamennyi lehetőségét terjedelmi okok miatt nem tudtam bemutatni. Ilyen pl. a modulok keresésére alkalmazható módszerek tárháza. Itt most csak az alapértelmezett Leiden-módszert ismertettem, de implementálva számos további módszer található, amelyek alkalmazása esetén némiképp eltérően kell interpretálni az eredményeket. Az alkalmazott modulkeresési eljárások közül a legtöbb többrétegű hálózatra is kiterjeszthető. Igaz ez a többi javasolt lépésre is, így lehetőség nyílna további dimenziók, pl. az idő figyelembevételére. Csupán utaltam rá, de ugyancsak terjedelmi okok miatt nem mutattam be részletesen az előszűrés során kapott klaszterhierarchiát. Itt, hasonlóan egy dendogramhoz, a modulok felbomlása miatt az előszűrés paraméterét növelve több kisebb modult kaphatunk, így vizsgálhatóvá válik a látens változók robusztussága. A modulok keresésénél a klasszikus konfigurációs modellből indultam ki, ahol egy nullmodell segítségével úgy optimálunk, hogy a modulon belül a kapcsolatok sűrűbbek legyenek, mint modulon kívül. Ugyanakkor ma már léteznek ennél sokkal szofisztikáltabb nullmodellek is, amelyek pl. a csomópontok földrajzi elhelyezkedését vagy gazdasági vonzóképességét is figyelembe veszik. Ha a csoportosítandó adatokhoz ilyen *a priori* információt is hozzárendelhetünk, akkor a modulokat ennek függvényében kereshetjük, így a módszert területi adatok vizsgálatára is kiterjeszthetjük. Ezeknek a kérdéseknek a részletes tárgyalására egy későbbi tanulmányban vállalkozom.

Melléklet

M1. táblázat

A *CWTS Leden Ranking 2020* indikátorai*
Employed indicator set of CWTS Leden Ranking 2020 *dataset*

Tudományos hatás	Az egyetem publikációinak száma. Ha itt a frac = 0, akkor = collab_P = oa_P					
	Az egyetem publikációira kapott hivatkozások (citációk)	száma				
			– szakterületre és a publikáció évére normalizálva.			
		átlaga				
			– szakterületre és a publikáció évére normalizálva.			
	Az egyetem publikációinak	száma	a szakterület azonos évében megjelent, a kapott hivatkozásaik alapján a felső	1	%-ba tartozó egyéb publikációkhoz képest.	
				5		
				10		
				50		
		aránya		1		
				5		
				10		
				50		
Együttműködési indikátorok	Az egyetem publikációinak száma. = oa_P = impact_P, ha az 1-es counting					
	Az egyetem olyan publikációinak	száma	, amelyek olyan társszerzőségben készültek, amelyek tagjai	1 vagy annál több	szervezettől	származnak.
		aránya				
		száma		2 vagy annál több	országból	
		aránya				
		száma		1 vagy annál több	ipari szervezettől	
		aránya				
		száma	, amelyek földrajzi együttműködési távolsága (ami egy publikáció esetén a legnagyobb földrajzi távolságot jelenti a publikációban említett két székhely között)	kevesebb mint 100 km.		
		aránya				
		száma		több mint 5000 km.		
		aránya				

(A táblázat folytatása a következő oldalon)

(folytatás)

Open access	Az egyetem publikációinak száma. = collab_P = impact_P, ha az 1-es counting					
	Az egyetem		open access publikációinak	száma		
				aránya		
		gold		száma	Ezek azok a publikációk, amelyek olyan folyóiratban jelentek meg, amely	open access.
				aránya		
		hibrid		száma		előfizetéses open access.
				aránya		
		bronz		száma		nyílt hozzáférésű open access.
				aránya		
		green		száma		open access repozitóriumban is elérhető.
				aránya		
		ismeretlen open access státuszú publikációk		száma	Ezeknek a publikációknak tipikusan nincs DOI-ja a Web of Science adatbázisban.	
				aránya		
Nemek	Az egyetem	összes	szerzőjének	száma	pl. egy ötszerzős publikációnál, ahol 3 szerző az X egyetemet jelöli meg az affiliációjában, kettő pedig az Y egyetemet. Ez a publikáció 3 szerzőséget jelent az X egyetemnek és 2-t az Y-nak.	
		férfi és női			, azaz az egyetem azon szerzőségeinek száma, ahol a szerzők neme ismert,	
		ismeretlen nemű				
		férfi				
		női				
		ismeretlen nemű		aránya	az összes szerzőséghez képest.	
		férfi				
		női				
		férfi női			a férfi és női szerzőségekhez képest.	

* Saját szerkesztés a Leiden-rangsor *(CWTS Leiden Ranking 2020)* indikátorai alapján.

Irodalom

Abonyi J. – Czvetkó T. – Kosztyán Zs. T. – Héberger K. (2022): Factor analysis, sparse PCA, and Sum of Ranking Differences-based improvements of the Promethee-GAIA multicriteria decision support technique. *Plos One.* Vol. 17. No. 2. e0264277. https://doi.org/10.1371/journal.pone.0264277

Abonyi J. – Ipkovich Á. – Dörgő Gy. – Héberger K. (2021): A leideni egyetemi rangsor több szempontú döntéselemzése. *XXXIV. Magyar Operációkutatási Konferencia*, Cegléd, 2021. augusztus 31. – 2021.szeptember 2. 26. o.

Aguirre, O. – Taboada, H. (2011): A clustering method based on dynamic self organizing trees for post-pareto optimality analysis. *Procedia Computer Science*. Vol. 6. pp. 195–200.

Aittokoski, T. – Äyrämö, S. – Miettinen, K. (2009): Clustering aided approach for decision making in computationally expensive multiobjective optimization. *Optimization Methods & Software.* Vol. 24. No. 2. pp. 157–174.

Banász Zs. – Csányi V. V. – Telcs A. – Kosztyán Zs. T. (2021): Hazai felsőoktatási intézmények a nemzetközi rangsorokban. *Educatio.* 29. évf. 3. sz. 495–508. o.

Barthélemy, M. (2011): Spatial networks. *Physics Reports*. Vol. 499. No. 1–3. pp. 1–101. https://doi.org/10.1016/j.physrep.2010.11.002

Bartlett, M. S. (1950): Tests of significance in factor analysis. *British Journal of Psychology*. Vol. 3. pp. 77–85.

Bartlett, M. S. (1951): A further note on tests of significance in factor analysis. *British Journal of Psychology*. Vol. 4. pp. 1–2.

Croux, C. – Filzmoser, P. – Fritz, H. (2013): Robust sparse principal component analysis. *Technometrics*. Vol. 55. No. 2. pp. 202–214.

Driver, H. E. – Kroeber, A. L. (1932): Quantitative Expression of Cultural Relationships. *University of California Publications in American Archaeology and Ethnology*. University of California Press. Quantitative Expression of Cultural Relationships: Berkley University Press. Vol. 31. No. 4. pp. 211–256.

Fauzi, M. A. – Tan, C. N. L. – Daud, M. – Awalludin, M. M. N. (2020): University rankings: A review of methodological flaws. *Issues in Educational Research*. Vol. 30. No. 1. pp. 79–96.

Gadár L. – Kosztyán, Zs. T. – Abonyi J. (2018): The Settlement Structure Is Reflected in Personal Investments: Distance-Dependent Network Modularity-Based Measurement of Regional Attractiveness. *Complexity*, Article ID 1306704. 1–16. https://doi.org/10.1155/2018/1306704

Gallegos, M. T. – Ritter, G. (2018): Probabilistic clustering via Pareto solutions and significance tests. *Advances in Data Analysis and Classification*. Vol. 12. No. 2. pp. 179–202.

Gorsuch, R. L. (1973): Using Bartlett's significance test to determine the number of factors to extract. *Educational and Psychological Measurement*. Vol. 33. No. 2. pp. 361–364. https://doi.org/10.1177/001316447303300216

García-Escudero, L. A. – Gordaliza, A. – Matrán, C. – Mayo-Iscar, A. (2010): A review of robust clustering methods. *Advances in Data Analysis and Classification*. Vol. 4. pp. 89–109.

Hair, J. F. J. – William, C. B. – Barry, J. B. – Rolph, E. A. (2020): *Multivariate Data Analysis* (*7th Edition*). Upper Saddle River, NJ: Prentice Hall.

Henning, C. (2015): Clustering strategy and method selection. *Handbook of cluster analysis*. Vol. 9. pp. 703–730.

Horn, J. L. (1965): A rationale and test for the number of factors in factor analysis. *Psychometrika*. Vol. 30. No. 2. pp. 179–185.https://doi.org/10.1007/BF02289447

Hu, J. – Pei, J. (2018): Subspace multi-clustering: A review. *Knowledge and Information Systems*. Vol. 56. No. 2. pp. 257–284. https://doi.org/10.1007/s10115-017-1110-9

Jacomy, M. – Venturini, T. – Heymann, S. – Bastian, M. (2014): ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PloS one. Vol. 9. No. 6. Article ID: e98679.

Jung, S. – Marron, J. S. (2009): PCA consistency in high dimension, low sample size context. *The Annals of Statistics*. Vol. 37. No. 6B. pp. 4104–4130. https://doi.org/10.1214/09-AOS709

Kosztyán Zs. T. – Kurbucz M. T. – Katona A. I. (2022a): Network-based dimensionality reduction of high-dimensional, low-sample-size datasets. *Knowledge-Based Systems*, Article ID: 109180. https://doi.org/10.1016/j.knosys.2022.109180

Kosztyán Zs. T. – Telcs A. – Abonyi J. (2022b): A multi-block clustering algorithm for high dimensional binarized sparse data. *Expert Systems with Applications*. Vol. 191. Article ID: 116219.

Li, Y. – Li, G. – Lian, H. – Tong, T. (2017): Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models. *Journal of Multivariate Analysis*. Vol. 155. pp. 133–150.

Mahmud, M. S. – Fu, X. – Huang, J. Z. – Masud, M. A. (2018): High-Dimensional Limited-Sample biomedical data classification using variational autoencoder. *Australasian Conference on Data Mining.* pp. 30–42. https://doi.org/10.1007/978-981-13-6661-1_3

Nagy M. – Molontay R. (2022): Network classification-based structural analysis of real networks and their model-generated counterparts. *Network Science*. Vol. 10. No. 2. pp. 146–169.

Nakayama, Y. – Yata, K. – Aoshima, M. (2021): Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings. *Journal of Multivariate Analysis*, Article ID: 104779. https://doi.org/10.1016/j.jmva.2021.104779

Newman, M. E. J. (2006): Modularity and community structure in networks. *Proceedings of the national academy of sciences*. Vol. 103. No. 23. pp. 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nunnally, J. C. – Bernstein, I. H. (1994): *Psychometric theory*. New York, McGraw-Hill.

Pearson, K. (1901): On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*. Vol. 2. No. 11. pp. 559–572. https://doi.org/10.1080/14786440109462720

Schölkopf, B. – Smola, A. – Müller, K. R. (1998): Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*. Vol. 10. No. 5. pp. 1299–1319. https://doi.org/10.1162/089976698300017467

Spearman, C. (1904): General intelligence objectively determined and measured. *American Journal of Psychology*. Vol. 15. No. 2. pp. 201–293. https://doi.org/10.2307/1412107

Székely G. J. – Rizzo, M. L. (2009): Brownian distance covariance. *The Annals of Applied Statistics*,Vol. 3. No. 4. pp. 1236–1265. https://doi.org/10.1214/09-AOAS312

Székely G. J. – Rizzo, M. L. (2013): The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*. Vol. 117. pp. 193–213. https://doi.org/10.1016/j.jmva.2013.02.012

Szüle B. (2019): Klaszterszám-meghatározási módszerek összehasonlítása. *Statisztikai Szemle.* 97. évf. 5. sz. 421–438. o.

Traag, V. A. – Waltman, L. – van Eck, N. J. (2019): From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*. Vol. 9. Article ID: 5233. https://doi.org/10.1038/s41598-019-41695-z

Tran, U. S. – Formann, A. K. (2009): Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement*. Vol. 69. pp. 50–61.

Velicer, W. F. (1976): Determining the number of components from the matrix of partial correlations. *Psychometrika*. Vol. 41. No. 3. pp. 321–327. https://doi.org/10.1007/BF02293557

Velicer, W. F. – Eaton, C. A. – Fava, J. L. (2000): Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. *Problems and solutions in human assessment* pp. 41–71. https://doi.org/10.1007/978-1-4615-4397-8_3

Wasim, M. – Brereton, R. G. (2004): Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy. *Chemometrics and intelligent laboratory systems*. Vol. 72. No. 2. pp. 133–151. https://doi.org/10.1016/j.chemolab.2004.01.008

Wilkinson, L. – Friendly, M. (2009): The history of the cluster heat map. *The American Statistician*. Vol. 63. No. 2. pp. 179–184.

Zelditch, M. L. – Goswami, A. (2021): What does modularity mean? *Evolution & Development*. Vol. 23. No. 5. pp. 377–403.