A depresszió környezeti faktorainak vizsgálata oksági elemzési módszerekkel

Vetró Mihály[1], Dr. Hullám Gábor[1], Dr. Juhász Gabriella[2], Dr. Antal Péter[1]
[1]BME, Méréstechnika és Információs Rendszerek Tanszék,
hullam.gabor@mit.bme.hu
1117 Budapest, Magyar Tudósok Körútja 2.
[2]Semmelweis Egyetem, Gyógyszerhatástani Intézet,
juhasz.gabriella@pharma.semmelweis-univ.hu
1089 Budapest, Nagyvárad tér 4.

Összefoglaló: **Kutatómunkánk során globális és lokális oksági feltáró algoritmusokat alkalmazunk a depresszióhoz kapcsolódó környezeti és egyéb tényezőket közötti oksági kapcsolatok azonosítására.**

Bevezető

A depresszió nagyszámú genetikai és környezeti tényező komplex kölcsönhatásának eredménye. Ebből adódóan a mai napig kihívást jelent a hozzájáruló környezeti és egyéb tényezők, úgymint a stressz, a negatív életesemények, a szocioökonómiai és életmódbeli tényezők, a testmozgás, a mentális egészségi állapot leírók együttes elemzése, különösen a jelenségek mögött meghúzódó mechanizmusok megértése, új oksági összefüggések feltárása. Ennek elsődleges oka, hogy a megfigyelési adatokon alapuló vizsgálatoknál a változók közötti függőségi mintázatok oksági értelmezése korlátozott, legtöbbször csak megfelelő háttértudás, például időbeliséget meghatározó információ birtokában lehetséges. Más esetekben az oksági kapcsolatok egy részének azonosítására van csak lehetőség.

Célkitűzés

Az itt bemutatott munka célja összehasonlítani különböző oksági-függőségi struktúra tanuló algoritmusok eredményeit a depresszióhoz kapcsolódó környezeti faktorok oksági kapcsolatainak elemzése során. Az oksági kapcsolatok feltárásának célja a depresszió és más multifaktoriális betegségek elemzése során az, hogy a betegséget direkt módon befolyásoló faktorokat azonosíthassuk.

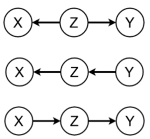
Módszer

A probléma kezelésére számos módszert hoztak létre, melyeket csoportosíthatunk megközelítés szerint globális vagy lokális szemléletű algoritmusokra. Az előbbiek a teljes oksági-függőségi struktúrát kívánják

rekonstruálni a rendelkezésre álló adatok alapján, míg az utóbbiak az oksági-függőségi struktúra kisméretű lokális egységeit vizsgálják, és ezek együtteséből következtetnek a függőségek rendszerére.

A **lokális módszerek** közül a *lokális oksági feltárás* algoritmus (*local causal discovery - LCD*) [1] egy módosított változatát implementáltuk, amely hatékonyan alkalmazható nagy dimenziójú adatok esetén is, ami orvosbiológiai tárgyterületeken gyakori. Ugyanakkor lokális jellege miatt a többváltozós feltételes függőségek egy kis részét veszi csak figyelembe, emiatt csak közelíteni képes a valós függőségi mintázatokat. A módszer alapját egy függőségi (függetlenségi) teszt adja, amely diszkrét változókból álló adathalmaz esetén lehet a Khí-négyzet (χ^2) próba [2]; esetünkben egy empirikusan választott $\alpha = 0{,}001$ szignifikancia küszöb és Bonferroni korrekció [3] alkalmazása mellett. Ennek segítségével meghatároztuk az összes változó páronkénti függetlenségét, majd a nem-független változóknak megfelelő csomópontpárokat összekötve kialakítottuk a reprezentációs modellként szolgáló Bayes-háló vázát [4]. Ezt követően e váz tripletein, azaz három csomópontból és a köztük futó két élből álló alstruktúráin végig iterálva azonosítottuk az ezekben található két él orientációját, ahol ez lehetséges volt. Ahogyan az 1. Táblázatban is látható, ha az adott triplet (*X*, *Z*, *Y*) két legtávolabb eső csomópontjához tartozó változója (*X* és *Y*) marginálisan független ($X \perp Y$), és a közbülső változó (*Z*) ismeretében feltételes függőséget mutat $\neg(X \perp Y \mid Z)$, akkor a három változó egyértelműen „V-struktúrát" alkot [4]. Ettől eltérő esetben azonban a másik három lehetséges él-elrendezés bármelyike előfordulhat feltéve, hogy a két szélső változó feltétel nélküli függést mutat $\neg(X \perp Y)$, a köztes változó ismeretében azonban feltételesen függetlenek $(X \perp Y \mid Z)$ (tehát a közbülső változó „D-szeparálja" a két szélsőt [4]).

1. táblázat: Lehetséges háromelemű irányított struktúrák a változók függetlenségének ismeretében.

Feltétel	Struktúra
$(X \perp Y), \neg(X \perp Y \mid Z),$ $\neg(X \perp Z), \neg(Y \perp Z)$	
$\neg(X \perp Y), (X \perp Y \mid Z),$ $\neg(X \perp Z), \neg(Y \perp Z)$	

Emiatt a „közös szülő” és a két lehetséges irányú „lánc” struktúra egymáshoz képest megfigyelési ekvivalensek. Ebből kiindulva az LCD módszernél első lépésként detektáljuk az összes lehetséges V-struktúrát ($X \rightarrow Z \leftarrow Y$), majd megkeressük az olyan háromelemű láncokat ($X \rightarrow Z — W$), amelyek sorrend szerinti első éle ($X \rightarrow Z$) része egy V-struktúrának, amelyből a másik él is egyértelműen irányítható ($Z \rightarrow W$). Ezen lánckeresést ismételten futtatjuk addig, amíg a legutóbbi futtatás új beillesztett élt eredményez. A megmaradt irányítatlan éleknél, melyek egyik csomópontjából egy irányított él vezet tovább ($U — Z \rightarrow W$), „közös szülő” struktúra meglétét vizsgáljuk ($U \leftarrow Z \rightarrow W$), feltéve, hogy ezáltal nem hoz létre új V-struktúrát. Végül a hamis pozitív találatok szűrése érdekében töröljük a háromelemű lánc struktúrák ($X \rightarrow Z \rightarrow Y$) első és utolsó elemeit összekötő tranzitív éleket ($X \rightarrow Y$) feltéve, hogy az X és Y csomópontot a lánc középső változója (Z) D-szeparálja.

A **globális** oksági-függőségi struktúra tanuló **algoritmusok** közül egy olyan bayesi modellátlagoláson alapuló módszert alkalmaztunk [5], amely a lehetséges struktúrák terében történő véletlen bolyongást végez egy Markov-lánc Monte Carlo (MCMC) technikára épülő módszerrel [6]. Ennek a folyamatnak az eredménye egy irányított körmentes gráf (DAG) struktúra halmaz (az MCMC mintavételezés által nyert minták), amely felhasználható különféle egyszerűbb strukturális tulajdonságok, például irányított élek meglétének, vagy összetettebb tulajdonságok, például Markov-takarók kiértékelésére. Ezek az összegyűjtött minták használhatók fel az egyes tulajdonságok *a posteriori* valószínűségeinek becsléséhez. Az ezen az elven alapuló rendszeralapú módszertant - melynek fejlesztése a *BME Méréstechnika és Információs Rendszerek Tanszékén,* a *Számítógépes Biomedicina és Bioinformatika Munkacsoportban* zajlott - bayesi többszintű relevancia analízisnek (*Bayesian Multi-Level Analysis of relevance – BMLA*) nevezzük [7]. Ezt a módszert korábban több kutatás vizsgálataiban alkalmaztuk a *Semmelweis Egyetem Gyógyszerhatástani Intézetében* működő *Új Antidepresszív Gyógyszercélpont Kutatócsoporttal* együttműködésben a depressziót befolyásoló genetikai variánsok [8], illetve a környezeti-életviteli faktorok relevanciájának, oksági-függőségi mintázatainak vizsgálatára [9].

Mindezek mellett lényeges, hogy bár mind a lokális, mind a globális módszerek a legtöbb esetben elfogadható eredményt adnak a függőségi struktúrát illetően, oksági értelmezésük feltételekhez kötött. Ezek alapja többek közt az oksági Markov-feltétel [10], az oksági hűség (faithfulness), a szükségesség (sufficiency), a stabilitási feltétel, torzításmentességi feltétel

és az elégséges minta. Ezek teljesülésekor kijelenthetjük, hogy minden függőségi gráfból kiolvasható függőség az együttes valószínűség-eloszlásnak megfelelő valós függőséget reprezentál és nem szerepel benne más.

Eredmények

A lokális módszer (LCD) hatékonyságát megvizsgáltuk 10.000, 20.000, 50.000, 100.000 és 150.000 mintapontból álló mesterséges mintahalmazokon, melyek mindegyikéből 25 darabot állítottunk elő, mindegyiket egy-egy önálló, 60 csomópontos, véletlenszerűen kialakított Bayes-hálóból mintavételezve. Ezek alapján megállapítottuk, hogy a módszer alacsony mintaszám (10.000) mellett átlagosan az élek 55%-át találta meg, viszont a jósolt élek 85%-a helyes. Magasabb mintaszám mellett az eredeti modell éleinek 60%-át képes megtalálni, emellett azonban az összes jósolt él mindössze 60-70%-a lesz helyes. Végül fontos megfigyelésnek bizonyult, hogy mintaszámtól függetlenül a jósolt élek mindössze 10%-a fordított (tehát az él helye az eredeti modellel konzisztens, iránya viszont helytelen), így a függőségek irányát elfogadható megbízhatósággal képes detektálni a módszer.

Korábbi kutatómunkánk folyamán a depresszió kialakulásával kapcsolatos faktorok összefüggésrendszerét tártuk fel egy bayesi függőségi térkép segítségével, amely egy több mint 110 000 mintát tartalmazó adathalmazon alapult [9]. Az ezen az adathalmazon – bayesi relevancia analízissel [7] – azonosított többváltozós modelleket vetjük össze a lokális oksági kapcsolatokat feltáró módszer hasonló eredményeivel. Ahogyan a 2. táblázat: A lokális feltáráson alapuló módszer (LCD) eredményeinek összehasonlítása a globális bayesi módszerrel (BMLA) a vizsgált adaton.-ban is látható, az LCD módszernél a tranzitív függőséget reprezentáló élek utólagos törlése jelentősen csökkenti a hamis pozitív találatok (FP) számát, és ezáltal növeli a módszer specifikusságát (TNR - True Negative Rate) a globális bayesi módszer eredményét referenciaként alkalmazva.

2. táblázat: A lokális feltáráson alapuló módszer (LCD) eredményeinek összehasonlítása a globális bayesi módszerrel (BMLA) a vizsgált adaton.

	Alapeset				Tranzitív élek szűrése			
Küszöb	**TP**	**TPR**	**TN**	**TNR**	**TP**	**TPR**	**TN**	**TNR**
0.5	118	0,544	1911	0.669	114	0,525	2302	0.806
0.2	226	0,483	1768	0.679	218	0,466	2155	0.828
0.1	230	0,481	1762	0.679	220	0,46	2147	0.828

A 3. táblázat: A lokális feltáráson alapuló módszer (LCD) eredményeinek összehasonlítása a globális, rendszeralapú, bayesi módszer (BMLA) által becsült élvalószínűségekkel, a depresszióhoz kötődő 10 legvalószínűbb élre. mutatja a lokális oksági feltárás révén előálló, depresszió leíró változóhoz kapcsolódó élek azon halmazát, melyek a globális, bayesi módszer szerint a legvalószínűbbek. Az eredmények szerint a neuroticizmus személyiségjegy (*neuroticism*), a jelenlegi depresszív tünetek (*current depression*) és a testzsír-arány (*body fat*) áll közvetlen oksági kapcsolatban a depresszióval, mely kapcsolatokat mind a lokális, mind a globális módszerek alátámasztják. A negatív életesemények (*life stress*) depresszióval való kapcsolatát szintén detektálja mindkét módszer, bár a globális szerint kisebb valószínűséggel van jelen közvetlen oksági él. A táblázatban található további hat él azonban csak az LCD szerint közvetlen kapcsolat.

3. táblázat: A lokális feltáráson alapuló módszer (LCD) eredményeinek összehasonlítása a globális, rendszeralapú, bayesi módszer (BMLA) által becsült élvalószínűségekkel, a depresszióhoz kötődő 10 legvalószínűbb élre.

Feltárt él (LCD)	**Élvaló-színűség (BMLA)**	**Feltárt él (LCD)**	**Élvaló-színűség (BMLA)**
Neuroticism → depression	9.99E-01	Back pain → depression	1.88E-04
Current depression → depression	9.99E-01	Vigorous physical activity→ depression	1.73E-04
Body fat → depression	8.00E-01	Headache → depression	1.73E-04
Life stress → depression	2.00E-01	Alcohol intake →depression	8.97E-05
Falls → depression	3.50E-04	Insomnia → depression	6.81E-05

Következtetések

A kiértékelés folyamán három lényeges szempontot kell figyelembe vennünk. (1) A megfigyelési ekvivalencia miatt egyes ok-okozati kapcsolatok egyértelműen nem azonosíthatóak, így irányítottságuk esetleges. Ezek az esetleges irányítású élek hatékonyan megkereshetőek [11], és ennek figyelembevételével a bayesi modellátlagolás során a globális módszer pontosabbá válik. (2) A bayesi függőségi térkép modellátlagolás révén jött létre, tehát nem egy lehetséges modellt ír le, hanem számos modellt. Az egyes élek előfordulási gyakoriságát a lehetséges modellekben egy a posteriori valószínűség reprezentálja, amihez nincs rögzített elfogadási küszöb. Egy él akár mindkét irányban előfordulhat a modellekben valamekkora valószínűséggel. Ezzel szemben a lokális okságsági feltárás egy lehetséges modellt azonosít a

függetlenségi/függőségi vizsgálathoz kapcsolódó küszöbnek megfelelően. (3) A lokális módszerek nem vesznek figyelembe magasabb rendű függőségi viszonyokat, így sűrűn összekötött oksági-függőségi struktúrák esetén a kapcsolatok egy részét nem képesek azonosítani. Mindezek ismeretében elfogadható a különbség a lokális és globális módszerek eredményei között, különösen, hogy az eredeti munka legrelevánsabb eredményeit megerősíti.

Köszönetnyilvánítás

A kutatómunka az Innovációs és Technológiai Minisztérium Tématerületi Kiválósági Programjának (TKP2020, BME), valamint az ÚNKP-20-5 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával, továbbá az Országos Tudományos Kutatási Alapprogramok (OTKA-119866) támogatásával és a Bolyai János Kutatási Ösztöndíj segítségével valósult meg.

Hivatkozások

[1] S. Mani és G. F. Cooper, A Study in Causal Discovery from Population-Based Infant Birth and Death Records, Pittsburgh, 1999.

[2] A. Agresti, An Introduction to Categorical Data Analysis, Wiley, 2007.

[3] „Bonferroni Correction," [Online]. Available: https://mathworld.wolfram.com/BonferroniCorrection.html.

[4] J. Pearl, Causality: Models, Reasoning and Inference, C. U. P., 2000.

[5] D. Madigan, S. A. Andersson, M. Perlman és C. T. Volinsky, „Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs," in *Comm. Statist. Theory Methods*, 1996, p. 2493–2520.

[6] N. Friedman és D. Koller, „Being Bayesian about network structure," in *Mach. Learn.*, 2003, p. 95–125.

[7] P. Antal, A. Millinghoffer, G. Hullám, C. Szalai és A. Falus, „A bayesian view of challenges in feature selection: feature aggregation, multiple targets, redundancy and interaction," *FSDM,* pp. 74-89, 2008.

[8] G. Juhasz, Brain galanin system genes interact with life stresses in depression-related phenotypes, 2014.

[9] G. Hullam, P. Antal és P. Petschner, „The UKB envirome of depression: from interactions to synergistic effects," *Sci Rep,* 2019.

[10] J. Pearl, „Causal inference in statistics: An overview," *Statistics Surveys,* pp. 96-146, 2009.

[11] M. C. David és M. Christopher, „Selective Greedy Equivalence Search: Finding Optimal Bayesian Networks Using a Polynomial Number of Score Evaluations,” 2015.