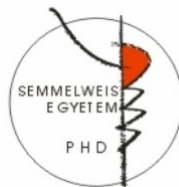


KLASZTEREZÉSTŐL A KAUZALITÁS- ANALÍZISIG: JELFELDOLGOZÓ MÓDSZEREK ALKALMAZÁSA ÉS FEJLESZTÉSE IDEGTUDOMÁNYI KONTEXTUSBAN

Tézisek

Benkő Zsigmond

Semmelweis Egyetem
Szentágotthai János Idegtudományi Doktori Iskola



Témavezető: Somogyvári Zoltán, Ph.D, Tudományos főmunkatárs

Hivatalos bírálók: Kerepesi Csaba, Ph.D, Tudományos munkatárs
Dr. Ráczy Frigyes Sámuel, Ph.D, Egyetemi tanársegéd

Szigorlati bizottság elnöke: Dr. Kamondi Anita, D.Sc, Egyetemi tanár
Szigorlati bizottság tagjai: Mészéna Géza, D.Sc., Egyetemi tanár
Barthó Péter, PhD, Tudományos főmunkatárs

Budapest
2020

1. Bevezetés

Az idegrendszeri mérésekből származó adatok elemzése összetett feladat, ezért minden új adat-elemzési módszer lehetőségeket nyit meg az agy anatómiájának és működésének felderítésében.

A disszertációban egy adatelemzési utat járok be: egy klasszikus módszert alkalmazok neuroanatómiai metszeteken és új módszereket fejleszték ki dinamikus idegi jelenségeket vizsgálatára. Ehhez felhasználom a valószínűségszámítás és a nemlineáris dinamikai rendszerek elméletének eszköztárát.

Először retinális metszeteket vizsgálok nem-felügyelt tanulási megközelítésből. Majd egy új anomália-detekciós eljárást fejleszték ki, hogy alvási apnoét, gravitációs hullámot és benfentes kereskedést azonosítsak vele. Ezek után egy sokaságdimenzió becslő módszert fejleszték tovább, hogy epilepsziás rohamok dinamikáját térképezem fel. Végül pedig megvizsgálom az idegszövetben mérhető optikai és elektromos jelek közötti ok-okozati összefüggéseket.

2. Célkitűzések

2.1. Csapsejt szinaptikus talpak nem felügyelt klaszterezése

Az emberi retina külső szinaptikus rétegében a csapsejtek réskapcsolatokkal is kapcsolódnak egymáshoz és a bipoláris sejtekhez. A dolgozatban megpróbálunk statisztikailag robusztus strukturát találni csapsejt-talp metszetekben. E feladathoz lognormális keverék eloszlást illesztünk az adatokra, az optimális modellt Bayes-i információs kritérium segítségével választjuk meg.

2.2. Ritkaság alapú anomália-detekció

Az anomáliák ritkán fordulnak elő és a megszokottól különböző mintapontokat generálnak az adatsorban. A legtöbb detektáló-algoritmus az utóbbi kritériumot – a különbözőséget – használja ki, hogy beazonosítsa a nem megszokott eseményeket.

A dolgozatban egy olyan új detektáló algoritmust fejlesztünk ki, mely az anomáliák ritkaságát használja ki. Az új módszert Temporális Outlier Faktornak (TOF) kereszteljük el, majd összehasonlítjuk teljesítményét egy sűrűség-alapú anomália-detektáló algoritmussal, a Loká-

lis Outlier Faktorról (LOF). Megvizsgáljuk a TOF viselkedését valóvilágbeli adatsorokon, smegállpítjuk, hogy ez a ritkaság-alapú algoritmus hozzáad-e többlet-információt az eddigiekhez képest, avagy sem.

2.3. Sokaságdimenzió-becslés 2.0

Az adatsorok sokaság-dimenziója információval szolgál az adat komplexitásáról, és megbízható ismerete sokszor elengedhetetlen feltétele az adatok sikeres elemzésnek.

A dolgozatban kiszámoljuk a Farahmand-Szepesvari-Audibert dimenzióbecslő valószínűségi sűrűségét – azzal a céllal – hogy megbízhatóbb dimenzióbecsléseket adhassunk a jövőben. Ezek után – a sűrűségfüggvény ismeretében – megkíséreljük a maximum likelihood megoldás levezetését, majd körüljárunk pár – a globális becslést érintő – fejlesztési lehetőséget. Összemérjük a módosított becslő algoritmus teljesítményét a klasszikus Levina-Bickel és DANCo algoritmusok teljesítményével. Végül alkalmazzuk az új algoritmust, hogy kirajzoljuk segítségével az epileptikus dinamika térképét intrakraniális EEG felvételeken.

2.4. Optikai és elektromos jel közötti összefüggés feltárása egér "élő" agyszeletben

Az agyszövetben mérhető optikai jel a szövet fényáteresztő és elnyelési tulajdonságainak változásait követi. Ez főképp a sejtek méretének – a közeg ozmotikus ingadozásai következtében megjelenő – változásai miatt jön létre. Ebből fakadóan az optikai jel az elektromos potenciál bonyolult módon összefüggésben vannak egymással, de ez a viszony gyakorlatban még feltáratlan.

A dolgozatban a konvergens keresztleképezés – kauzalitás-elemző – módszert használjuk fel arra, hogy megvizsgáljuk az optikai jel és a lokális mezőpotenciál kapcsolatát rágcsáló "élő" in vitro agyszeleteken kiváltott epileptikus aktivitás közben. Összevetjük a kapott eredményeket az egyszerű keresztkorrelációs függvény által kapott eredményekkel és levonjuk az összehasonlítás tanulságait. Végül egy empirikus modellben sűrítjük össze az analízis során megfigyelteteket, ha elég tapasztalat akkumulálódott a vizsgálatok során.

3. Módszerek

3.1. Retinális réskapcsolatok statisztikai elemzése

A csap talp terület, konvex burkoló és konnexin36 plakk szám/talpmetszet eloszlásokat lognormális keverékeloszlásokkal közelítettük.

A modellt a várakozás-maximalizációs algoritmussal illesztettük az adatokra. A modell-paramétereket random inicializáltuk és 800 lépést iteráltuk, ezt 20-szor végeztük el és a legjobb eredményt fogadtuk el. Az előbbi folyamatot különböző klaszterszámmal megismételtük (1 – 9). Az optimális klaszterszámot Bayesi információs kritériuma alapján választottuk ki.

3.2. A TOF analízis lépései

1. Előfeldolgozás és alkalmazhatóság ellenőrzése:

Ez a lépés estről esetre változhat, az adat típusától függ, hogy milyen előfeldolgozási lépések szükségesek ahhoz, hogy alkalmazni lehessen a módszert. Általában ajánlott az adatsort stacionerré alakítani. Például a jelben található legalacsonyabb frekvenciájú komponensnek is több periódussal kell szerepelnie. Ha ez nem teljesül, akkor Fourier szűrést lehet alkalmazni a túl alacsony frekvenciájú komponensek eltávolítására.

2. Beágyazás időeltolással:

Az idősort beágyazzuk egy magasabb dimenziós térbe az időeltolások módszerével. E módszernek két paramétere a beágyazási dimenzió illetve az időeltolás. A beágyazási paraméterek értéke előzetes tudás, illetve optimalizációs módszerekkel állítható be. Ilyen optimalizációs technika az eltolás meghatározására az autokorrelációs függvény első nullapontja, a hamis legközelebbi szomszéd módszer, avagy a differenciálenrópia módszer.

3. k - legközelebbi szomszéd keresése:

Megkeressük minden adatpont k legközelebbi szomszédját az állapotterben és elmentjük a szomszédok távolságát és időindexét.

4. TOF érték kiszámítása:

Kiszámítjuk a TOF értékeit az alábbi egyenletnek megfelelően:

$$\text{TOF}(t) = \sqrt[q]{\frac{\sum_{i=1}^k |t - t_i|^q}{k}}. \quad (1)$$

Ahol t az aktuális minta időpontja, t_i a minta állapotterbeli szomszédjának időpontja, k a szomszédok száma és $q = 2$ egy tetszőleges konstans.

5. Küszöb alkalmazása a TOF értékeken:

A küszöb értéke előzetes tudás, klaszterezés, vagy felügyelt tanulás segítségével állítható be. Előzetes tudás alapján, a maximális eseményhossz paraméter segítségével beállítható a küszöb értéke, ahol az 1 egyenletbe a különbség helyére a eseményhosszt helyettesítjük be. A küszöbölés után opcionálisan egy $-k/2$ mintapontnak megfelelő nagyságú – kiszélesítés alkalmazható a detektált pontok körül szimmetrikusan.

3.3. FSA algoritmus és bias-korrekción

3.3.1. Az FSA és mFSA algoritmus

Az FSA algoritmus során először a lokális dimenzióbecsléseket végezzük el az adat-térben:

$$d(x) = \frac{\ln(2)}{\ln(R_{2k}/R_k)} \quad (2)$$

Ahol $d(x)$ a lokális becslés x helyen, R_k a k . szomszéd távolsága x -től.

Majd második lépésként a minta-átlagot, avagy a módot vesszük, hogy megkapjuk a globális becslést. Az mFSA algoritmus a minta mediánt használja a globális dimenzió becslésére.

3.3.2. A cmFSA algoritmus

Hiperkocka addatszeteket szimuláltunk különböző sokaság-dimenzióval (2–30) és mintamérettel ($n = 10, 100, 1000, 2500, 10000$), minden adatszettből $N = 100$ realizációt generálva. Kiszámítottuk az mFSA értékeket minimális szomszédságméretre ($k = 1$) és periódikus peremfeltételre, valamint nem-periódikus peremfeltételre.

Egy exponenciális korrekciós formulát illesztettünk a relatív hibára, így kapva pontosabb dimenzióbecslést az eddieknél.

$$\hat{d}_c = \hat{d} \exp\left(\sum \alpha_i \hat{d}^i\right) \quad (3)$$

Az illesztés egy koefficienssel is jó illeszkedést mutatott a 2–30 dimenzió-intervallumon, azonban 2 – 80 intervallumon több koefficiensre volt szükség a pontos illesztéshez. Az illesztést legkisebb négyzetek módszerével és ortogonális távolság illesztéssel is elvégeztük a 100 realizációt felhasználva.

3.4. A konvergens kereszt-leképezés

A konvergens kereszt-leképezés két idősor között vizsgálja a kauzalitási reláció létét és minőségét. Kezdőlépése az időeltolásos beágyazás:

$$X(t) = [X(t - D\tau), \dots, X(t), \dots, X(t + D\tau)] \quad (4)$$

Ahol $x(t)$ az eredeti idősor és $X(t)$ az rekonstruált állapotterbeli adatsor.

Majd az algoritmus a beágyazott idősorokat (X és Y) felhasználva vizsgálja meg, hogy létezik-e keresztleképezés a terek között. Ahhoz, hogy az $X \rightarrow Y$ irányú kapcsolatot teszteljünk, kikeressük Y -ban az összes pont $2D + 2$ (beágyazási dimenzió plusz 1) legközelebbi szomszédját, és exponenciális kernelt használva súlyokat rendelünk az adott időindexű szomszédokhoz az alábbi formula szerint:

$$W_i(t) = \frac{e^{-\frac{|Y(t_i) - Y(t)|}{|Y(t_1) - Y(t)|}}}{\sum_i^{E+1} e^{-\frac{|Y(t_i) - Y(t)|}{|Y(t_1) - Y(t)|}}} \quad (5)$$

ahol, $|Y(t_i) - Y(t)|$ az Euklideszi távolság i . legközelebbi szomszéd és az eredeti pont között és $W_i(t)$ az adott időponthoz és szomszédhoz tartozó súly.

Az időindexek (t_i) és a súlyok ($W_i(t)$) felhasználásával egy-egy becült értéket számolhatunk a másik idősorhoz:

$$\hat{x}(t) = \sum_i W_i(t) * x(t_i) \quad (6)$$

Végül a becült és az eredeti idősor közötti korrelációs koefficiens szám-szerűsíti, hogy mennyire jó a leképezés az adott irányba:

$$CCM_{X \rightarrow Y} = r_{x, \hat{x}} \quad (7)$$

Ye és munkatársai kiterjesztették a fenti procedúrát időben késleltetett hatások kimutatására is: ha az idősorokat eltoljuk egymáshoz képest – a keresztkorrelációs függvény kiszámításához hasonlóan – akkor konvergens kereszt-leképezési függvényhez juthatunk.

4. Eredmények

1. **Csoportok elkülönítése csap szinaptikus talpak metszetein.**
 - 1.1. A konnexin36 réskapcsolatok vizsgálatával foglalkozó projekt részeként megvizsgáltam az emberi retina külső szinaptikus rétegében készült metszeteken a csap talpak eloszlását. Az eloszlást lognormális keverékmodellel modelleztem és Bayesi információs kritériummal megállapítottam az optimális klaszterszámot. Megutattam, hogy a sejttalpak két fő szubpopulációra oszthatók a vizsgált talp terület, talp konvex burkoló terület, és konnexin36 plakk szám/talp-metszet tulajdonságok alapján.
2. **Az Időbeli Anomália Faktor kifejlesztése egyedi események detektálására.** Az anomáliák detektálása kritikus feladat számos területen, és különösen nagy kihívást jelent, ha a detektálandó esemény nem ismert.
 - 2.1. Bevezettük az "egyszarvú" avagy egyedi esemény anomália-konceptiót és egy modellfüggetlen algoritmust a detektálásra. A koncepció nagyban különbözik az egyszerű outlier fogalmától: az ismétlődő anomáliák nem egyedi események többé, valamint egy egyedi esemény nem feltétlenül lóg ki a normális adateloszlásból.
 - 2.2. Megalkottuk a TOF algoritmust, mellyel az idősorokban bekövetkező események egyedi voltát lehet mérni. A módszer teljesítményét szimulált adatokon vizsgáltuk, valamint összehasonlítást végeztünk a LOF módszerrel. A TOF módszer jobb teljesítményt mutatott, mint a LOF módszer a szimulált adatokon: hagyományos outliereket is megtalált, valamint olyan egyszarvúakat, melyek nem minősülnek outliernek.
 - 2.3. Igazi adatsorokon bemutattam, hogy milyen előnyökkel bír az egyszarvú koncepció és a TOF algoritmus. Sikeresen demonstráltam a TOF módszer hasznát olyan esetekben, amikor ismert volt az anomália: gravitációs hullámot azonosítottam a LIGO detektor adataiban és alvási apnoé-t detektáltam EKG adatsoron. Ezeken kívül alkalmaztam a módszert

teljesen nem-felügyelt felállásban: a London InterBank Offered Rate adatsoron találtam új egyedi eseményeket.

3. **A korrigált medián FSA dimenzióbecslő algoritmus.** A dimenzionalitás az adatok komplexitásának mértéke. A dolgozatban továbbfejlesztettem az FSA sokaság-dimenzió becslő algoritmust ezáltal az élvonalba emelve azt.
 - 3.1. Kiszámoltam a lokális becslések valószínűségi sűrűségfüggvényét a lokális pontsűrűséget konstansnak véve. A sűrűségfüggvény alapján a lokális becslések minta-mediánját javasoltam globális dimenzióbecslésnek az eddig javasolt módszerrel és meannel szemben. Továbbá a sűrűségfüggvény alapján felírtam a maximum likelihood globális becslést, ha a minták függetlennek tekinthetők.
 - 3.2. A szélhatást és a véges minta effektust hiperkocka adatszetteken kalibrált exponenciális korrekciós-formulával orvosoltam. Az így kapott algoritmus teljesítményét az irodalomban javasolt teszt-adatszetteken vettem össze a Levina-Bickel illetve a DANCó módszer teljesítményével. Megmutattam, hogy a kmFSA algoritmus megveri az LB módszert és DANCóval egyenlő félként teljesít átlagos hibaszázalék és hiba ráta mértékek alapján.
 - 3.3. Az mFSA algoritmus segítségével szerteágazó változásokat mutattam ki az idegi dinamikában nyugalmi állapotban és epileptikus rohamok közben intrakraniális EEG felvételeken. Azonosítottam olyan csatornákat, melyeken – a többi csatornához viszonyítva – alacsony-dimenziójú dinamika volt megfigyelhető, ezek lehetséges okozói és kiindulási pontjai a rohamoknak.
4. **A saját optikai jel és az elektromos aktivitás közötti kauzális viszony.** Górcső alá vettük egér túlélő agyszövetekben a lokális mező-potenciál és a belső optikai jel kapcsolatát a konvergens keresztlekepezés módszerével.
 - 4.1. A vizsgálatok során az optikai jelen két komponenst különítettünk el egymástól: egy gyors aktivitásfüggő komponenst és egy lassú komponenst. A gyors komponens különböző előjellel jelent meg az áteresztési és visszaverődési jelen ezáltal a szövet fényvisszaverő és fényszórási tulajdonságával

áll kapcsolatban. A lassú komponens mind az áteresztett, mind a visszavert jelen negatív előjellel megjelenő alapvonal-eltolódásként manifesztálódott, tehát a szövet fényelnyelésével kapcsolatos.

- 4.2. A konvergens keresztleképezést alkalmazva azt találtuk, hogy a lokális mezőpotenciál erős egyirányú hatással van az optikai jel gyors komponensére 0.5-1 s időkéssel, visszacsatolását nem figyeltük meg. A az optikai jel és a mezőpotenciál keresztkorrelációs függvényén található csúcsok nem voltak összeegyeztethetőek az előző megfigyeléssel, azonban a optikai jel deriváltja és a mezőpotenciál közti keresztkorrelációs függvény szintén egyirányú kapcsolatot sugallt ugyanekkora időkéssel.
- 4.3. Az előbbi megfigyeléseket egy empirikus modellbe építettük, amelyben egy differenciálegyenlet segítségével modelleztük a mezőpotenciál és az optikai jel közötti kapcsolatot.

5. Következtetések

A dolgozat statikus neuroanatómiai adatok nem-felügyelt analizésétől indulva, a nemlineáris dinamikai rendszerek elméletének idősorokon való alkalmazásán keresztül jut el, agy-szeleteken mért optikai és elektromos jelek kapcsolatának feltárásáig.

Az emberi retina egy rendkívül összetett és jól szervezett anatómiai struktúra, mely a fényérzékelésért felelős és a vizuális feldolgozás első lépéseit is elvégzi. E jelfeldolgozásban játszanak fontos, de nem teljes egészében feltárt szerepet a rés-kapcsolatok és ezen kapcsolatokat alkotó konnexinek, például a connexin36. A réskapcsolatok és konnexin fehérjék anatómiai leírása segítheti az általuk betöltött funkciók azonosítását. Lognormális keverék eloszlások illesztésével és modell-szelekció alkalmazásával klasztereztük a sejt talpak eloszlását az emberi retina első szinaptikus rétegében készült metszeteken. Azt tapasztaltuk, hogy két fő talpacska-típus különíthető el a talpméret, konvex burkoló méret és végződésenkénti connexin36 szám alapján.

Az idősorokban fellelhető anomáliák ritka és a szokványostól eltérő mintázatok, melyek észlelése nagy gyakorlati jelentőséggel bírhat. Azonban ezen ritka események detektálása korántsem egyszerű feladat – különösen akkor – ha nem sztereotíp jelenségeket észlelnénk. A dolgozat ezen részében kifejlesztettünk egy anomália detektáló módszert, ami extrém ritka egyedi események – úgynevezett unikornisok – megtalálását teszi lehetővé. Megmutattuk, hogy az algoritmus és unikornis-konceptió többlet értéket ad a tradicionális anomália - detekció irodalmához és demonstráltuk a módszer potenciális gyakorlati hasznát EKG, gravitációs hullám és LIBOR adatsorokon.

A dimenzió az adatok komplexitásának mértéke, utal a megfigyelt változók közötti redundanciákra és kapcsolatokra. Hogy ezen információhoz hozzáférjünk, a dolgozatban a Farahmand - Szepesvári - Audibert dimenzióbecslő algoritmust vizsgáltuk meg és fejlesztettük tovább. Kiszámoltuk a lokális becslések valószínűsége-sűrűségfüggvényét és megállapítottuk, hogy az eloszlás mediánja megegyezik a dimenzióval. A véges mintaszám és a szél-hatás okozta torzítást korrigálni igyekeztünk a becslés relatív hibájára illesztett exponenciális formulával, melynek kalibrációját véletlenmintavételezett hyperkockákon végeztük. A korrekció egy szintre emelte az algoritmus teljesítményét DANCo-éval, mely

az egyik legmegbízhatóbb dimenzióbecslő algoritmus. A korrigált medián FSA módszert ezután epilepsziás betegből származó intrakraniális mezőpotenciál adatokon alkalmaztuk és azt találtuk, hogy a rohamok alatt a jobb temporális kéreg mutatta a legalacsonyabb dimenziójú dinamikát, mely azt sejteti, hogy a terület rohamindulásért felelős zóna része.

Megvizsgáltuk az egér ex vivo agyszeleteinek lokális mező potenciálját és intrinsic optikai jelét indukált epilepsziás rohamok során. Megállapítottuk, hogy az epilepsziás aktivitás alatt reflektált optikai jel egy lassú és gyors komponensből áll. A lassú komponens egy negatív alapvonal-eltolódás, míg a gyors komponens csökkenő negatív hullámok formájában jelenik meg az epilepsziás kisülések alatt, amit konvergáló relaxáció követ a kisülésmentes szakaszokon. A későbbiekben a késleltetett konvergens keresztleképezési módszert alkalmaztuk a mezőpotenciál és az optikai jel közötti ok-okozati összefüggések feltárására. Kiderült, hogy az mezőpotenciál hajtja meg az optikai jelet, és ez a hatás hozzávetőlegesen fél másodperc időkéssel jelenik meg. Úgy találtuk, hogy a mezőpotenciál és az optikai jel deriváltjának keresztkorrelációs függvénye csúcsot mutat fél másodperces késleltetésnél, ami a kapcsolat lineáris voltát sugallja. Ezen megfigyelések alapján egy empirikus differenciálegyenletben modelleztük a mezőpotenciál és az optikai jel kapcsolatát.

6. Saját publikációk

Kapcsolódó közlemények

Kántor O, Benkő Zs, Énzsöly A, Dávid Cs, Naumann A, Nitschke R, Szabó A, Pálfi E, Orbán J, Nyitrai M, Németh J, Szél Á, Lukáts Á, and Völgyi B. (2016) Characterization of connexin36 gap junctions in the human outer retina. *Brain Structure and Function*, 2216:2963–2984

Benkő Zs, Moldován K, Szádeczky-Kardoss K, Zalányi L, Borbély S, Világi I, and Somogyvári Z. (2019) Causal relationship between local field potential and intrinsic optical signal in epileptiform activity in vitro. *Scientific Reports*, 91:5171

Benkő Zs, Fabó D, and Somogyvári Z. Time Series and Interactions: Data Processing in Epilepsy Research. In P Érdi, B Sen Bhattacharya, and A Cochran (editors), *Computational Neurology and Psychiatry*, Springer, Cham, 2017: 73–91.

Benkő Zs, Bábel T, and Somogyvári Z. (2020) How to find a unicorn: a novel model-free, unsupervised anomaly detection method for time series. *arXiv*

Benkő Zs, Stippinger M, Rehus R, Bencze A, Fabó D, Hajnal B, Erőss L, Telcs A, and Somogyvári Z. (2020) Manifold-adaptive dimension estimation revisited. *arXiv*

Egyéb közlemények

Kántor O, Szarka G, Benkő Zs, Somogyvári Z, Pálfi E, Baksa G, Rác G, Nitschke R, Debertain G, and Völgyi B. (2018) Strategic Positioning of Connexin36 Gap Junctions Across Human Retinal Ganglion Cell Dendritic Arbors. *Frontiers in Cellular Neuroscience*

Benkő Zs, Zlatniczki Á, Stippinger M, Fabó D, Sólyom A, Erőss L, Telcs A, and Somogyvári Z. (2018) Complete Inference of Causal Relations between Dynamical Systems. *arXiv*