K 106154

Strukturális hatások keresztosztályozott adatokban

A kutatás fő célja változók közötti struktúrák értelmezésére és adatok alapján történő tesztelésére vonatkozó eljárások fejlesztése volt. Ezek a struktúrák vagy előre kijelölt változók hatásai más, szintén előre kijelölt változókra vagy pedig felcserélhető szerepű változók közötti szimmetrikusnak tekinthető kapcsolatok. E közül a két összefüggő, de eltérő kérdés közül az első különös fontosságú az oksági vizsgálatokkal kapcsolatban az elmúlt két évtizedben keletkezett új eredmények, és az irántuk megnövekedett érdeklődés miatt. A második kérdés pedig azért különösen aktuális, mert az adatelemzés, ezen belül a szociológiai adatelemzés korábban nem vizsgált struktúrájú és nagyságú adatállományokkal dolgozik napjainkban.

Ugyan a kutatás kategoriális adatokra, azaz többszempontos csoportosításokra vonatkozik, az eredmények jelentős része nemcsak ilyen, hanem folytonos adatokra is érvényes. A kategoriális környezetben azonban sok strukturális kérdés jobban látszik, mintha az adatokat folytonosnak tételeznénk fel. A folytonos eloszlások esetén szinte automatikusan feltételezett többváltozós normalitás a magasabb rendű interakciók hiánya miatt sok strukturális kapcsolatot eleve kizár, míg a kategoriális környezet lehetővé teszi ezek vizsgálatát.

A strukturális hatások elmélete szorosan kapcsolódik az okság modern elméleteihez, amelyek közül mind az elsősorban Rubin munkásságához kapcsolódó kontrafaktuális elméleten alapuló propensity score alapú illesztés, mind pedig a Pearl nevével fémjelzett manipuláció alapú grafikus modellezés releváns. Mindkét elmélet korlátjának tűnik, hogy a Simpson paradoxon jelenségére nem adnak kielégítő magyarázatot, és csak részben tudják megválaszolni, hogy ennek előfordulása esetén melyik kezelés hatékonyabb. A Simpson paradoxon általános formában azt jelenti, hogy két kezelés hatékonyságának összehasonlításában az adatokat valamilyen szempont szerint szétbontva, minden keletkező csoportban az egyik kezelés hatékonyabb, de az adatok aggregálása után a másik kezelés tűnik jobbnak. Tehát például az új gyógyszer férfiakra is és nőkre is jobb, mint a régi, de az összes embert tekintve mégis a régi gyógyszer a jobb. Ami meglehetősen nehezen értékelhető következtetés, különös tekintettel arra, hogy nemcsak felmérések, de másféle adatgyűjtési eljárások esetén s előfordulhat. Szélsőséges esetben a Simpson paradoxon előfordulása még tervezett kísérletek esetén is lehetséges, ami azért meglepő, mert a klasszikus elmélet szerint a tervezett kísérlet az az adatgyűjtési eljárás, amely alapján oksági következtetések levonhatóak.

A Simpson paradoxon nemcsak kategoriális adatokkal fordul elő, de a strukturális kérdés jobban vizsgálható ebben az esetben. Folytonos adatok esetén jól ismert, és a Simpson paradoxonnal rokon jelenség, hogy egy lineáris regresszió elemzésben egy magyarázó változó együtthatója, sőt

még annak előjele is megváltozhat, attól függően, hogy milyen más magyarázó változók vannak jelen a regressziós egyenletben.

A Simpson paradoxonnal foglalkozó irodalom olyan kiterjedt, hogy itt még csak felületes áttekintésére sincs lehetőség. A létező irodalom túlnyomó többsége a paradoxon előfordulását vagy az adatok hibájának, vagy az adatok gyűjtésére szolgáló eljárás hibájának tudja be.

A két kezelés közül a jobbik kiválasztására szolgáló eljárás mindkét kezelés esetében a pozitív és negatív kimenetelek számának hányadosát, azaz a pozitív kimenetel feltételes esélyét nézi, és azt a kezelést tekinti jobbnak, amelynek esetében a feltételes esély nagyobb. Ez az eljárás ekvivalens az esélyhányados vagy más néven a keresztszorzat hányados (CPR) használatával.

A 2010-ben a *Statistical Methodology* c. folyóiratban megjelent cikkem volt az első, amely a paradoxon előfordulását – a korábban meg nem kérdőjelezett – CPR használatához kötötte. Kimutatta, hogy a következtetés tulajdonságaira vonatkozó egyszerű feltételek mellett csak egyetlen olyan döntési eljárás van, amelyik semmilyen adatok esetén nem követi el a paradoxont. Ez a döntés nem a pozitív és negatív kimenetelek arányát, hanem különbségét hasonlítja össze a két kezelés esetén és ekvivalens a keresztkülönbség hányados (CSR) alapú döntéssel.

A CPR és CSR közötti különbség kicsinek látszik, de viselkedésük között alapvető eltérés van. Nemcsak a Simpson paradoxon előfordulása szempontjából különböznek, de a CSR értéke nem függ attól, hogy az egyes kezeléseket hányan kapták, míg a CSR értéke függ ettől. Az előbbi tehát nem alkalmas felmérésekből származó adatokra, amelyek esetében az, hogy hányan választották a kezelést hordoz információt, az utóbbi nem alkalmas kísérleti adatokra, ahol a kezelési csoportokba történt allokáció nem tartalmaz információt.

Az OTKA kutatás támogatásának felhasználásával ezekről a kérdésekről adtam elő a 6th European Congress of Methodology és a XIV Congreso de Metodología de las Ciencias Sociales y de la Salud, plenáris előadójaként. Az előadások kibővített anyagát közölte a *Methodology* c. folyóirat is. Ebben a cikkben úgy érvelek, hogy a paradoxon okának az a tény tekinthető, hogy ugyanazzal az eljárással próbálunk válaszolni igazából különböző kérdésekre. Sem a CPR sem a CCSR nem ajánlható általánosan, és a jobb kezelés kiválasztását célzó eljárásnak az értékelés számos körülményére tekintettel kell lennie az adatgyűjtés módján túlmenően a pozitív és negatív kimenetelek valódi jelentésére, de a meghozandó szakpolitikai döntés jellegére (a jobbnak tűnő kezelés kötelező lesz-e vagy választható) is.

A háttérben lévő matematikai eredményeket a *Statistical Methodology* c. folyóiratban általánosítottam több változóra. Ezek az eredmények a Simpson paradoxon többdimenziós változatának elkerülésére alkalmasak. Ez például az a helyzet lenne, amelyben ugyan a régi kezelés jobb, mint az új, de azért mind férfiakra, mind nőkre az új kezelés jobb, de ha ezeket a csoportokat tovább bontjuk életkor szerint, megint mindben a régi kezelés tűnik jobbnak, stb. A cikk egyúttal bemutatja a többdimenziós kontingencia táblázatnak egy olyan paraméterezését, amelyből a hatások iránya kiolvasható, és egy változó vagy változócsoport hatásának iránya nem

függ attól, hogy még hány másik változót veszünk figyelembe, azaz a paradoxon soha nem fordul elő. Ráadásul azt is bizonyítottam, hogy lényegében csak egyetlen ilyen paraméterezés lehetséges.

Ehhez kapcsolódóan Németh Renáta konferencia előadása kategoriális esetre alkalmazta a primer oksági hatás és a valódi oksági hatás eltérésének szokásos dekompozícióját. A primér oksági hatás a kezeltek és nem kezeltek válaszainak eltérése, de ezt nemcsak a két kezelés különbözősége, hanem a két kezelt csoport eltérő összetétele, sőt a kezelésekre vonatkozó eltérő reakciója miatt is lehet. Az említett dekompozíció éppen ezt a két hatást veszi figyelembe. Az eredmények azt is mutatják, hogy még az irodalomban hagyományosan oksági hatások megállapítására javasolt tervezett kísérleti elrendezés is csak határértékben alkalmas az oksági hatások megállapítására.

A *Computational Statistics and Data Analysis* c. folyóiratban társszerzőkkel megjelent cikk azt vizsgálja, hogy az adatok kategoriális jellegének, amely – szemben a többváltozós normális eloszlással --, magasabb rendű interakciót is lehetővé tesz, milyen hatása van a Pearl-féle oksági elméletben központi szerepet játszó gráfokra vonatkozó adatalapú kereső eljárásokra. A fő eredmény az, hogy helyesen gráfok helyett hipergráfok között kellene a keresési algoritmus alkalmazni, és bizonyos esetekre becsléseket tudtunk adni a helyes eljárások teljesítményére vonatkozóan.

Milibák Eszter szakdolgozata az oksági elemzések Rubin-féle elméletéhez kapcsolódik, és azt vizsgálja, hogy a gyakorlatban a propensity score becslésére legtöbbször használt logisztikus regressziós eljárások eredményei mennyire tekinthetőek stabilnak. A dolgozat számos problémára fényt vetett, amelyek további elemzést igényelnek. Annyi biztosnak látszik, hogy mindkét oksági elmélet sikeres alkalmazásához szükséges az a feltevés, hogy valamennyi hatással rendelkező változóról rendelkezésre állnak megfigyelések, enélkül az eredmények meglehetősen esetlegesek lesznek.

A kutatás másik részét Anna Klimovával együttműködve végeztük. Ez a munka a több kategoriális változó közötti asszociációs vagy hatás struktúra elemzésére használatos loglineáris modellek osztályát bővítette ki. A kibővítés szükségessége több okra vezethető vissza. A loglineáris modellek olyan kontingencia táblákra alkalmazhatóak, amelyekben bármely változó bármely kategóriája előfordulhat bármely más változók bármely kategória kombinációjával. Az ilyem struktúrákat Descartes szorzatnak nevezik. Azonban a szociológia számos lényeges problémája nem írható le ilyen struktúrákkal.

Például egy 3x3 méretű apa-fiú mobilitási táblában megkérdezhetjük, hogy a fiú társadalmi pozíciója független-e az apáétól. (Természetesen a legtöbb országban nem független, de a függetlenség itt nem valóban érdekes kérdésként, hanem más, hasonló, de reálisabb struktúrák legegyszerűbb példájaként szerepel.) Ha azonban a táblázatot átírjuk úgy, hogy az apa pozíciója az egyik változó, és a fiú mobilitása (felfelé, lefelé, immobil) a másik, akkor már nem kapunk

egy 3x3 méretű táblázatot, mert magas pozíciójú apának nem lehet felfelé mobil fia és alacsony pozíciójú apának nem lehet lefelé mobil fia. Ugyan a függetlenség kérdése ebben az esetben is felvethető, a létező módszerekkel nem megválaszolható, igazából nem is értelmezhető pontosan.

A relációs modellek ezekre a nem teljes Descartes szorzatokra is kiterjesztik a loglineáris modell fogalmát. További általánosítás a hagyományos loglineáris modellekhez képest, hogy a modell által megengedett hatások nem feltétlenül a változók valamilyen részhalmazához, hanem a táblázat celláinak tetszőleges tulajdonsággal definiált részhalmazához kapcsolódatnak. Ennek illusztrálására tekintsünk egy kereszttáblát, amely a férj és feleség iskolai végzettségét, valamint azt tartalmazza, hogy a férj tervez-e állást változtatni az elkövetkező évben. Itt előfordulhat, hogy az állásváltoztatási szándékára sem a férj, sem a feleség iskolai végzettsége, sem pedig ezeknek hagyományos értelemben vett interakciója nincs hatással, viszont azokban a cellákban, amelyekbe a férj iskolai végzettsége azonos a feleségével, a változtatás szándéka nagyobb, míg eltérő iskolai végzettség esetén kisebb.

Érdekes technikai következmény, hogy sok ilyen modell esetén, nem tehető fel egy minden cellában jelen lévő közös hatás létezése a modell megváltoztatása nélkül. Ezekben az esetekben a modell és a becslések tulajdonságai igen meglepőek lesznek.

A relációs modelleket még az OTKA kutatás megkezdése előtt egy a *Journal of Multivariate Analysis* c. folyóiratban vezettük be, és egy másik, a *Journal of Applied Statistics* c. folyóiratban mutattunk be egy mobilitási elemzést relációs modelleket felhasználva.

Az OTKA kutatás keretében ennek a munkának a folytatásaként csomagot publikáltunk R programozási környezetben a becslési eljárások végrehajtására. Ez a létező algoritmusok nagymértékű általánosítását kívánta meg.

A becslési eljárás elméleti hátterét, az iteráció konvergenciáját és a paraméterbecslések létezésére vonatkozó bizonyításokat egy a *Scandinavian Journal of Statistics* c. folyóiratban publikált cikkben jelentettük meg. Az elmélet az általánosan használt iteratív arányos illesztési eljárás olyan kiterjesztésén alapul, amely a modellel duális lineáris családokba nem a szokásos Kullback-Leibler divergencia minimalizálásával vetít, hanem az ezt kiterjesztő, általunk definiált Bregman divergencia szerint.

Munkánk következő kérdése az volt, hogy a becslési eljárás hogyan terjeszthető ki arra a gyakorlatban sokszor előforduló esetre, amikor a megfigyelt cellagyakoriságok között nullák is vannak. Annak érdekében, hogy ezekben az esetekben is találjunk az adatok likelihoodját maximalizáló becslést, maguk a modellek is általánosításra szorultak. Mindenesetre bizonyítható volt, hogy az eredeti becslési algoritmus ebben az általánosabb esetben is helyesen működik. Ezek az eredmények a *Journal of Multivariate Analysis c.* folyóiratban jeletek meg.

A relációs modellekkel való közvetlen munka utolsó állomásaként a modellek adatalapú tesztelésének kérdéseivel foglalkoztunk. Hagyományos loglineáris modellek esetén az illeszkedés

vizsgálata általában aszimptotikusan khi-négyzet eloszlású statisztikákkal történik. Ezek azonban a becslések eltérő természete miatt nem alkalmazhatóak automatikusan relációs modellekre. Elemzésünk eredménye az, hogy a Pearson statisztika ugyanúgy használható, de a likelihood hányados statisztikát módosítani kell, méghozzá éppen azzal a taggal, amely a Kullback-Leibler divergencia és a Bregman divergencia különbsége. Az erről szóló kézirat jelenleg lektorálás alatt áll.

A relációs modellek kifejlesztésének egyik célja a táblázat üres celláinak pontosabb kezelése volt. Hagyományosan, véletlen vagy mintavételi és strukturális nullákat szoktak megkülönböztetni. Az előbbiek csak a mintavételből származó ingadozások, esetleg a kis mintanagyság miatt nem tartalmaznak megfigyelést, az utóbbiak viszont a populációban is üresek. A relációs modellek elméletében háromféle üres cellát különböztetünk meg. Logikailag lehetetlen, logikailag lehetséges, de az adott populációban üres, az adott populációban nem üres, de a mintában nem megfigyelt. Mivel ebben a megközelítésben nem teljes Descartes szorzatokon megfigyelt adatok is modellezhetőek, a logikailag lehetetlen megfigyeléseknek megfelelő cellák kihagyásra kerülhetnek.

A hiányzó adatok struktúrájának elemzéséhez tartozik az utolsó tevékenység, amely a kutatás keretében folyt, és eredményei még nincsenek publikálva. Gyakorlatilag minden felmérésből származó adatállomány hiányos, és az elemzés többnyire a Missing at random vagy Missing completely at random feltevések alapján zajlik. Az előbbi feltevés lényege, hogy az összes többi változó értékének rögzítése mellett, azaz feltételesen, az, hogy egy változót megfigyeltünk-e, független a változó tényleges értékétől. Az utóbbi feltevés ezt feltétel nélkül állítja. Ezek a feltevések lényegében azt eredményezik, hogy a részlegesen megfigyelt adatállomány elemzése ugyanazokat az eredményeket adja, mintha a teljes adatállományt tudnánk elemezni. Ezért ezek a feltevések nagyon hasznosak, ugyanakkor helyességüket a ténylegesen megfigyelt adatokból nem lehet ellenőrizni.

A Missing at random és Missing completely at random feltevések csak egyetlen változóra vonatkoznak és a marginális modellek elméletének felhasználásával, amely kutatási irányt 2002-ben kezdeményeztük Wicher Bergsmaval, sokféle általánosításuk definiálható. Természetesen ezek sem tesztelhetők a megfigyelt adatokból, de sokféleségük talán alkalmas arra, hogy a felhasználókat gondolkodásra késztesse, mielőtt automatikusan élnének a MaR vagy a MCaR feltevéssel. Például két változó esetén egy lehetséges általánosított modell azt tételezi fel, hogy a két változó megfigyeltsége egymástól független, az egyik MCaR, a másik MaR. Ez nyilvánvalóan egy marginális modell, és az ilyen modellek összes jó tulajdonsága automatikusan teljesül rá nézve. Ezeket az eredményeket a jövőben fogom publikálni, az OTKA támogatásra való hivatkozással.

Összességében a kutatás eredményeiből 5 impakt faktoros publikáció született, ezekből 4 Q1, 1 pedig Q2 besorolású, továbbá 3 plenáris és 9 meghívott előadás nemzetközi konferenciákon.

Írott publikációk

Klimova, A., Rudas, T. (2013): Iterative Scaling in Curved Exponential Families (arXiv: 1307.3282)

Klimova, A., Rudas, T. (2015): On the closure of relational models. arXiv: 1408.2489

Klimova, A., Rudas, T. (2015): Iterative Scaling in Curved Exponential Families. Scandinavian Journal of Statistics 42, 832-847.

Klimova, A., Rudas, T. (2015): On the closure of relational models. arXiv Preprint: 1408.2489

Rudas T. (2015): Directionally collapsible parameterizations of multivariate binary distributions Statistical Methodology, 27, 132-145.

Klimova, A., Rudas T. (2016) Testing the fit of relational models. arXiv:1612.02416

Rudas T, Klimova A (2016) On the closure of relational models. Journal of Multivariate Analysis, 143: pp. 440-452.

Klimova, A., Uhler, C., Rudas, T. (2014): Faithfulness and learning of hypergraphs from discrete distributions. arXiv: 1404.6617

Klimova, A., Uhler, C., Rudas, T. (2015) Faithfulness and learning of hypergraphs from discrete distributions.Computational Statistics and Data Analysis 87, 57-72.

Milibák Eszter (2015) Okságmegközelítések a statisztikában - Elemzés és kritika. Szakdolgozat. ELTE TÁTK, Survey Statisztika MSc.

Rudas, T. (2014): Log-linear and marginal models. In: Wright, J (ed) International Encyclopedia of Social and Behavioral Sciences 2nd ed, Elsevier

Rudas, T. (2015): Effects and interactions. Methodology - European Journal of Research Methods for the Behavioral and Social Sciences. 142-149.

Rudas, T. (2014): Directionally collapsible parameterizations of multivariate binary distributions. arXiv: 1408.2489

Rudas T. (2015): Directionally collapsible parameterizations of multivariate binary distributions Statistical Methodology, 27, 132-145.

Plenáris előadások

Tamás Rudas (2014): Effects and interactions. VI. European Congress of Methodology, July 23-25, 2014 Utrecht

Tamás Rudas (2015): On the measurement of effects and interactions XIV. Congreso de Metodología de las Ciencias Sociales y de la Salud, Palma, July 22-24.

Tamás Rudas (2016): Some current challenges for statistical methodology. Conference of European Statistics Stakeholders 2016, October 20-21, Budapest

Meghívott előadások

Rudas, T., Klimova, A. (2013): Independence on non-product spaces. In Programme and Abstracts of the 7th International Conference on Computational and Financial Econometrics (CFE 2013) and 6th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on Computational and Methodological Statistics (ERCIM 2013) p126. Senate House, University of London, UK, Dec. 14-16, 2013.

Klimova, A., Rudas, T. (2013): Extended relational models. In Programme and Abstracts of 7th International Conference on Computational and Financial Econometrics (CFE 2013) and 6th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on Computational and Methodological Statistics (ERCIM 2013) p125. Senate House, University of London, UK, Dec. 14-16.

Tamás Rudas (2014): On directionally collapsible parameterizations of multivariate binary distributions, ERCIM 2014, December 6-8, Pisa

Anna Klimova, Caroline Uhler, Tamás Rudas (2014): Faithfulness of discrete distributions to graphs and hypergraphs, ERCIM 2014, December 6-8, Pisa

Tamás Rudas (2014): Directionally Collapsible Measures of Association, May 14, Center for Statistics and the Social Sciences, Universiy of Washington

Tamás Rudas (2015): Testing the Fit of Relational Models for Contingency Tables. Eighth International Workshop on Simulation, Bécs, szeptember 21-25.

Tamás Rudas (2015): Simpson's paradox and a linear concept of association. 8th International Conference of the ERCIM WG on Computational and Methodological Statistics London, december 12-14.

Tamás Rudas (2016): Estimation and testing in relational models. IX. International Conference of the ERCIM WG on Computational and Methodological Statistic, December 9-11, Seville

Tamás Rudas (2016): Independence models for non rectangular data. Cost Action Workshop, April 16-17, Paris (2016)

Tamás Rudas (2016): On variants of the iterative scalling algorithm. XXII. International Conference on Computational Statistics, August 23-26, 2016 Oviedo

Tamás Rudas (2016): On generalizations of the log-linear model. October 14, University of Washington

Tamás Rudas (2016): Model based analysis of incomplete data with non-ignorable missing data mechanism. October 10, Columbia University

Tamás Rudas (2016): Model based analysis of incomplete data with non-ignorable missing data mechanism. October 7,New York University

Egyéb előadások

Rudas, T., Klimova, A. (2013): Log-linear models on non-product spaces. In 45e Journées de Statistique de la SFdS (JDS 2013) p40, Toulouse, France, May 27-31, 2013

Rudas, T. (2013): How to avoid Simpson's paradox in treatment selection based on observational data. In Causal inference in health and social sciences (UK-CIM 2013) p17, Manchester, UK, May 14-15, 2013

Rudas, T., Klimova, A. (2013): Log-linear models on non-product spaces. In Márkus, L. & Prokaj, V. (Eds), Abstracts of the 29-th European Meeting of Statisticians (EMS 2013) p258, Budapest, Hungary, July 20-25, 2013

Németh, R., Rudas, T. (2013): On sociological applications of discrete graphical models. In Márkus, L. & Prokaj, V. (Eds), Abstracts of the 29-th European Meeting of Statisticians (EMS 2013) p227. Budapest, Hungary, July 20-25, 2013

Klimova, A., Rudas, T. (2013): Iterative scaling in curved exponential families. In Márkus, L. & Prokaj, V. (Eds), Abstracts of the 29-th European Meeting of Statisticians (EMS 2013) p159. Budapest, Hungary, July 20-25, 2013

Anna Klimova, Caroline Uhler, Tamás Rudas (2014): Parametric faithfulness of discrete data, Algebraic Statistics 2014, May 19-22, Illinois Institute of Technology, Chicago.

Anna Klimova (2014): Learning Hypergraphs from Discrete Data, Prague Stochastics, August 25-29, Prague.

Tamás Rudas (2014): Properties of Measures of Association for Binary Distributions, Prague Stochastics, August 25-29, Prague.

Renáta Németh, Tamás Rudas (2015): Confounding in Causal Analysis in Case of Binary Responses 12th Conference of the European Sociological Association, Prága, augusztus 25-28.

Tamás Rudas (2016): Model based analysis of incomplete data with non-ignorable missing data mechanism. VII Conference of the European Association of Methodology, July 27-29, 2016 Palma