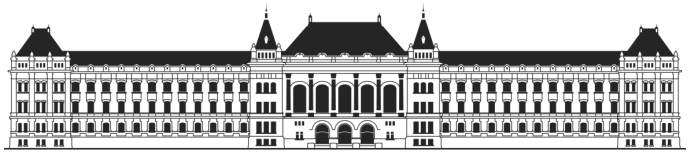


MŰEGYETEM 1782
Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

Bayesi relevancia és hatáserősség mértékek

PhD tézisfüzet

Hullám Gábor

Témavezető:
Dr. Strausz György , PhD (BME)

Budapest, 2016

1. Bevezetés

A különféle jelenségek mögött meghúzódó mechanizmusok megértése mindig is a tudományos kutatás egyik alapvető célja volt. Az egyes kutatási területen a mechanizmusok legtöbbször a változók közötti összetett kapcsolatok és interakciók összességeként írhatók le. A megfigyeléseken alapuló vizsgálatok általános célja ezen kapcsolatok feltárása a függőségi mintázatok vizsgálata által, melyre számos különböző módszert alkalmaznak. Általában egy vagy néhány kiválasztott faktor áll a vizsgálatok célpontjában. Ezek jellemzően státusz leíró változók, melyek az adott területre jellemző besorolással (címkével) látják el a mintákat, ami alapján azok külön osztályokba sorolhatók. Emiatt e változókat osztály- vagy célváltozóknak szokás nevezni. A változók kapcsolatainak elemzése során számos szignifikáns kérdés merülhet fel:

- Hogyan lehet a kapcsolatokat jellemezni?
- Elégséges-e a kapcsolatok meglétének kvalitatív vizsgálata vagy egy kvantitatív elemzés is szükséges?
- Elfogadható-e ha egy analízis csak egyváltozós kapcsolatokat (célváltozó - magyarázó változó párokat) vizsgál?
- Milyen mértékben szükséges a többváltozós kapcsolatok vizsgálata?

A válaszok legtöbbször a vizsgált tárgyterülettől függenek, emiatt nagyszámú, különböző megközelítést alkalmazó módszer jött létre erre a célra, melyeket együttesen jegykiválasztási (feautre subset selection - FSS) módszereknek nevezünk [KJ97].

A jegykiválasztás egy széles körben alkalmazott technika számos területen, mint például a gépi tanulás és a statisztika [BL97; RN10]. Az FSS általános célja egy vagy több célváltozó szempontjából releváns, prediktív erővel bíró jegyek, azaz változók kiválasztása. Az FSS eredménye a releváns változók egy halmaza, melyet többféleképpen lehet definiálni [BW00]. Az egyik lehetőség, hogy egy adott relevancia mérték szerint felállított rangsorból egy előre meghatározott számú elemet (például a 10 legjobbat) választunk ki, és ezek alkotják a releváns változók halmazát. Egy másik opció, hogy egy előzetesen meghatározott küszöbérték alapján választjuk ki az adott relevancia mérték szerinti legrelevánsabb változókat [HS97]. A megfelelő FSS módszer kiválasztása egy adott alkalmazáshoz kihívást jelenthet, mivel jelentős számú mérték és kiválasztási metódus közül lehet választani [GE03].

A változók közötti kapcsolatok leírásának egy további lehetséges megközelítési módja, hogy modellalapú explorációs módszerek segítségével a változók közötti kapcsolatok és interakciók részletes jellemzésére törekszünk. Ezek a módszerek olyan asszociációs mértékeket használnak, melyek amellett, hogy lehetővé teszik a releváns változók azonosítását, további információt szolgáltatnak a változók közötti kapcsolatokról. Az eredmény a tárgyterület egy kifinomult, tudásgazdag, rendszeralapú függőségi modellje, amelynek azonban a magas számítási komplexitás az ára.

A disszertációban bemutatott kutatómunka ez utóbbi, röviden felvázolt rendszeralapú megközelítést követte. A fő motiváció az volt, hogy új megoldási lehetőségeket és módszereket biztosítson olyan a genetikához és orvosbiológiához tartozó alkalmazási területek számára, melyek igénylik a függőségek részletes modellezését. Idővel a génasszociációs vizsgálatok (genetic association studies - GAS) váltak a kutatás központi fókuszává, mivel olyan elemzési módszereket igényelnek, melyek képesek rendszeralapú modellezésre és a statisztikai eredmények konzisztens kezelésére. A bayesi statisztikai keretrendszerben alkalmazott Bayes-háló alapú módszerek eleget tesznek ezeknek a kívánalmaknak, mivel a függőségek részletes jellemzését teszik lehetővé

modellátlagoláson és valószínűségi következtetésen alapuló metódusok révén. A disszertáció fő célkitűzései a következők: (1) a meglévő bayesi relevancia analízis módszerek alkotta keretrendszer kiterjesztése új, a változók hatáserősségét kvantifikáló módszerekkel, és (2) ezen módszerek jelölt génasszociációs vizsgálatokban való alkalmazásához szükséges irányelvek és ajánlott paraméter beállítások ismertetése.

2. Létező módszerek és megközelítések

Az alapvető FSS megközelítés lényege megtalálni azt a minimális számú változót, amely lehetővé teszi a célváltozó predikcióját adott pontossággal [GE03; Inz+00]. Ennek egy jellemző megvalósítása, hogy egy választott tanuló módszer alkalmazásával osztályozót építenek a rendelkezésre álló adat alapján. Az osztályozó két alapvető tényezőjét: a pontosságát és a komplexitását kell ilyen esetben figyelembe venni, melyek között egy megfelelő kompromisszumot kell találni.

Az FSS módszerek két alapvető komponense a változó kiválasztási algoritmus (keresés a lehetséges változók között) és a kiértékelő algoritmus, amely az osztályozó építéséért felelős. E komponensek implementációjától és integráltságának mértékétől függően [LY05] az FSS módszereket az alábbi három csoportba oszthatjuk: (1) szűrők (filters) [HS97; KS96], (2) burkolók (wrappers) [Inz+00; Mao04], és (3) beágyazott módszerek [LY05].

A szűrő módszerek között a legegyszerűbb megoldások egyváltozós megközelítést alkalmaznak, vagyis egy választott relevancia mértéket számítanak ki minden egyes változóra, és az alkalmazott mérce szerint legjobb értékű (legrelevánsabb) változók kerülnek az eredményhalmazba [GE03]. Ezek a módszerek az egyszerűségükből fakadóan hatékonyak lehetnek egyes esetekben, és képesek azonosítani a releváns változók egy részét. Azonban nem képesek többváltozós függőségi mintázatok vizsgálatára, azaz a változók közötti összetett kapcsolatok feltárására. Emiatt értékes információ maradhat figyelmen kívül, ami gátolhatja a vizsgált tárgyterület mechanizmusainak tényleges megismerését.

Az esetek egy jelentős részében átfogó vizsgálat szükséges, ami többváltozós függőségek modellezésére képes módszereket igényel, melyek lehetővé teszik a változók közötti interakciók elemzését.

2.1. Többváltozós modellezési módszerek

A többváltozós modellezést lehetővé tévő módszerek többféleképpen csoportosíthatók [DGL96; Gel+95; Ste09; Esb+02]. Például Devroye et al. a módszerek három csoportját különbözteti meg: feltételes modellezés, sűrűségfüggvény tanulás, és diszkrimináns függvény tanulás [DGL96]. Esbensen et al. pedig a módszereket elsődleges alkalmazási céljuk szerint különbözteti meg, így például külön csoportot alkotnak az osztályozó módszerek és a diszkrimináns módszerek [Esb+02].

Ebben az alfejezetben a többváltozós módszerek két csoportjának: a (1) *feltételes modellezés* alapú módszerek (conditional modeling) és a (2) *rendszeralapú modellezés*t (systems-based modeling) megvalósító módszerek ismertetésére kerül sor. A disszertációban leírt új eredmények és módszerek az utóbbi csoporthoz kapcsolódnak. Az alábbi szakaszban a rendszeralapú módszerek főbb jellemzőinek bemutatása és a feltételes modellezés tulajdonságaival való összehasonlítása következik.

A feltételes modellezés alapú megközelítés főleg burkoló módszerek esetén jellemző, célja a célváltozóra nézve nagy predikciós erővel bíró változóhalmaz azonosítása, tekintet nélkül a változók közötti függőségekre, és a célváltozóhoz kötődő kauzális mechanizmusokban betöltött lehetséges szerepre. Bár a feltételes modellezés lehetővé teszi interakciók vizsgálatát, de nem biztosít részletes jellemzést a változók közötti kapcsolatok típusáról [PCB06]. Például a logisztikus

regresszió, ami egy gyakran alkalmazott feltételes modellezési módszer, lehetővé teszi interakciós termek hozzá vételét a regressziós modellhez, de a célváltozó és a magyarázó változók közötti kapcsolat jellegéről nem nyújt információt.

Tekintsünk példaként egy elképzelt tárgyterületet, amelynek változói az 1. ábrán látható függőségi mintázatot alkotják. Tegyük fel, hogy az X_0 változót szignifikánsnak találjuk az Y célváltozó szempontjából. Feltételes modellezést alkalmazó módszereket használva az a tény, hogy e két változó között tranzitív kapcsolat áll fenn, vagyis X_0 hatását Y-ra mediálják más változók ($X_0 \to X_6 \to Y$ and $X_0 \to X_3 \to X_7 \to Y$), rejtve marad. Egy rendszeralapú megközelítésben, ahol a hangsúly a részletes eredmények magyarázatán és további feldolgozásán van, ez hátránynak számít, mivel az elsődleges cél ekkor a tárgyterület mechanizmusainak feltárása, vagyis az egyes változók szerepeinek azonosítása. Másfelől, egy olyan megközelítésben, ahol a predikción van a hangsúly, az az eredmény, hogy X_0 egy releváns változó önmagában elegendő, ekkor további részletek feltárására nincs szükség.

Ezzel szemben a rendszeralapú modellezést megvalósító módszerek célja az összes változó közötti függőségi kapcsolat feltárása (mind a célváltozók és a magyarázó változók között, mind a célváltozók között) [12]. Ezek a függőségi mintázatok egy irányított aciklikus gráf (DAG) segítségével vizualizálhatóak, ahol a csomópontok változókat jelölnek, az irányított élek pedig a köztük lévő függőségi kapcsolatokat [Lun+00; OLD01]. Ez a DAG struktúra egybeeshet a tárgyterület mechanizmusait leíró kauzális modellel. Egy kiválasztott változóhoz képest (ami jellemzően a célváltozó) vizsgált relatív pozíció alapján a változók közötti kapcsolatok típusokba sorolhatók, úgy mint: közvetlen ok, közvetlen hatás (okozat), interakció és tranzitív kapcsolat.

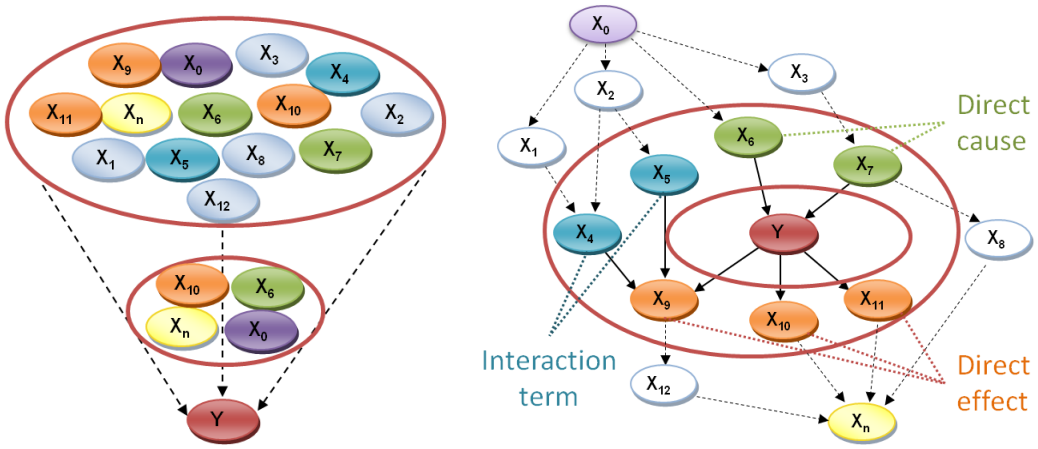
Az 1. ábrán látható függőségi mintázatnál Y a célváltozó, X_6 és X_7 zöld csomópontokkal jelölt közvetlen okok, melyek közvetlenül hatnak Y-ra, azaz nincs közbenső változó. Hasonlóképpen, X_9, X_{10} és X_{11} narancsszínű csomóponttal jelölt közvetlen hatások, melyekre közvetlenül hat Y. Ezzel ellentétben az interakciós tagok (halványkék színű csomóponttal jelölt X_4, X_5) csak feltételesen függenek Y-tól egy közös okozat által, azaz ezeket a függőségi kapcsolatokat egy harmadik változó mediálja. A másik indirekt típus a tranzitív kapcsolat, melynél egy irányított út vezet a célváltozó és a függőségi kapcsolat másik változója között, és ez a változó nem szomszédos a célváltozóval [8]. Két kitüntetett szerepű tranzitív függőségi kapcsolat az ős ok X_0 és a közös hatás X_n. Az előbbi esetben az ős ok hatással van a célváltozóra és más változókra több irányított útvonalon keresztül, míg a közös hatásra számos változó és a célváltozó is hatást gyakorol. A közvetlen okok, közvetlen hatások és interakciók a strukturális relevancia szempontjából jelentősek, mivel elszigetelik a célváltozót a többi változó közvetlen hatásától. Másfelől viszont a tranzitív kapcsolatok gyakorlati szempontból lehetnek hasznosak, például mérhetőség szempontjából könnyebben elérhetőek.

A rendszeralapú modellezést megvalósító módszerek hátránya a magas számítási komplexitás [Coo90; CHM04] és mintakomplexitás [FY96]. Mivel a cél ilyen esetekben a vizsgált tárgyterület részletes függőségi modelljének létrehozása, ezért e módszerek számítási igénye jelentősen meghaladja más megközelítésű módszerekét. Emiatt egyes gyakorlati esetekben, ahol a cél néhány releváns változó meghatározása, melyek megközelítőleg meghatározzák a célváltozót, a rendszeralapú megközelítés szükségtelen.

2.2. A Bayes-statisztikai megközelítés

A rendszeralapú megközelítést megvalósító módszerek legfőbb kihívása az, hogy a teljes modell meghatározása a legtöbb gyakorlati esetben nem megvalósítható annak számításintenzív volta miatt. Ennek elsődleges oka, hogy legtöbbször relatíve magas a változók száma a relatíve alacsony számú mintához képest (elégtelen mintaszám) [FY96].

E problémának két lehetséges megoldása van: (1) egy szakértők által konstruált (fix) modell



1. ábra. (a) feltételes modellezés alapú megközelítés illusztrációja, amely figyelmen kívül hagyja a változók közötti függőségi struktúrát, (b) a rendszeralapú megközelítés illusztrációja, amely egyúttal a lehetséges strukturális kapcsolattípusokat ismerteti. Y a célváltozó, míg $X_0, X_1, \ldots, X_n$ a többi megfigyelt változót jelöli. A célváltozóval fennálló különböző kapcsolattípusokat eltérő színek jelölik: X_0 – ős ok (lila), X_n – közös hatás (sárga), X_6, X_7 – közvetlen ok (zöld), $X_9 - X_{11}$ – közvetlen hatás (narancs), X_4, X_5 – interakciós tag (türkizkék), $X_1 - X_3, X_8, X_{12}$ – további változók (fehér). Azok a változók, melyek közvetlen ok, közvetlen hatás vagy interakciós kapcsolatban álló csomópontoknak felelnek meg, az erősen releváns változók halmazát alkotják (lásd az 1. definíciót). E halmaz elemei (az ábrán a vörös gyűrűvel jelölt rész) statisztikailag izolálják a célváltozót a többi változótól.

illeszkedését vizsgálják a rendelkezésre álló adathoz, vagy (2) lehetséges modelleket (vagy modell részeket) tanulnak az adat alapján. Az előbbi klasszikus statisztikai megközelítést igényel, mivel egy hipotézisre van szükség a modell kialakításához, amelynek az adathoz való illeszkedését statisztikai hipotézistesztelés módszerével lehet vizsgálni. A strukturális egyenletrendszer alapú modellezés a társadalomtudományokban gyakran alkalmazott módszer, amely ezt a megközelítést követi [Pea00].

A másik megközelítés esetében nincs szükség a priori struktúrára, viszont a lehetséges modellek tanulása bayes-statisztikai megközelítést igényel. Ez azt jelenti, hogy egyetlen modell kiértékelése helyett számos lehetséges modellt kell megvizsgálni, azaz minden egyes M modell valószínűségét meg kell határozni a rendelkezésre álló D adat alapján. A Bayes-szabály alapján egy többváltozós függőségi modell *a posteriori* valószínűsége $P(M|D)$ az alábbiak szerint becsülhető [Ber95]:

$$P(M|D) \propto P(D|M) \cdot P(M), \tag{1}$$

ahol $P(D|M)$ a *likelihoodot* jelöli, ami a D adat (létrejöttének) valószínűségét adja meg az M modellt feltételezve, továbbá $P(M)$ jelöli a modell *a priori* valószínűségét. Az adat marginális valószínűsége ($P(D)$) normalizációs konstansként funkcionál, így a kifejezésből elhagyható (további részletek a 5.4 szekcióban találhatók). Ennek a kifejezésnek a lényege az, hogy a modellek felett definiálható egy *a posteriori* eloszlás [Mad+96; HGC95].

Mindemellett a bayesi modellátlagolási technikára alapozva a modellek egyes jellemző tulajdonságai (például a célváltozó környezetét leíró részmodellek) meghatározhatóak [Mad+96; Hoe+99]. Az egyes változók relevanciája is megadható *a posteriori* valószínűség formájában, ami azzal áll kapcsolatban, hogy mennyire jellemzően van jelen egy változó a célváltozó környezetét

leíró lehetséges részmodellekben (például egy magas relevanciájú változó a legtöbb modellben megtalálható).

Több lényeges különbség húzódik a klasszikus statisztika hipotézistesztelési paradigmája és a Bayes-statisztikai paradigma között. A főbb eltéréseket az 1. táblázat foglalja össze. Az első fontos differencia, hogy a bayesi megközelítés a tárgyterület hipotézismentes vizsgálatát teszi lehetővé, szemben a hipotézistesztelésen alapuló klasszikus statisztikai megközelítéssel, amely előírja egy hipotézis meglétét, melyet statisztikai teszttel vizsgálni lehet. A vizsgált hipotézist többnyire a függetlenséget leíró nullhipotézissel állítják szembe. Bár a bayesi módszerek nem hipotézis által vezéreltek, de lehetőséget adnak az *a priori* szakértői tudás felhasználására priorok formájában, amely adott esetben részletesebb lehet egy hipotézisnél.

1. táblázat. A klasszikus statisztikai és a bayesi megközelítés összehasonlítása modellezési tulajdonságok alapján. *A priori tudás* – a felhasznált *a priori* tudás jellege, *Kiértékelés módszere* – eredmények (modellek) kezelése, *Kiértékelés alapja* – a modellek kiértékeléséhez használt mérték, *Eredmény* – a modellezés kimenete, *Variancia leírása* – a variancia mértékét megadó mennyiség, *Döntés alapja* – a végső modell meghatározásának alapja, *Probléma* – a megközelítéshez kapcsolódó jellemző probléma.

Tulajdonság	**Klasszikus megközelítés**	**Bayesi megközelítés**
A priori tudás	Hipotézis (egyetlen modell)	Több lehetséges modell (*a priori* valószínűségek)
Kiértékelés módszere	Modellkiválasztás	Modellátlagolás
Kiértékelés alapja	Statisztikai teszt	Bayes faktor
Eredmény	p-érték (nullhipotézis elvetése v. elfogadása)	A posteriori valószínűségek
Variancia leírása	Konfidencia intervallum	Megbízhatósági intervallum
Döntés alapja	Szignifikancia szint	Optimális döntés a várható hasznosság alapján
Probléma	Többszörös hipotézis tesztelés	Számítási komplexitás

Egy további lényeges kérdés a modellek validációjához kapcsolódik. A klasszikus statisztikai hipotézistesztelő keretrendszerben egy (pl.: két változó között függőséget leíró) modell akkor kerülhet elfogadásra, ha a kapcsolódó nullhipotézis (pl.: a két változó független) elvetésre kerül. Leegyszerűsítve, ebben az esetben a számított statisztikához tartozó p-értéknek egy meghatározott küszöbértéknél alacsonyabbnak kell lennie. Ezt a küszöbértéket nevezik szignifikancia szintnek, jellemzően $\alpha = 0.05$. Ellenkező esetben az alternatív hipotézist kell elvetni függetlenül attól, hogy a p-értéke közel volt a küszöbértékhez (pl.: p-érték$= 0.052$) vagy sem (pl.: p-érték$= 0.92$).

Ezzel szemben a bayes-statisztikai keretrendszer a modellekkel kapcsolatos állításokat (az adat által nyújtott megerősítést) *a posteriori* valószínűségként kvantifikálja, amely alkalmas a relevancia közvetlen mérésére. Ez lehetővé teszi a modellek valószínűségeinek összehasonlítását és a modellátlagolást. Információ elhagyása nélkül, minden eredmény kezelése konzisztensen megvalósítható.

2.3. A Bayes-háló modellosztály

A valószínűségi gráfmodellek (probabilistic graphical models - PGM) ideális eszközök a rendszeralapú többváltozós modellezés megvalósításához, mivel lehetővé teszik valószínűségi változók

feltételes függetlenségeinek és függőségeinek a reprezentálását egy irányított gráf formájában [12], [FK03; CH92; Mad+96]. A Bayes-háló modellosztály az egyik leggyakrabban alkalmazott PGM, mely széleskörűen alkalmaznak számos területen, úgymint a gépi tanulás, számítógépes biológia (computational biology), és a képfeldolgozás [Bar12; Mit07]. A Bayes-hálók három fő tulajdonsága, melyek lehetővé teszik, hogy egy sokoldalú modellezési eszközként funkcionáljanak a következőek: (1) képes hatékonyan reprezentálni az együttes valószínűség-eloszlást, (2) lehetővé teszi a valószínűségi véletlen változók közötti feltételes függetlenségek hálózatának reprezentálását (conditional independence map), és ha kauzális értelmezés megengedhető, akkor (3) képes ábrázolni az ok-okozat kapcsolatokat [Pea00]. A Bayes-háló alapú (tanuló) módszerek lehetővé teszik a többváltozós függőségi kapcsolatok reprezentálását és detektálását, továbbá egy gazdag eszközkészletet nyújtanak az asszociációk részletes jellemzésére [8], [12], [MSL12].

A disszertációban bemutatott módszerek és eredmények a rendszeralapú többváltozós modellezésen alapuló megközelítéshez kötődnek. Az implementáció alapját egy bayes-statisztikai keretrendszerben alkalmazott Bayes-háló modellosztály képezte.

3. Alkalmazási területek

A disszertáció az alábbiakban bemutatott alkalmazási területekre helyezi a hangsúlyt, melyek rendszeralapú többváltozós modellezést igényelnek, és melyekhez kötődően számos megoldandó probléma merült fel. Mindemellett, a lehetséges modellek nagy száma miatt a bayes-statisztikai keretrendszer hatékony alkalmazása is egy lényeges kihívást jelent, mivel kezelnie kell a többszörös tesztelés okozta problémát és lehetővé kell tennie a bayesi modellátlagolást.

3.1. Génasszociációs vizsgálatok

A közelmúltban az orvosbiológia és genetikai méréstechnológia terén lezajlott gyors ütemű fejlődés lehetővé tette a multifaktoriális (azaz több géncsoport és környezeti tényező együttese által indukált) betegségek (például: rheumatoid arthritis, depresszió, asztma) genetikai hátterének kutatását. Ez az alkalmazási terület megköveteli a komplex függőségi kapcsolatok modellezését, amely elengedhetetlen az ilyen összetett hátterű betegségek mechanizmusainak megértéséhez [Ste09]. A génasszociációs vizsgálatok (GAS) olyan egynukleotidos polimorfizmusok (single nucleotide polymorphism- SNP [Bal07]) azonosítását tűzte ki, melyek befolyásolják a betegségek kialakulását vagy lefolyását (súlyosságát). Egy tipikus GAS öt szakaszból áll: (1) vizsgálat megtervezése, (2) minták begyűjtése, (3) genetikai mérések, (4) statisztikai elemzés, és (5) az eredmények feldolgozása, értelmezése.

A kezdeti időszakban (2000-2005) egy egyszerű megközelítés volt a jellemző, lényegében (változó) páronkénti asszociációs teszteket végeztek, azaz a jellemzően bináris, betegségstátuszt leíró célváltozó és SNP-knek megfelelő változók között vizsgáltak statisztikai függőséget. Ha egy adott SNP esetében a genotípusok (a SNP lehetséges variánsai, azaz a változó lehetséges értékei) eloszlása különböző volt esetek (beteg páciensek) és a kontrollok (egészséges páciensek) csoportjában, akkor az azt jelentette, hogy az SNP valamilyen szerepet játszik a vizsgált betegség mechanizmusaiban.

A nagy áteresztőképességű genotipizáló (genotípus meghatározó) technológiák megjelenése lehetővé tette a teljes genom asszociációs vizsgálatokat (genome-wide association studies - GWAS), melyek 10^4-10^5 SNP lemérésére képesek (egy vizsgálat keretében). Egyes területeken a GWAS teljesen felváltotta a korábbi kisebb léptékű vizsgálatokat, ahol egy vizsgálat néhányszor tíz-száz SNP mérését jelentette. Ez utóbbira újabban kandidáns (jelölt) génasszociációs vizsgálatként (candidate gene association study - CGAS) hivatkoznak.

A GWAS-ok többsége azonban csak mérsékelten volt sikeres. A fő célkitűzésük az elemzés tekintetében az volt, hogy egy egységes szemléletben kezeljék a statisztikai teszteket, azaz páronkénti vizsgálatokat az összes SNP-re elvégezzék ugyanolyan módszerekkel, azonos beállításokkal és korrekciókkal, ahelyett, hogy külön-külön elemeznék a SNP-ket. Sajnálatos módon a várakozásoktól elmaradó eredmények egyik oka pont az egységes statisztikai tesztek kapcsán alkalmazott többszörös hipotézis tesztelés miatti szigorú korrekció volt. A többszörös hipotézis tesztelés miatti korrekció a hipotézistesztelésen alapuló statisztikai keretrendszer miatt szükséges a véletlen hamis pozitív eredmények kiküszöbölésére, amelynek mértéke nem elhanyagolható ha ugyanazon az adathalmazon végzünk el akár több ezer statisztikai tesztet. A problémát az jelenti, hogy a korrigált szignifikanciaszint elég alacsony: 10^7-10^8, ami jelentősen korlátozza a tesztek által detektálható hatáserősséget (a SNP variánsok célváltozóra vonatkoztatott hatáserősségét), és növeli a szükséges mintaszámot [GSM08].

A mérsékelt siker egy további feltételezett oka a túlzottan leegyszerűsített megközelítés volt a célváltozó és a környezeti változók tekintetében, mivel jellemzően egyetlen betegségstátuszt leíró célváltozót alkalmaztak, a kapcsolódó környezeti és klinikai változókat viszont figyelmen kívül hagyták. Mivel a multifaktoriális betegségek esetében a környezeti változók szerepe jelentős, ezért több tanulmány is javasolta ilyen jellegű faktorok vizsgálatba való bevonását [Man+09; EN14]. Ennek köszönhetően a CGAS-ok ismét előtérbe kerültek, mint (GWAS-ok eredményeit) megerősítő vizsgálatok, ezúttal azonban többrétű környezetleíró és fenotípus (külsőleg megfigyelhető jegyek, mint például: nem) változókat alkalmazva [PCB13]. Mindazonáltal a korábban alkalmazott egyváltozós módszerek nem alkalmasak több genetikai és környezeti változó együttes vizsgálatára, ehhez mindenképp többváltozós módszerekre van szükség.

Ezek az akadályok inspirálták a kutatást új statisztikai módszerek kifejlesztésére. A módszerekkel szemben támasztott főbb követelmények az alábbiakban foglalhatók össze:

Komplex függőségi kapcsolatok elemzésének képessége
: A komplex fenotípus megközelítés lényege a genetikai, környezeti és klinikai változók együttes elemzése. Ebből kifolyólag olyan módszerekre van szükség, melyek többváltozós statisztikai elemzésre képesek, többek között alkalmasak interakciók detektálására [Ste09; Com+00; Sto+04].

Többszörös hipotézistesztelés optimális megoldása
: A többszörös hipotézistesztelés alapvető fontosságú a hamis pozitív eredmények elkerülése érdekében, azonban a jelenlegi formájában (a hipotézistesztelő keretrendszerben) túlzottan szigorú [GSM08]. Mérsékelten szignifikáns eredményeket is gyakran elutasít, holott lehet, hogy érdemes lenne azokat további vizsgálni. Továbbá, több fenotípus együttes vizsgálata esetén több statisztikai tesztet kell elvégezni, ami további korrekciót jelent, így összességében még inkább korlátozó tényezővé válik a korrekció. Az új módszereknek kvantifikálniuk kellene a közepes és gyenge relevanciájú eredményeket, a potenciálisan hasznos információ elhagyása nélkül.

Eredmények kiértékelésének támogatása
: Az alapvető elemzés mellett az új módszereknek preferáltan további eszközöket kellene nyújtaniuk, például kiegészítő mértékek formájában, az eredmények vizualizációjának és értelmezésének támogatására.

A Bayes-hálókon alapuló rendszeralapú többváltozós modellezés eleget tesz ezeknek a kívánalmaknak, mivel lehetővé teszi a függőségi kapcsolatok elemzését, konzisztens korrekciót biztosít többszörös hipotézis tesztelés esetén, valamint egy gazdag eszközkészletet nyújt az eredmények kiértékeléséhez és magyarázatához.

3.2. Betegségek modellezése

A betegségek modellezésének a célja klinikai, környezeti és genetikai változók együtteséből álló részletes paraméterezéssel ellátott függőségi modellek létrehozása, melyek elősegítik a terápiaválasztást, lehetővé teszik a betegségek kezeléséhez kapcsolódó kockázatbecslést, és végül lehetővé teszik az orvosi döntéstámogatást [LT06; SP78; OLD01; Bel+96; Mik+95]. Ez utóbbi alkalmazáshoz azonban már szükség van a modellek további összetevőkkel való bővítésére, úgymint hasznosságok definiálása, költséghatékonyság érvényesítése, és további döntéselméleti elemekre. Ez az alfejezet a modellkonstrukcióra helyezi a hangsúlyt.

A modellépítést különféle módokon kezelhetjük, melyek mindegyike rendelkezik előnyökkel és hátrányokkal.

Modellépítés szakértői tudás alapján
Jól ismert tárgyterületek esetén, ahol jelentős mértékű *a priori* tudás áll rendelkezésre, létrehozhatók modellek pusztán szakértői tudást alapul véve [Wei+78]. Ezek a modellek később validálhatók, ha rendelkezésre állnak kapcsolódó adathalmazok. Értelemszerűen ez a metodológia nem alkalmazható egy teljesen ismeretlen területen.

Modelltanulás adat alapján
Ha nincs elérhető *a priori* tudás, akkor a modellt az adatok alapján kell megtanulni. Bayeshálók esetén erre a célra számos struktúratanuló algoritmus alkalmazható [FK03; CH92]. Ez követően pedig a tanult struktúrát alapul véve paramétertanulást kell végrehajtani. Egy lehetséges probléma ott merülhet fel, hogy a tanulóalgoritmusok paraméterezésétől függően akár jelentősen eltérő modellek is adódhatnak. Emiatt a paraméterezés helyességét vizsgálattal kell alátámasztani az eredmények jóságának biztosítása érdekében.

Modelltanulás adat és *a priori* tudás alapján
A harmadik lehetőség, hogy az adat alapján történő modelltanulás során felhasználjuk a rendelkezésre álló *a priori* tudást. A gyakorlati esetek többségében az elérhető *a priori* tudás nem elégséges egy teljes modell létrehozásához, de segítheti a modelltanulást. Ezen *a priori* tudás felhasználását a bayesi módszerek lehetővé teszik *informatív* és *nem informatív* priorok formájában. Az *informatív priorok* tárgyterület specifikus információkat tartalmaznak, mint például egy célváltozó szempontjából várhatóan releváns változók és kapcsolataik. Az ilyen jellegű információk felhasználhatók a függőségi struktúra tanulásánál például úgy, hogy egyes változók között feltételezünk kapcsolatot, míg esetleg más változókat figyelmen kívül hagyunk [AC08]. A *nem informatív priorok* általános jellegű feltevéseket írnak le, úgymint egy változóval közvetlen függőségben álló változók maximális száma, egy releváns változó megtalálásának az esélye, vagy a teljes modell komplexitására vonatkozó becslés [Kon+98].

A gyakorlatban alkalmazott módszerek nagy része a harmadik megközelítést követi, azaz valamennyi *a priori* tudásra támaszkodva modelltanulást valósítanak meg. Az esetek többségében nem informatív priorokat használnak, amelyek kapcsolata a vizsgált tárgyterülettel nem egzakt. A tárgyterülettel kapcsolatos tudás lefordítása modellezésnél használható tudássá (egy adott formába) legtöbbször jelentős kihívást jelent. Például ha egy Bayes-háló konstruálásánál alkalmazunk egy küszöbértéket az egy csomópontba bejövő élek számára, akkor nem triviális, hogy ez hogyan köthető a tárgyterület egy jellemzőjéhez. A struktúratanulásnál számos olyan paraméter ragadható meg, melyeket nem lehet egyértelműen az adott tárgyterület tulajdonságaihoz kötni.

A modelltanulás során egy további kihívást jelent az adat elégségessége, ami legfőképp a változók számától és a mintaszámtól függ. Emellett a változók kardinalitásától (lehetséges értékei

számosságától) és az elemzési módszertől is függ az elégségesség, mivel ha több változót együttesen vizsgálunk, akkor a több mintára van szükség, hogy elkerüljük a statisztikailag inadekvát mintaszámot. Emiatt célszerűnek látszik különböző modellezési előfeltevéseket megfogalmazni adekvát mintamennyiségnél, illetve kis mintás esetben.

3.3. Intelligens adatelemzés

Bár az intelligens adatelemzés szerves részét képzi az előbbiekben ismertetett GAS és betegségmodellezés alkalmazási területeknek [BZ08], ez az alfejezet egy általánosabb aspektusra helyezi a hangsúlyt.

Az intelligens adatelemzés hipotézismentes, adatvezérelt, statisztikai adatelemzésként definiálható (gépi tanulási módszerek felhasználásával), melynek célja a változók közötti (függőségi) kapcsolatok feltárása, és a háttérben meghúzódó mechanizmusok részletes leírása [Kon01; LKZ00; CY08]. A hipotézismentesség egyes esetekben azonban lehet kevésbé előnyös is, például egyes ritka többváltozós függőségi kapcsolatok vizsgálata esetén, illetve az alábbiakban:

Kontextuálisan releváns változók
: A GWAS vizsgálatokkal szembeni egyik legfőbb kritika az volt, hogy nem vették figyelembe a vizsgált betegséghez tartozó jellemző kontextust, azaz a releváns környezeti és fenotípus leírókat. Az új kontextus gazdag génasszociációs vizsgálatokban az ilyen jellegű leíró változók szerepe egy olyan kontextus azonosítása, amelyben a genetikai és klinikai változók releváns hatást mutatnak a vizsgált betegség kapcsán. Mindemellett olyan eset is előfordulhat, hogy egy változó csak egy egyedi kontextusban, azaz az érintett leírók egy adott értékkombinációja mellett mutat releváns hatást, míg minden más esetben irreleváns [Bou+96]. Ha a kontextusnak megfelelő alpopuláció a teljes mintánál jóval kisebb, akkor fennáll a lehetősége annak, hogy a változó releváns volta nem mutatható ki általános megközelítéssel, különösen akkor, ha nem áll rendelkezésre *a priori* információ a lehetséges kontextust alkotó változókról.

Változók együttes hatása
: A gyakorlati esetek többségében a környezeti és fenotípus leíró változók erős függést mutatnak a célváltozóval. Ezzel szemben a genetikai és klinikai változók jellemzően jóval gyengébb függőségekkel rendelkeznek a célváltozó viszonylatában, ami releváns változóként való azonosításukat megnehezítheti. Továbbá, ha egy változó csak egy másik változóval együttesen releváns, és önmagában elhanyagolható a hatása, akkor a változó relevánsként való azonosítása még nehezebb.

Az egyik lehetséges megoldás az intelligens explorációs elemzőeszközök alkalmazása, amelyek lehetővé teszik priorok alkalmazását. Azonban a háttértudás átalakítása priorrá nem egy egyértelmű folyamat, egyes esetekben akadályokba ütközhet. Egy további lehetséges megoldás olyan relevancia mértékek alkalmazása, melyek a háttértudást felhasználva hatékonyan feldolgozhatók.

4. Célkitűzések

Kutatásom fő célkitűzései az alábbi pontokban foglalhatóak össze:

1. cél: Relevancia mértékek kialakítása intelligens adatelemzéshez.
: A relevancia különféle aspektusokból értelmezhető, úgymint parametrikus, strukturális és

kauzális aspektusból. Egy bayes-statisztikai keretrendszerben a Bayes-háló tulajdonságaira alapozva létrehozhatók új relevancia mennyiségek, melyek egyesítik a parametrikus és strukturális aspektusokat, és átfogó képet nyújtanak a relevancia relációkról.

2. cél: Nem informatív priorok hatásának vizsgálata.
Az *a priori* tudás átalakítása Bayes-háló alapú többváltozós modellezés esetén nem informatív priorokká jelentős kihívásokat rejt. Különösképp paraméter priorok esetén, melyek megfelelő megválasztása alapvető fontosságú, mivel komplexitás regularizáló szerepet játszanak. Számos alkalmazási területen a modelltanulást elősegítené, ha a paraméter prior a tárgyterülethez kapcsolódó (parametrikus) tulajdonságokhoz lenne köthető.

3. cél: Bayes-háló alapú többváltozós modellezés alkalmazása GAS esetén.
A Bayes-háló alapú többváltozós modellezést megvalósító *Bayesi relevancia analízis* mesterséges adatokon történő kiértékelése bizonyította a módszer alkalmazhatóságát génasszociációs vizsgálatok (GAS) elemzésére. A bayesi relevancia analízis keretrendszer az új kiterjesztésekkel és a paraméterezéséhez kapcsolódó új eredmények figyelembe vételével együttesen egy értékes GAS elemző eszközként szolgálhat. A modelltanulás folyamán figyelembe vehető számos paraméter és gyakorlati megfontolás miatt egy alkalmazási útmutató nagy mértékben elősegítené a módszer GAS-beli alkalmazását.

A disszertáció ezeket a célokat helyezi a középpontba, és a továbbiakban bemutatott eredményekből és új módszerekből áll.

5. Kutatási módszer és új tudományos eredmények

A disszertációban bemutatott kutatás alapjai a Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszékének Számítógépes Biomedicina és Bioinformatika (COMBINE) munkacsoportjának kutatási tevékenységeihez kapcsolódnak. A COMBINE munkacsoport kutatásának középpontjában Bayes-statisztikai keretrendszerben alkalmazott bayesi modellalapú módszerek fejlesztése és felhasználása áll olyan szerteágazó területeken, mint a genetika és az orvosbiológia. E kutatómunkának egyik jelentős mérföldköve a bayesi többszintű relevancia analízis (Bayesian multilevel analysis of relevance - BMLA), amely egy bayesi metodológiai keretrendszer a relevancia elemzésére. (Az egyszerűség kedvéért a továbbiakban *bayesi relevancia analízis*ként (Bayesian relevance analysis) hivatkozik rá a dolgozat.) A módszer elsőként egy petefészekrákhoz kapcsolódó adatelemzéshez készült, mely első alkalmazási területeként szolgált [Ant07]. Ezt követően a COMBINE csoport tagjai bővítették és adaptálták nagy számítási teljesítményű gridekre [12], [8]. A későbbiekben számos kollaboráció jött létre több Semmelweis egyetemi kutatócsoporttal abból a célból, hogy bayesi elemző módszereket tegyünk elérhetővé a jelölt génasszociációs vizsgálatok kiértékelésére [9]. Ennek következtében a jelölt génasszociációs vizsgálatok váltak a bayesi relevancia analízis fő alkalmazási területévé, továbbá e disszertációban leírt módszerek és eredmények létrejöttét a jelölt génasszociációs vizsgálatok által nyújtott kihívások motiválták.

A doktori kutatásom a bayesi relevancia analízis egy valós jelölt génasszociációs vizsgálat (CGAS) elemzésére történő adaptációjával kezdődött. Először a bayesi relevancia analízis alkalmazhatóságát vizsgáltam mesterséges adathalmazon egy összehasonlító elemzés keretében. A módszer jóságát több klasszikus statisztikai és bayesi módszerrel vetettem össze [25]. A második lépés a bayesi relevancia analízis alkalmazása volt egy asztma CGAS-hoz kapcsolódó valós adathalmazon, melynek során a CGAS elemzési metodológia első változatának kidolgozásra került sor [3]. Ezt követően számos CGAS elemzésben vettem részt, melyek a rheumatoid arthritis,

allergia, leukémia, hypodontia és depresszió betegségekhez kapcsolódtak. Ezek során számos új kihívás merült fel, melyek ösztönözték a bayesi relevancia analízis kiterjesztését a relevancia parametrikus aspektusát feltáró új mértékekkel, illetve a nem informatív priorok hatásának vizsgálatát. Végül az elemzési metodológia kiegészült az új eredmények felhasználásával.

A következő fejezetrész összefoglalja a disszertációban bemutatott új tudományos eredményeket. Elsőként a 5.1 alfejezet egy átfogó képet ad az eredményekről a korábban ismertetett célkitűzések tükrében. Ezt követően a 5.2 - 5.4 alfejezetekben kerül sor az eredmények részletes bemutatására, melyet minden esetben a tématerülethez kapcsolódó rövid áttekintés előz meg.

5.1. Eredmények áttekintése

1. Tézis - áttekintés

A relevancia mértékek jellemzően a relevancia egy adott aspektusát vizsgálják. Az asszociációs mértékek egy adott változó hatáserősségét mérik a célváltozó szempontjából, ami lehetővé teszi a prediktív erő meghatározását. Például egy genetikai változó hatáserősségének a meghatározása egy betegség kapcsán egy eset-kontroll vizsgálat keretében lehetővé teszi annak meghatározását, hogy mely genetikai variáns növeli a betegség előfordulását. Ilyen típusú mértékek azonban nem nyújtanak információt a változók közötti függőségi viszonyokról, tehát ezek alapján nem lehet meghatározni, hogy az adott genetikai változó közvetlen hatással rendelkezik a betegségre vagy csak közvetítő változókon keresztül hat. Ezzel ellentétben a strukturális relevancia mértékek a változók háttérben meghúzódó függőségi mintázatát vizsgálják, és ennek reprezentálására Bayes-hálókat alkalmaznak. Például, ha egy klinikai tanulmány célja azonosítani egy kezelési célpontot a vizsgált változók között (azaz a kezelés a választott klinikai változó befolyásolására irányul), akkor annak az ismerete, hogy az egyes változók milyen szerepet töltenek be a függőségi viszonylatrendszerben, alapvető fontosságú. Egy a célváltozóra közvetlen hatást gyakorló változó befolyásolása hathatósabb eredményeket hozna, mint egy csupán tranzitívan kapcsolódó (más változók által mediált) változó befolyásolása. Ugyanakkor ezek a mértékek nem veszik figyelembe a parametrikus aspektust, vagyis nem határozzák meg a változó hatásának nagyságát a célváltozóra.

Ebben a tézisben javaslom a relevancia hibrid szemléletű megközelítését, és erre a célra egy új bayesi relevancia mértéket, mely lehetővé teszi a strukturális és parametrikus relevancia aspektusok együttes vizsgálatát. A kapcsolódó publikációk a következők: [1], [2], [4], [6], [7], [11].

- A javasolt hibrid megközelítéshez szükség van egy *hatáserősség mérő komponensre*, mely a parametrikus relevanciát kvantifikálja, illetve egy *strukturális relevanciát vizsgáló komponensre*, amely a függőségi kapcsolatok strukturális tulajdonságainak a kvalitatív meghatározását teszi lehetővé. Ez utóbbi komponens ahhoz szükséges, hogy a hatáserősség számítása során a strukturális relevancia is figyelembe legyen véve. Például egy naív hibrid mértéket meg lehetne úgy valósítani, hogy a hatáserősséget kiszámítja minden lehetséges függőségi struktúra mellett. Azonban ebben az esetben ez a mérték nem fejezné ki megfelelően a strukturális relevanciát, mivel olyan függőségi struktúrákat is figyelembe venne, melyek a vizsgált célváltozó szempontjából strukturálisan irrelevánsak.
- Mindezek miatt egy hibrid bayesi hatáserősség mértéket javasolok: a *strukturális feltételű bayesi hatáserősséget* (structure conditional Bayesian odds ratio), mely csak olyan függőségi struktúrákat vesz számításba, ahol a vizsgált változó a célváltozóhoz képest strukturálisan releváns [1], [2], [11].

- Megmutattam, hogy a *strukturális feltételű bayesi hatáserősség* kiszámítható az adathalmazból tanult Markov-takaró gráfok *a posteriori* valószínűségeinek a felhasználásával. A lényegi kapcsolat az, hogy a Markov-takaró gráfok csak erősen releváns változókat tartalmaznak. [1] Tehát ha csak azokat a függőségi struktúrákat vesszük figyelembe a hatáserősség számításánál, amelyeknél a vizsgált változó a célváltozó Markov-takarójában található, akkor a strukturális relevancia fennáll. Mindezt alapul véve definiálható egy gyakorlatban jól alkalmazható hibrid bayesi hatáserősség mérték: a *Markov-takaró alapú bayesi esélyhányados* (MBG-based Bayesian odds ratio) [2], [7], [11].

- Kiterjesztettem a *Markov-takaró alapú bayesi esélyhányados*t változók egy halmazára, ami lehetővé teszi több változó együttes hatáserősségének meghatározását [2].

- A Markov-takaró alapú bayesi esélyhányados alkalmazására több génasszociációs vizsgálatban is sor került, úgymint a folát útvonalhoz tartozó gének leukémiában betöltött szerepének vizsgálatában [4], a szerotonin transzporter (5-HTTLPR) variánsok depresszióra kifejtett hatását elemző vizsgálatban [6], a rheumatoid arthritis genetikai és klinikai faktorainak vizsgálatánál [7], és az impulzivitás és a HTR1A kapcsolatát vizsgáló esettanulmányban [1].

2. Tézis - áttekintés

A tudományos eredmények értelmezése, és az erre támaszkodó döntéshozatal a tudományos munka alapvető elemei. Döntést hozni arról, hogy egy eredmény elfogadható-e függ az alkalmazott benchmark(ok)tól, a tárgyterülethez kapcsolódó általánosan elfogadott ismeretektől, és további preferenciáktól. A vizsgálatot követő lépéseket, úgymint az eredmények publikálását, további kísérletek végrehajtását, vagy új vizsgálat megtervezését az eredmények alapján, mind befolyásolják a preferenciák. A probléma nem az, hogy preferenciák léteznek, hanem hogy nincsenek formálisan meghatározva. Egy lehetséges megoldás egy meghatározott preferenciákat kezelő döntéselméleti keretrendszer alkalmazása. A bayesi keretrendszer egyik jelentős előnye, hogy lehetővé teszi egy döntéselméleti komponens integrálását, illetve olyan speciális mennyiségek kialakítását, amelyek a preferenciáknak való megfelelést vizsgálni képesek. Például egy preferencia a releváns hatáserősségek tekintetében lehet egy küszöbérték, amelyet egy tárgyterületen általánosan elfogadott releváns változó hatáserőssége alapján állapítunk meg. A bayesi döntéselméletet alkalmazva a kapcsolódó várható veszteség alapján meghatározható az optimális akció, azaz például hogy jelentsük-e az adott változót releváns elemként vagy sem.

A kontextuális relevancia vizsgálatának kérdése egy másik lehetséges preferencia, amit a döntéselméleti keretrendszer képes megvalósítani. A kontextuális relevancia azt jelenti, hogy a vizsgált változó csak egy adott kontextusban, azaz más (kontextust formáló) változók meghatározott értékkonfigurációja esetén releváns a célváltozó szempontjából. Például egy stressz szint leíró környezeti változó ('alacsony', 'közepes' és 'magas' értékekkel) egy lehetséges kontextusformáló változó, és a 'magas' értéke egy lehetséges kontextus, ahol a stresszhez kötődő változók hatáserőssége releváns. A kontextuálisan releváns függőségi kapcsolatok feltárása hagyományos módszerekkel nem mindig lehetséges, főleg ha a releváns kontextus ritka a vizsgált populációban. Emiatt a kontextuálisan releváns kapcsolatok feltárása érdekében kontextusfüggő döntéselméleti mértékre és kontextusformáló változó jelöltekre van szükség. Ez a megközelítés releváns eredményeket szolgáltathat adatelemzéshez, különösen komplex függőségi hálózattal rendelkező területek esetén.

[1] Az erős relevancia a strukturális relevancia egy formája.

E tézisben javaslom a döntéselméleti bayesi relevanciamértékek alkalmazását az előzetesen meghatározott preferenciák kvantitatív kiértékelésére. A kapcsolódó publikációk a következők: [5], [10], [12].

- Létrehoztam a *hatáserősségen alapuló egzisztenciális relevancia* (effect size conditional existential relevance - ECER) mértéket, ami lehetővé teszi a kiértékelés specifikus a priori tudás közvetlen alkalmazását. ECER számításához szükséges a (szakértői megítélés szerint) elhanyagolható mértékű hatáserősségek intervallumának megadása, amely implicit módon megadja a releváns hatáserősségek tartományát [10].
- Megvalósítottam az ECER mérték kontextuális kiterjesztését (C-ECER), amely lehetővé teszi a kontextuálisan releváns függőségi kapcsolatok feltárását [12].

3. Tézis - áttekintés

A bayesi megközelítés alapvető paradigmája szerint a függőségi struktúrák Bayes-háló struktúrákként való tanulása esetén a $P(G|D)$ *a posteriori* valószínűsége a G struktúrának D adat esetén a likelihood score $P(D|G)$ és az a priori valószínűség $P(G)$ alapján számítható [Ber95]. Ez utóbbi komponensek teszik lehetővé különböző típusú apriori tudás beépítését. A felhasznált háttértudás lehet informatív (például egyes struktúrákat preferál, így azokhoz magasabb a priori valószínűséget rendel) vagy nem informatív (például minden lehetséges struktúrához ugyanakkora a priori valószínűséget rendel). E struktúra priorok mellett a likelihood score szabad paraméterei is állíthatóak a priori, ezeket nevezzük paraméter prioroknak [HGC95]. Az informatív priorokat tárgyterület specifikus tudás alapján határozzák meg, amely viszont rendszerint nem elérhető. Azonban az a priori valószínűségi eloszlást mindenképpen definiálni kell, ekkor jutnak szerephez a nem informatív priorok. Tehát ha nem áll rendelkezésre a priori tudás, akkor egy semleges a priori eloszlást kell definiálni a nem informatív priorok révén. Azonban paraméter priorok esetén, különösen a gyakran alkalmazott bayesi Dirichlet prior esetén, még a nem informatív esetben is számos lehetőség adódik. A Dirichlet prior szabad paraméterét *virtuális mintaszám*nak nevezzük, amelyet egyfajta komplexitás regularizációnak lehet tekinteni. Habár a virtuális mintaszám megadása önmagában egy nem informatív priornak tekinthető, mégis jelentős mértékben befolyásolja a bayesi struktúratanulást, és ezáltal a bayesi relevancia analízist is.

Ebben a tézisben meghatároztam a nem informatív paraméter priorok hatását a bayesi relevancia analízisre. A kapcsolódó publikációk a következők: [2], [3], [7], [8], [9].

- Levezettem egy kifejezést, amely összeköti egy adott W változóhoz köthető paraméterek a priori eloszlásának virtuális mintaszám paraméterét a W változó várható hatáserősségével [2].
- Gyakorlati megközelítésben ez a kapcsolat felhasználható arra, hogy a paraméter priort a virtuális mintaszám meghatározása által úgy definiáljuk, hogy az összhangban legyen a hatáserősségek a priori (várható) eloszlásával [2], [7].

Az eredmények részletes bemutatása

5.2. Bayesi hibrid relevancia mértékek

Egy prediktor (változó) relevanciája a célváltozó szempontjából a gépi tanulás egy alapvető koncepciója. Azonban értelmezése többféleképpen történhet, továbbá az olyan további fogalmakkal való kapcsolata, mint asszociáció, prediktív erő, és hatáserősség gyakran nem egyértelmű. A relevancia egy feltételes valószínűségeloszlás alapú általános definíciója a következőképpen fogalmazható meg [KJ97]:

1. Definíció (Erős és gyenge relevancia)**.** *Egy X_i jegy (változó) erősen releváns Y szempontjából, ha létezik $X_i = x_i, Y = y$ és $\mathbf{s_i} = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, és $p(x_i, \mathbf{s_i}) > 0$ úgy, hogy igaz $p(y|x_i, \mathbf{s_i}) \neq p(y|\mathbf{s_i})$. Egy X_i jegy (változó) gyengén releváns, ha nem erősen releváns, és létezik a jegyek (változók) egy részhalmaza $S_i' \subset S_i$, továbbá x_i, y és $\mathbf{s_i'}$, és $p(x_i, \mathbf{s_i'}) > 0$ úgy, hogy igaz $p(y|x_i, \mathbf{s_i'}) \neq p(y|\mathbf{s_i'})$. Egy jegy releváns, ha vagy gyengén vagy erősen releváns, máskülönben irreleváns.*

Ezzel szemben az asszociáció a feltétel nélküli statisztikai függéshez köthető [UC14].

2. Definíció (Asszociáció)**.** *Asszociációs kapcsolat áll fenn Z és W változók között, ha $p(Z, W) \neq p(Z)p(W)$, vagyis Z és W statisztikailag függőek.*

Mindezek miatt a feltételes valószínűségeloszlás alapú relevancia mértékek és asszociáción alapuló (asszociációs) mértékek a relevanciát különböző nézőpontból írják le. A megközelítések közötti különbség miatt Z és W közötti asszociáció nem vonja maga után azt, hogy Z erősen releváns W szempontjából, és fordítva, azaz ha W erősen releváns Z szempontjából, az nem jelenti azt feltétlenül, hogy W és Z között asszociációs kapcsolat áll fenn.

Mivel a fő célkitűzések egyike a relevancia analízis kiterjesztése volt a többváltozós függőségi kapcsolatok (hatáserősséget is figyelembe vevő) részletes elemzésével, ezért számos relevancia feltáró módszer vizsgálatára került sor, melyek alapján három lényegi megközelítés különböztethető meg:

Asszociáció alapú (parametrikus) megközelítés.
Hatáserősség mértékekre helyezi a hangsúlyt, melyek kvantifikálják egy változó másik változóra vonatkoztatott predikciós erejét [Agr02; HTF01]. A változók közötti kapcsolatok jellegét nem vizsgálja.

Strukturális megközelítés.
A változók közötti függőségi kapcsolatok mintázatának feltárására és a kapcsolatok típusainak azonosítására helyezi a hangsúlyt. Az idetartozó módszerek Bayes-hálók strukturális tulajdonságait használják fel a kapcsolatok reprezentálására a rendelkezésre álló adat alapján [Nea04; FK03; Hec99; SGS01]. Ez a megközelítés a kapcsolatok kvalitatív megismerésére összpontosít, azaz egy strukturális tulajdonság meglétére, vagyis a relevancia strukturális aspektusát vizsgálja.

Kauzális megközelítés.
Egy kauzális (funkcionális) modellt feltételez, amely ok-okozati kapcsolatokat definiál a változók között, és megadja egy változó hatásának jellegét más változó(k)ra. A strukturális egyenlet modellezés (structural equation modeling) [Pea00] és más hasonló módszerek a kauzális relevanciára helyezik a hangsúlyt, ami egyrészt egzisztenciális abban a tekintetben, hogy két változó között vagy van ok-okozati kapcsolat vagy nincs, másrészt parametrikus, mivel számszerűsíti az egyik változó által a másikon okozott változás mértékét (a változó értékei tekintetében).

E megközelítések közül az asszociáció alapú módszereket alkalmazzák a leggyakrabban, mivel többnyire egyszerű, egyváltozós hatáserősség mértéket szolgáltatnak, amely a relevancia parametrikus (prediktív) aspektusát ragadja meg. Ezzel szemben a strukturális bizonytalanságon alapuló módszereket akkor alkalmazzák, amikor a függőségi kapcsolatok modelljére, illetve a többváltozós függőségi kapcsolatok elemzésére van szükség. Ezen módszerek hátránya az, hogy csak a relevancia egy adott aspektusára fókuszálnak, holott gyakorlati esetekben mind a strukturális, mind a parametrikus aspektus ismeretlen. Ilyen helyzetekben több relevancia aspektus integrált elemzése, azaz például a strukturális és a parametrikus aspektusé részletesebb képet képes adni a változók közötti függőségi kapcsolatokról. A Bayes-hálók, mint átlátható (white-box) modellosztály egy ideális jelölt erre a célra, mivel rendelkeznek egy G irányított aciklikus gráf struktúrával és egy kapcsolódó θ parametrizációs réteggel.

A létező struktúratanuló algoritmusok [Hec99; HMC97; CH92; FGW99; ATS03; Pen+07] lehetővé teszik a strukturális relevancia elemzését különböző strukturális tulajdonságok azonosítása által, azonban figyelmen kívül hagyják a parametrikus aspektust. [2]

Azonban a parametrizációs réteg felhasználható lenne a hatáserősség, vagyis a változók parametrikus relevanciájának meghatározására.

1. Tézis ***Javaslom olyan új bayesi relevancia mértékek alkalmazását, melyek Bayes-háló alapú Bayes-statisztikai keretrendszerre épülnek, és lehetővé teszik a relevancia strukturális és parametrikus aspektusainak az együttes elemzését.***

E tézisben a Bayes-hálók strukturális és parametrikus tulajdonságainak felhasználását javaslom a változók hatáserősségének meghatározására egy Bayes-statisztikai keretrendszerben. Elsőként javaslom egy hibrid bayesi hatáserősség mérték, a *strukturális feltételű bayesi esélyhányados* (structure conditional Bayesian odds ratio - SC-BOR) alkalmazását, amely egyesíti a relevancia strukturális és parametrikus aspektusait (1.1 altézis). Másodsorban javaslom a Markov-takaró gráfok (Markov blanket graph - MBG) alkalmazását a relevancia strukturális aspektusának reprezentálására az SC-BOR mértékben (1.2 altézis), továbbá javaslok egy algoritmust a *Markov-takaró alapú bayesi hatáserősség* mérték (MBG-based Bayesian odds ratio - MBG-BOR) számítására. Harmadlagosan javaslom az MBG-OR többváltozós kiterjesztését, ami lehetővé teszi több változó együttes hatásának vizsgálatát (1.3 altézis).

Az 1. Tézishez kapcsolódó eredmények bemutatására a disszertáció 4. fejezetében kerül sor. A kapcsolódó publikációk a következők: [1], [2], [4], [6], [7], [11].

1.1 Altézis: Strukturális feltételű bayesi esélyhányados (SC-BOR)

Javaslom a hibrid bayesi hatáserősség mérték, a strukturális feltételű bayesi esélyhányados $\mathrm{OR}(X_i, Y|\theta, G)$ alkalmazását, amely a háttérben lévő (függőségi viszonylatokat leíró) Bayes-háló $\mathbf{BN}(G, \theta)$ mindkét komponensére, azaz a G gráfstruktúrára és a hozzátartozó θ parametrizációra egyaránt támaszkodik.

1. Javaslat. *Annak érdekében, hogy együttesen lehessen vizsgálni a relevancia strukturális és parametrikus aspektusait javaslom, hogy X_i változó hatáserősségének a kiszámítása csak olyan G_j struktúrák figyelembe vételével történjen, ahol X_i strukturálisan releváns, azaz erősen releváns.*

[2] Pontosabban a parametrikus réteg analitikusan ki van átlagolva, ami akkor lehetséges, ha a paraméter függetlenségi feltevés (parameter independence assumption) megállja a helyét [HGC95]. Ez Dirichlet paraméter priorok (lásd 5.4 alfejezet) alkalmazása és teljes (hiányzásmentes) adat esetén valósulhat meg. Máskülönben, például ha a paraméter priort egy szakértő határozza meg egyedileg, vagy ha az adat nem teljes, akkor a paraméterek analitikus kezelése nem lehetséges.

Az erős relevancia Bayes-háló alapú értelmezése a Markov-takaró halmazokhoz köthető [Pea88]:

3. Definíció (Markov-takaró halmaz). *Egy* $\mathbf{M}$ *változóhalmazt, amire fennáll* $\mathbf{M} \subseteq \mathbf{V}$, $\mathbf{V} = \{X_1, X_2, ..., X_n\}$, *azt* Y *változó* Markov-takaró halmaz*ának nevezzük a* $p(\mathbf{V})$ *eloszlás tekintetében, ha* $\perp\!\!\!\perp(Y, \mathbf{V} \setminus \mathbf{M}|\mathbf{M})$ *teljesül, ahol* $\perp\!\!\!\perp$ *a feltételes függetlenséget jelöli.*

A Markov-takaró halmazok és a Bayes-hálók más strukturális tulajdonságainak a jelentősége az, hogy képesek strukturális relevanciára vonatkozó információt megragadni. A minimális Markov-takaró és az erős relevancia közötti kapcsolatot Tsamardinos et al. teremtette meg [TA03], ami megadja azokat a feltételeket, amelyek mellett a releváns strukturális tulajdonságok egyértelműen reprezentálhatók. Feltéve az egyértelmű reprezentációt, egy adott G struktúra esetén Y szempontjából az összes erősen releváns X_i változó része Y Markov-takaró halmazának, melyet $\mathrm{MBS}(Y, G)$ jelöl.

Tehát $\mathrm{MBS}(Y, G)$ erősen releváns változók halmaza, ami felhasználható egy páronkénti reláció definiálására.

4. Definíció (Markov-takaró tagság). *Az* $\mathrm{MBM}(X_i, Y)$ *páronkénti relációt, amely azt jelzi, hogy* X_i *tagja-e* $\mathrm{MBS}(Y, G)$*-nek (azaz* Y *Markov-takaró halmazának) Markov-takaró tagságnak nevezzük.*

A strukturális feltételű bayesi esélyhányados (SC-BOR) a Markov-takaró tagsági reláció felhasználásával az alábbi módon definiálható:

1. Javasolt definíció (Strukturális feltételű bayesi esélyhányados). *Jelölje a Markov-takaró tagság indikátorfüggvényét egy adott* G *struktúra esetén* $I_{\mathrm{MBM}(X_i,Y|G)}$*, ami csak akkor veszi fel az 1 értéket, ha* $X_i \in \mathrm{MBS}(Y, G)$*, és 0 máskülönben. Ekkor a strukturális feltételű esélyhányados* $\mathrm{OR}(X_i, Y|I_{\mathrm{MBM}(X_i,Y|G)})$ *egy véletlen változó* $P(\mathrm{OR}(X_i, Y|I_{\mathrm{MBM}(X_i,Y|G)}))$ *valószínűségi eloszlással, amely az alábbi kifejezés szerint számítható*

$$P(\mathrm{OR}(X_i, Y|I_{\mathrm{MBM}(X_i,Y|G)})) = \frac{P(\mathrm{OR}(X_i, Y, I_{\mathrm{MBM}(X_i,Y|G)}))}{P(I_{\mathrm{MBM}(X_i,Y|G)})}. \tag{2}$$

Ez azt jelenti, hogy az esélyhányados számítása minden lehetséges G struktúrára megtörténik, de csak azok járulnak hozzá érdemben a $p(\mathrm{OR}(X_i, Y|I_{\mathrm{MBM}(X_i,Y|G)}))$ eloszláshoz, amelyeknél $I_{\mathrm{MBM}(X_i,Y|G)} = 1$, azaz X_i erősen releváns az adott struktúrában.

1.2 Altézis: MBG-alapú bayesi esélyhányados (MBG-OR)

Az SC-BOR megvalósítható struktúrák és paraméterek feletti bayesi modellátlagolással számításintenzív Markov-lánc Monte Carlo szimuláció segítségével [Mad+96]. Azonban a lehetséges struktúrák számának és azok lehetséges parametrizációinak tekintetében ez a számítás jelentős mértékben redundáns.

A teljes struktúra tanulása (és az abból való mintavételezés) helyett javaslom a paraméterek mintavételezését a Bayes-háló 'releváns részére' alapozva megvalósítani. A strukturális relevancia szempontjából a *Markov-takaró gráf* (Markov blanket graph - MBG) az a strukturális tulajdonság, amely az összes erősen releváns elemet tartalmazza.

5. Definíció (Markov-takaró gráf). *„Egy* Y *változó Markov-takaró gráfja* $\mathrm{MBG}(Y, G)$ *egy olyan részgráfja a* G *Bayes-háló struktúrának, ami tartalmazza* Y *Markov-takaró halmazának* $\mathrm{MBS}(Y, G)$ *csomópontjait, valamint* Y *változóba és gyermekeibe befutó éleket. Egy célcsomópontra nézve, ami* Y *célváltozónak feleltethető meg, az* $\mathrm{MBG}(Y, G)$ *részgráf azon csomópontokból áll, melyek (1)* Y *szülei, (2)* Y *gyermekei vagy (3)* Y *gyermekeinek további szülei" [27].*

2. Javaslat. *A strukturális feltételű bayesi esélyhányados kiszámítható a lehetséges Markov-takaró gráfok* a posteriori *valószínűségi eloszlását felhasználva, ahol a Markov-takaró gráfok felparaméterezése az adat alapján történik.*

Mindez egy gyakorlatban alkalmazható bayesi hibrid hatáserősség mérték kialakításához: a Markov-takaró alapú bayesi esélyhányadoshoz (MBG-based Bayesian odds ratio) vezet.

2. Javasolt definíció (Markov-takaró gráf alapú bayesi esélyhányados)**.** *Az MBG-alapú bayesi esélyhányados (MBG-BOR) a lehetséges MBG-k alapján számított esélyhányadosok feletti átlagolt értékként számítható a következőképpen:*

$$\text{MBG-BOR}(X_i, Y|D) = \sum_{j=1}^{m} \text{OR}(X_i, Y|\,\text{MBG}_j(Y,G)) \cdot p(\text{MBG}_j(Y,G)|D) \cdot I_{(X_i \in \text{MBG}_j(Y,G))},$$

ahol m *azon* MBG*-k száma, melyek esetében az* a posteriori *valószínűség* $p(\text{MBG}_j(Y,G)|D) > 0$*. Az indikátor függvény* $I_{(X_i \in \text{MBG}_j(Y,G))}$ *az* 1 *értéket veszi fel, ha* $X_i \in \text{MBG}_j(Y,G)$*, és* 0 *máskülönben.*

Az MBG-BOR-t megvalósításának lépéseit az 1. Algoritmus írja le.

Algorithm 1 Az $\text{MBG-BOR}(X_i, Y)$ számítása

Require: $n, m, \text{MBG}(Y,G), D$
  for $\text{MBG}_{1\ldots n}$ **do**
    for $\theta_{1\ldots m}$ **do**
      draw parametrization $\theta_k = (X_{k1} = x_{k1}, \ldots, X_{kr} = x_{kr})$
      for all $X_k \in \text{MBG}_j$, so that $X_k \neq X_i$.
      estimate $P(Y = 0|X_i = x_i, \theta_k)$
      compute $\text{Odds}(X_i = x_i, \theta_k) = \frac{P(Y=1|X_i=x_i,\theta_k)}{P(Y=0|X_i=x_i,\theta_k)}$
      compute $\text{OR}(X_i, \theta_k) = \frac{\text{Odds}(X_i=x_i^1,\theta_k)}{\text{Odds}(X_i=x_i^0,\theta_k)}$
    end for
    compute $\text{OR}(X_i|\,\text{MBG}_j) = \sum_{\theta_k=1}^{m} \text{OR}(X_i, \theta_k)$
    update $\text{OR}_{\text{histogram}}(X_i)$
  end for
  $\text{MBG-BOR}(X_i, Y|D) = \sum_{\text{MBG}_j=1}^{n} \text{OR}(X_i, Y|\,\text{MBG}_j) \cdot p(\text{MBG}_j\,|D)$
  calculate credible interval for $\text{MBG-BOR}(X_i, Y)$ based on $\text{OR}_{\text{histogram}}(X_i)$

1.3 Altézis: MBG-alapú többváltozós bayesi hatáserősség

Az MBG-alapú bayesi esélyhányados mérték kiterjeszthető változók halmazára, amely lehetővé tesz több változó együttes hatásának vizsgálatát.

3. Javasolt definíció. *Feltéve a változók egy* $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$ *halmazát a többváltozós* MBG-BOR *az alábbiak szerint számítható*

$$\text{MBG-BOR}^*(\mathbf{V}, Y) = \sum_{j=1}^{m} \text{OR}(\mathbf{V}, Y|\,\text{MBG}_j(Y,G)) \cdot p(\text{MBG}_j(Y,G)) \cdot I^*_{(\mathbf{V} \in \text{MBG}_j(Y,G))}, \quad (3)$$

ahol az indikátorfüggvény $I^*_{(\mathbf{V} \in \text{MBG}_j(Y,G))}$ *az* 1 *értéket veszi fel, ha bármely* $X_i \in \mathbf{V}$ *változóra igaz, hogy* $X_i \in \text{MBG}_j(Y,G)$*, és* 0 *máskülönben.*

Ennek megfelelően, az MBG-alapú odds a változók egy $\mathbf{V}$ halmazára a következő:

$$Odds^{*}_{\mathrm{MBG}_j(Y,G)}(\mathbf{V}, Y) = \frac{p(Y = 1|\,\mathrm{MBG}_j(Y,G), X_{n1} = x_{n1}, \ldots, X_{nr} = x_{nr})}{p(Y = 0|\,\mathrm{MBG}_j(Y,G), X_{n1} = x_{n1}, \ldots, X_{nr} = x_{nr})}, \qquad (4)$$

ahol $x_{n1} \ldots x_{nr}$ az $X_{ni} \in \mathbf{V}$ változók értékei, melyek elemei $\mathrm{MBG}_j(Y,G)$-nek.

5.3. A priori tudás által vezért bayesi relevancia mértékek

A klasszikus statisztika hipotézistesztelő keretrendszere lehetővé teszi az *a priori* tudás konzisztens hipotézisekké történő átalakítását, majd vizsgálatát [Bor98]. Azonban a többszörös hipotézistesztelés (multiple hypothesis testing - MHT) miatti korrekció limitálja e keretrendszer alkalmazhatóságát olyan tárgyterületek esetén, melyeket komplex, többváltozós függőségi mintázatok jellemeznek [Nob09]. Az orvosbiológiai tárgyterületeken ez egy jelentős korlátozó tényező, holott egy alkalmas hipotézis az adatelemzési folyamat számára lényeges iránymutatást jelentene.

Számos klasszikus statisztikai [BH95; Dun61] és bayesi [Sto02] megoldás született az MHT probléma kezelésére. Viszont a többváltozós bayesi módszerek eleve egy „beépített" MHT korrekcióval rendelkeznek [GHY12], [NC06]. Például a bayesi relevancia analízis esetében egy strukturális tulajdonság *a posteriori* valószínűsége a bayesi modellátlagolás eredményeként jön létre, azaz számos lehetséges modellt (amely különböző hipotéziseknek feletethető meg) vesz figyelembe, ahelyett, hogy egyetlen legjobb hipotézist választana. Ez egyfelől azt jelenti, hogy a bayesi keretrendszer lehetővé teszi a függési kapcsolatok feltárását egy kiemelt modell nélkül, másfelől a korrekció abból adódik, hogy olyan lehetséges modelleket is figyelembe vesz, melyekben a vizsgált tulajdonság nem releváns.

Olyan esetekben, amikor az a priori tudás elérhető a többváltozós bayesi módszerek feltudják azt használni különböző struktúra és paraméter priorok formájában [HGC95], [PS12]. Azonban az *a priori* tudásnak léteznek olyan aspektusai, amelyet nem lehet struktúrához vagy paraméterhez kötődő *a priori* eloszlássá ('priorrá') transzformálni, mert a kutatás egy más fázisához tartoznak. Pontosabban az eredmények kiértékelési és jelentési fázisához tartozó tudás nem feltétlenül alkalmazható a feltáró elemzési fázisban. Ebben az esetben kiértékelés alatt az eredmények értelmezését és további felhasználását (transzlációját) kell érteni. Például egy vizsgálat eredményeit összevethetjük egy korábbi vizsgálat eredményeivel, benchmarkokkal (viszonyítási standardokkal), vagy más a tárgyterületre vonatkozó tudással, ami elérhető a kapcsolódó szakirodalomban. Ekkor az eredmények értelmezése (elvárásnak megfelelő vagy váratlan az eredmény, elfogadható vagy esetleg hibás) attól függ, hogy milyen referenciákhoz hasonlítjuk őket. A folyamat következő lépéseként az eredményeket felhasználják a kutatással kapcsolatos döntéshozatalban, azaz például el kell dönteni, hogy publikálják-e az eredményeket, tervezzenek-e új kísérletet, vagy kipróbáljanak-e egy alternatív megközelítést. A kutatók gyakran szembesülnek ilyen jellegű döntésekkel, mégis ezek a döntések jellemzően informálisak maradnak.

Erre egy lehetséges megoldás egy döntéselméleti keretrendszer alkalmazása, amely veszteségfüggvényeket és azokat implementáló mércéket használ fel. A veszteségfüggvények speciális hasznosságfüggvények, melyek egy adott szcenárió által okozott veszteséget határozzák meg [RN10]. Például egy kísérletileg validált (az adott területen elismerten releváns) változó hatáserősségének ismeretében létrehozható egy olyan veszteségfüggvény, ami ezt az értéket használja küszöbértékként annak eldöntésére, hogy más változók relevánsak-e parametrikusan vagy sem (egy változó ebben az esetben akkor lesz parametrikusan releváns, ha hatáserőssége az adott célváltozóra nézve legalább akkora, mint a küszöbérték).

Az *a priori* tudás egy másik lehetséges formája a releváns kontextus ismerete. Komplex mechanizmusok esetén a kontextualitás releváns szerepet játszhat, mivel lehetséges, hogy bizonyos hatások csak egy speciális környezetben jelennek meg [WB10], [Bou+96]. Például bizonyos genetikai faktorok hatása csak magas szintű stresszkitettség esetén érvényesül [6]. Ilyen esetekben a kontextualitással kapcsolatos tudás, vagyis annak ismerete, hogy egy hatás feltehetőleg kontextuális, illetve hogy mely változók alkothatják a kontextust, alapvető fontosságú, mivel lehetséges, hogy általános, kontextustól független mércék alkalmazásával a hatás nem detektálható. Például ha egyes függőségek csak az adat egy részében vannak jelen (azaz egy alpopulációban), akkor az asszociációs mércék lehet, hogy nem képesek detektálni, ha ez a részhalmaz relatíve kisméretű a teljes adathalmazhoz képest, amin a függőségeket vizsgálják. A kontextuális függőségek vizsgálatához tehát szükség van kontextusérzékeny veszteségfüggvényekre és ahhoz kapcsolódó mércékre.

A kiértékelés-specifikus *a priori* információ döntéselméleti keretrendszerben való felhasználásához szükség van specializált mértékekre, melyek közvetlenül támaszkodnak az *a priori* tudásra és alkalmazhatóak a meglévő bayesi keretrendszerben.

2. Tézis ***Egy egzisztenciális relevancia alapú bayesi, döntéselméleti mérték alkalmazását javaslom, amely a Bayes-hálók parametrikus tulajdonságaira támaszkodva lehetővé teszi kiértékelés-specifikus*** **a priori** ***tudás alkalmazását.***

E tézisben egy új megközelítést javaslok az egzisztenciális relevancia[3] mérésére. Javaslom a vizsgált változó (célváltozó szempontjából tekintett) bayesi hatáserősség eloszlásának a felhasználását a vizsgált változó (célváltozó szempontjából tekintett) egzisztenciális relevanciájának vizsgálatához. Erre a célra egy új mennyiséget, a *hatáserősség alapú egzisztenciális relevancia* mértéket (effect size conditional existential relevance - ECER) hoztam létre (2.1 altézis). Az ECER lehetővé teszi a kiértékelés-specifikus *a priori* tudás felhasználását, egy *a priori* definiált *elhanyagolható hatáserősség intervallum* (interval of negligible effect size - ϵ) megadásával.

A kontextuálisan releváns függőségi kapcsolatok feltárásának elősegítésére javaslom a ECER kontextuális kiterjesztését: C-ECER, amely lehetővé teszi a releváns kontextushoz kötődő *a priori* tudás felhasználását (2.2 altézis).

A 2. Tézis eredményeinek a bemutatására a disszertáció 4. fejezetében kerül sor. A kapcsolódó publikációk a következők: [5], [10], [12].

2.1 Altézis: Hatáserősség alapú egzisztenciális relevancia (ECER)

Létrehoztam egy bayesi egzisztenciális relevancia mértéket, a hatáserősség alapú egzisztenciális relevanciát (ECER), amely a bayesi hatáserősség eloszláson $p(\mathrm{OR}(X_i, Y|\Theta))$ alapszik. Megmutattam, hogy az ECER lehetővé teszi a kiértékelés-specifikus *a priori* tudás közvetlen alkalmazását egy szakértői megítélés szerint elhanyagolható hatáserősségek intervallumának megadásával.

Az ECER formális definiáláshoz szükség van egy C_ϵ elhanyagolható hatáserősségek intervallumának megadására, amely *a priori* tudás alapján például egy szakértő által meghatározható azzal a megkötéssel, hogy a semleges esélyhányadost (odds ratio = 1) tartalmaznia kell az intervallumnak. Például egy $\epsilon = 0,3$ méretű C_ϵ intervallum megadható a semleges esélyhányados körül szimmetrikusan $C_\epsilon = [0,85; 1,15]$, aszimmetrikusan $C_\epsilon = [0,9; 1,2]$, vagy egyoldalúan $C_\epsilon = [1,0; 1,3]$.

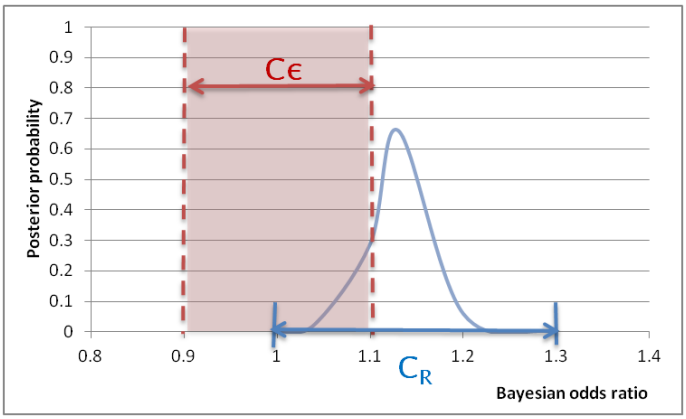
[3] A strukturális relevancia egy korlátozott formája.

4. Javasolt definíció (Hatáserősség alapú egzisztenciális relevancia - ECER)**.** *Feltéve, hogy adott egy C_ϵ elhanyagolható hatáserősségek intervalluma $\epsilon \geq 0$ mérettel, és az X_i változó Y célváltozóra vonatkozó hatáserősségének eloszlása $P(\mathrm{OR}(X_i, Y|\Theta))$, jelölje az $I_{\{ECER_\epsilon(X_i,Y)\}}$ indikátorfüggvény azt, hogy $\mathrm{OR}(X_i, Y|\Theta) \notin C_\epsilon$, azaz X_i változó hatáserőssége releváns, mivel kívül van a C_ϵ intervallumon. Ekkor a bayesi keretrendszerben a hatáserősség alapú egzisztenciális relevancia $ECER_\epsilon(X_i, Y)$* a posteriori *valószínűsége a $p(I_{\{ECER_\epsilon(X_i,Y)\}})$ kifejezés által definiálható [10].*

Az $\mathrm{OR}(X_i, Y|\Theta)$ kifejezés egy véletlen változó, amely X_i változó Y célváltozóra vonatkozó bayesi hatáserősségét jelöli, melynek eloszlását $P(\mathrm{OR}(X_i, Y|\Theta))$ jelöli, ahol Θ a lehetséges parametrizációkat reprezentáló véletlen változó.

3. Javaslat. *A $P(\mathrm{OR}(X_i, Y|\Theta))$ azon része, mely a C_ϵ intervallum és az $\mathrm{OR}(X_i, Y|\Theta)$ bayesi hatáserősség megbízhatósági intervallumának metszetében található, kvantifikálja X_i változó Y célváltozóra vonatkoztatott (szakértői tudáson alapuló) parametrikus irrelevanciáját [10].*

A C_ϵ tehát meghatározza az irreleváns hatáserősségek intervallumát vagyis a parametrikus irrelevanciát. Ha az $\mathrm{OR}(X_i, Y|\Theta)$ megbízhatósági intervalluma metszi a C_ϵ intervallumot, akkor az azt jelenti, hogy X_i parametrikusan releváns egy adott mértékben. Minél nagyobb $P(\mathrm{OR}(X_i, Y|\Theta) \in C_\epsilon)$ (az eloszlás C_ϵ-ba eső része), annál kevésbé releváns parametrikusan X_i (lásd illusztrációként az 2 ábrát).



2. ábra. Egy példa a C_ϵ szakértői megítélés szerint elhanyagolható hatáserősségek intervalluma és az X_i változó bayesi hatáserősségének *a posteriori* eloszlásához tartozó C_R megbízhatósági intervalluma közötti lehetséges viszonyra. Az illusztrált esetben C_R metszi C_ϵ-t, így X_i irreleváns parametrikusan egy adott mértékben a megadott C_ϵ esetében.

Ha $\mathrm{OR}(X_i, Y|\Theta)$ megbízhatósági intervalluma nem metszi C_ϵ-t, akkor X_i változó parametrikusan releváns, mivel a hatáserősség eloszlás $P(\mathrm{OR}(X_i, Y|\Theta))$ csak nem elhanyagolható értékekből áll.

4. Javaslat. *Ha $\mathrm{OR}(X_i, Y|\Theta)$ megbízhatósági intervalluma nem metszi C_ϵ-t akkor X_i változó 'egzisztenciálisan releváns' Y célváltozóra vonatkoztatva [10].*

Az a tény, hogy $p(\mathrm{OR}(X_i, Y|\Theta) \in C_\epsilon) = 0$ azt jelenti, hogy az adott C_ϵ mellett nincs parametrikusan kódolt függetlenség X_i és Y között. Feltéve egy függőségi kapcsolatokat leíró $\mathrm{BN}(G, \theta)$ Bayes-hálót (ahol G a gráfstruktúra és θ a kapcsolódó parametrizáció), amelynél G hűen reprezentál minden feltételes függetlenséget (a Markov-feltevés alapján [Nea04]), ez azt jelenti, hogy

X_i és Y nemcsak parametrikusan függenek, hanem strukturálisan is. Azonban a strukturális relevancia pontos típusa nem határozható meg egyedül $\mathrm{OR}(X_i, Y|\Theta)$ alapján, vagyis hogy X_i és Y között közvetlen kapcsolat áll fenn, vagy például interakció. Mivel csak a létezése (egzisztenciája) jelenthető ki az X_i és Y közötti strukturális kapcsolatnak ezért a strukturális relevancia e fajtáját *egzisztenciális relevanciának* nevezzük.

2.2 Altézis: Az ECER kontextuális kiterjesztése

Korábban a kontextualitás fogalmát a kontextuális irrelevancia nézőpontjából definiáltuk [12], vagyis egy $\mathbf{C} = \mathbf{c}$ kontextus esetén az X_i független Y-tól, máskülönben függőek.

A jelenlegi célkitűzés folytán javaslom kontextualitást a kontextuális relevancia nézőpontjából definiálni.

5. Javasolt definíció (Kontextuális relevancia)**.** *Tételezzük fel, hogy $X_i \cup \mathbf{C}$ irreleváns Y viszonylatában, vagyis $\perp\!\!\!\perp(Y, (X_i \cup \mathbf{C}))$, és $(X_i \cap \mathbf{C} = \emptyset)$. Ekkor X_i kontextuálisan releváns, ha létezik egy adott kontextus $\mathbf{C} = \mathbf{c}$, amire igaz $\not\perp\!\!\!\perp(Y, X_i|\mathbf{c})$ [10].*

Feltételezve egy *a priori* megadott $\mathbf{C}$ kontextust (formáló változóhalmazt), **javaslom az ECER kontextuális kiterjesztését (C-ECER) a kontextuálisan releváns függőségek feltárása céljából**.

6. Javasolt definíció (Kontextuális ECER: C-ECER)**.** *Feltéve, hogy adott C_ϵ az elhanyagolható hatáserősségek intervalluma $\epsilon \geq 0$ mérettel, és X_i változó Y-ra vonatkoztatott hatáserősségének az eloszlása $P(\mathrm{OR}(X_i, \mathbf{C} = \mathbf{c_j}, Y|\Theta))$, és egy $\mathbf{C}$ kontextus formáló változóhalmaz $\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_r}$ lehetséges értékekkel; jelölje $I_{\{ECER_\epsilon(X_i, \mathbf{C}=\mathbf{c_j}, Y)\}}$ indikátorfüggvény, hogy $\mathrm{OR}(X_i, \mathbf{C} = \mathbf{c_j}, Y|\Theta) \notin C_\epsilon$, azaz X_i változó hatáserőssége adott $\mathbf{C} = \mathbf{c_j}$ kontextus mellett releváns, mivel a C_ϵ intervallumon kívül található.*

Ekkor jelölje $I_{\{C-ECER_{\epsilon,\mathbf{C}=\mathbf{c_j}}(X_i,Y)\}}$ azt, hogy X_i kontextuálisan ECER releváns, ha létezik olyan $\mathbf{C} = \mathbf{c_j}$ kontextus, amire $I_{\{ECER_\epsilon(X_i, \mathbf{C}=\mathbf{c_j}, Y)\}} = 1$.

Ez tehát azt jelenti, hogy X_i C-ECER releváns, ha létezik kontextus, amiben ECER releváns: $I_{\{C-ECER_{\epsilon,\mathbf{C}=\mathbf{c_j}}(X_i,Y)\}} = \bigcup_{c_j=c_1}^{c_r} I_{\{ECER_\epsilon(X_i, \mathbf{C}=\mathbf{c_j}, Y)\}}$

5.4. A nem informatív paraméter priorok hatása

A bayesi módszerek alapvető paradigmája, hogy a $P(A|B)$ *a posteriori* valószínűség a Bayes-tétel szerint a $P(A)$ *a priori* eloszlás és a $P(B|A)$ likelihood alapján kiszámítható [Ber95]. Tételezzük fel, hogy adott egy $\mathbf{V} = X_1, ..., X_n$ diszkrét véletlen változók halmaza, melyek együttes valószínűségeloszlása $P(\mathbf{V})$ a $\mathrm{BN}(G, \theta)$ Bayes-háló által hűen reprezentálható. A struktúratanuló módszerek célja azonosítani a legvalószínűbb struktúrákat D adathalmaz alapján, amihez szükség lenne minden egyes G struktúra *a posteriori* valószínűségének a kiszámítására a következők szerint:

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}. \tag{5}$$

Mivel $P(D)$ modellezési konstansnak tekinthető, ezért $P(G|D)$ számításához a lehetséges struktúrák $P(G)$ *a priori* eloszlására és $P(D|G)$ likelihoodra van szükség.

A $P(G)$ kifejezés alkalmas *a priori* tudás kifejezésére, mivel képes azt a feltételezést kódolni, hogy egyes struktúrák valószínűbbek másoknál (ez a becslés származhat például egy tárgyterületi szakértőtől). Alapvetően az *a priori* eloszlások megadása kapcsán két fő megközelítési

módot különböztetjük meg: az informatív [AC08] és a nem informatív [HGC95; GH95; Bun91; CH92] módokat. Az informatív megközelítés lényege, hogy az *a priori* tudást a lehető legteljesebb mértékben fel kívánja használni, például a legvalószínűbb struktúrák keresésénél. Ennek eredményeként az informatív priorok nagy mértékben befolyásolják az elemzést és kiértékelést. A nem informatív megközelítés viszont azt a szemléletet követi, mely szerint az *a priori* tudást nem szabad arra használni, hogy korlátok közé szorítsa az új kutatásokat, mivel elvileg lehetséges, hogy a korábbi tanulmányok, kísérletek nem voltak teljesek, vagy akár tévesek voltak. Emiatt a nem informatív prioroknak korlátozott a hatása az elemzésre. Például a struktúrák *a priori* eloszlása (nem informatív esetben) jellemzően egyenletes, vagyis minden struktúra azonos *a priori* valószínűséggel rendelkezik.

Elviekben a prior hatása fokozatosan csökken, ahogy egyre több adat (evidencia) kerül feldolgozásra. Valós problémák esetében azonban a rendelkezésre álló mintaszám gyakran nem elégséges ahhoz, hogy teljesen felülírja a prior hatását. Ebből kifolyólag még a nem informatív prioroknak is döntő szerep juthat a tanulási folyamat során.

A $P(D|G)$ kifejezés a likelihood (score), amely azt méri, hogy a D adat mennyire valószínű egy adott G struktúra esetén. Erre a célra a *bayesi Dirichlet* (BD) mérték [Bun91] egy gyakori választás, melyet az alábbi fejezetrész részletesen ismertet. E mértéknek vannak szabad paraméterei, az úgynevezett hiperparaméterek, melyeket előre definiálni kell, tehát bizonyos szintű *a priori* tudást igényel a paraméterek tekintetében. Emiatt *paraméter prior*nak nevezzük [HGC95].

3. Tézis *Egy szisztematikus kiértékelési módszert követve azonosítottam a nem informatív priorok bayesi relevancia analízisre kifejtett hatását.*

E tézisben, a nem informatív paraméter priorok meghatározott kritériumok szerinti alkalmazását javaslom. Először összefüggést adtam a bayesi Dirichlet metrika virtuális mintaszám paramétere és az *a priori* hatáserősség eloszlás között (3.1 altézis). Ezt követően egy szisztematikus vizsgálatot végeztem el, amely alapján javaslom, hogy a virtuális mintaszám paraméter megválasztása az *a priori* várható hatáserősség értéke szerint történjen (3.2 altézis).

A 3. tézis eredményeinek bemutatására a disszertáció 5. fejezetében kerül sor. A kapcsolódó publikációk a következők: [2], [3], [7], [8], [9].

3.1 Altézis: A virtuális mintaszám paraméter és a hatáserősség közötti kapcsolat

Ebben az altézisben ismertetem az összefüggést egy W változóhoz tartozó *a priori* paramétereloszlás virtuális mintaszám paramétere és a W változó várható hatáserőssége között [2].

6. Definíció (Prior Dirichlet eloszlás)**.** *Legyen W egy diszkrét véletlen változó, mely multinomiális eloszlással rendelkezik k lehetséges értékkel. Továbbá jelölje ν_i annak a valószínűségét, hogy W a w_i értéket veszi fel, azaz $p(W = w_i)$. Ekkor az* a priori *valószínűségi sűrűségeloszlás a ν_i paraméterek felett W változó esetén Dirichlet eloszlást követ az $\mathbf{R}^{k-1}$ dimenziós euklideszi téren:*

$$Dir(\nu_1, ..., \nu_{k-1}|\alpha_1, ..., \alpha_k) = \frac{1}{\beta(\alpha)} \cdot \prod_{i=1}^{k} \nu_i^{\alpha_i - 1}, \tag{6}$$

ahol a ν_i paraméterhez tartozó virtuális mintaszámot α_i jelöli, és $\beta(\alpha)$ a többváltozós béta függvényt jelöli. A ν_i maximum likelihood becslése $\frac{N_i}{N}$, ahol N_i azon megfigyelések száma, ahol $W = w_i$ és N az adathalmaz mérete.

Megemlítendő, hogy gyakorlati esetekben α_i-kről feltesszük, hogy azonosak minden ν_i paraméterre, ami végül egyetlen α virtuális mintaszám paraméterhez vezet. Annak érdekében, hogy

az esélyhányados logaritmusa szerinti alakban kifejezett hatáserősség *a priori* eloszlását megbecsülhessük, a ν_i paraméterek valószínűségi sűrűségfüggvényét transzformálni kell a esélyhányados logaritmusa számítása szerint.

7. Definíció. *Legyen W változó $k = 2$ lehetséges értékkel és legyen Y egy bináris célváltozó. Ekkor a log odds ratio ν' feltételes valószínűség paraméterek felhasználásával az alábbiak szerint definiálható:*

$$\log \frac{p(Y = y_1|W = w_1)}{p(Y = y_0|W = w_1)} - \log \frac{p(Y = y_1|W = w_0)}{p(Y = y_0|W = w_0)} = \log \frac{\nu'_{1|1}}{\nu'_{0|1}} - \log \frac{\nu'_{1|0}}{\nu'_{0|0}}. \tag{7}$$

Mivel a célváltozó bináris, mindkét az esély logaritmusát leíró kifejezés egyszerűsíthető: $\log \frac{\nu'_{1|1}}{1-\nu'_{1|1}}$ és $\log \frac{\nu'_{1|0}}{1-\nu'_{1|0}}$. Tehát az *a priori* valószínűségi sűrűségfüggvényen elvégzendő transzformáció: $t(\nu) = \log \frac{\nu}{1-\nu}$.

A paraméter prior nem informatív alkalmazása miatt feltételezhető a virtuális mintaszámok uniformitása, azaz $\alpha_1 = \alpha_2 = ...\alpha_k$.

7. Javasolt definíció. *A transzformált Dirichlet eloszlás általános formája uniform virtuális mintaszám paramétereket feltételezve a következő alakban adható meg:*

$$g(z, \alpha) = \frac{Dir(\frac{1}{1+e^{-z}}|\alpha) \cdot e^{-z}}{(1 + e^{-z})^2} \tag{8}$$

Felhasználva azt a tényt, hogy a Béta függvény $\beta(\alpha)$ kifejezhető a Gamma függvény $\Gamma(.)$ segítségével:

$$\beta(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}(\alpha_i))}, \tag{9}$$

a transzformált Dirichlet eloszlás is kifejezhető $\Gamma(.)$ függvények segítségével.

8. Javasolt definíció. *A transzformált Dirichlet eloszlás egyszerűsített formája az alábbiak szerint adható meg:*

$$g(z, \alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \cdot (\frac{1}{1 + e^{-z}})^{\alpha+1} \cdot (1 - \frac{1}{1 + e^{-z}})^{\alpha-1} \cdot e^{-z}. \tag{10}$$

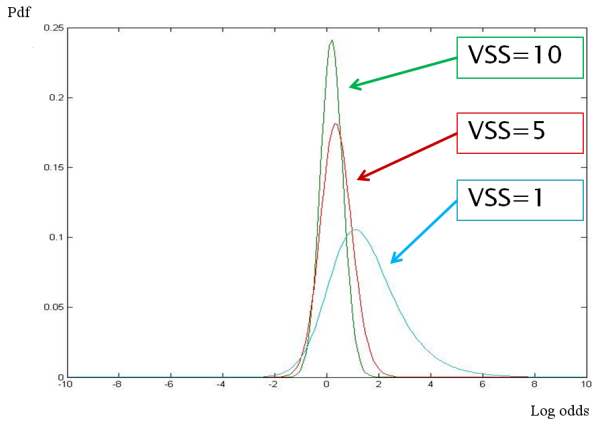
Ez a kifejezés definiálja az esélyek logaritmusának *a priori* eloszlását, és azt mutatja, hogy az α virtuális mintaszám meghatározó szereppel bír.

3.2 Altézis: A virtuális mintaszám paraméter megválasztása

A transzformált $g(z, \alpha)$ eloszlás analitikusan nem jól kezelhető, mivel az eloszlás nagy valószínűségi sűrűségű tartománya nem azonosítható. Viszont az eloszlás mintavételezhető különböző α virtuális mintaszám paraméterek mellett, ami lehetővé teszi a paramétereknek a log odds valószínűségi sűrűségfüggvényére gyakorolt hatásának vizsgálatát [2], [7].

A 3. ábrán látható mérési eredmények azt a korábban elvárt hatást mutatják, hogy a virtuális mintaszám növelésével a megbízhatósági tartománya a esély logaritmusa (log odds) eloszlásának csökken. Mindez azt jelenti, hogy a magas virtuális mintaszám azt az *a priori* valószínűséget fejezi ki, hogy magasabb esély értékek kevésbé valószínűek.

5. Javaslat. *Gyakorlati megközelítésben a bemutatott összefüggés felhasználható a paraméter prior meghatározására a virtuális mintaszám beállításával a várható hatáserősség eloszlása szerint. [2], [7].*



3. ábra. A log odds eloszlása α = 1, 5, és 10 virtuális mintaszám esetén.

6. Az új tudományos eredmények alkalmazása

Ez a fejezet a disszertációban bemutatott eredmények gyakorlati alkalmazását foglalja össze.

6.1. Génasszociációs vizsgálatok eredményeinek elemzése bayesi relevancia analízissel

A Bayes-háló alapú bayesi többszintű relevancia elemzés (röviden *bayesi relevancia analízis*) egy többváltozós, modellalapú bayesi módszer, amely lehetővé teszi a függőségi kapcsolatok elemzését [27],[26], [12]. Alkalmazható általános célú jegykiválasztó módszerként is, de ezen felül lehetővé teszi a komplex függőségi mintázatok részletes analízisét is. A bayesi relevancia analízis képes releváns változók, releváns változók halmazainak és releváns változók interakciós modelljeinek az azonosítására. Ezek a relevancia különböző absztrakciós szintjeihez köthetőek, melyek a háttérben lévő (függőségi kapcsolatokat leíró) Bayes-háló különböző strukturális tulajdonságai *a posteriori* valószínűségének kiértékelése alapján határozhatóak meg (például releváns változók halmazának meghatározásához a Markov-takaró halmazok valószínűségének vizsgálata szükséges). Az *a posteriori* valószínűségek számítása bayesi modellátlagoláson alapul (egy bayes-statisztikai keretrendszerben).

A 3. tézisben ismertetett eredmények szorosan kapcsolódnak a bayesi relevancia analízis gyakorlati alkalmazásához, különösen jelölt génasszociációs vizsgálatok elemzésénél (CGAS - candidate gene association study). Az 1. és a 2. tézisben bemutatott új bayesi hatáserősség mértékek pedig a bayesi relevancia analízis kiterjesztésének tekinthetők, melyek kiemelt alkalmazási területe szintén a jelölt génasszociációs vizsgálatok elemzése.

A génasszociációs vizsgálatok genetikai faktorok, legtöbbször egynukleotidos polimorfizmusok (SNP) és különféle betegségek közötti kapcsolatok feltárására irányul [Hir+02]. Ennek keretében egészséges és a vizsgált betegségben szenvedő páciensek biológiai mintái alapján genetikai faktorokat mérnek meg, és azonosítják a variánsokat. Ezt követően a mérési eredményeket integrálják egy adathalmazba, amelyen statisztikai elemzést hajtanak végre. Az analízis célja azon genetikai faktorok azonosítása, melyek statisztikai függőségi kapcsolatban állnak a célváltozóval, amely tipikusan egy betegségleíró [SB09; Lew02]. Összetett GAS esetén további klinikai, környezeti és fenotípusos változók (a betegséghez kapcsolódó szimptómák, jellemzők, vagy a populációra jellemző tulajdonságok) halmazait mérik meg és értékelik ki [EN14]. Mindezt annak érdekében, hogy elősegítsék a multifaktoriális (azaz több géncsoport és környezeti tényező

együttese által indukált) betegségek (úgymint asztma, allergia, rheumatoid arthritis) komplex genetikai hátterének feltárását.

Egy korábbi mesterséges adaton elvégzett összehasonlító elemzés keretében igazolást nyert, hogy a bayesi relevancia analízis egy alkalmas jegykiválasztó módszer, amely képes a releváns változók meghatározására jelölt génasszociációs vizsgálatok elemzése esetén [25].

Ezt követően a Semmelweis Egyetem, Genetikai, Sejt- és Immunbiológiai intézete (SE-DGCI) és a Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszéke kutatócsoportjainak közös kutatási projektje keretében végzett asztma jelölt génasszociációs vizsgálat elemzéséhez alkalmaztam a bayesi relevancia analízist [3]. A sikeres alkalmazást követően a módszert számos további CGAS-ban használtuk fel az alábbi kutatócsoportokkal együttműködve:

- *Genetikai, Sejt- és Immunbiológiai Intézet, Semmelweis Egyetem (SE-DGCI).* Az Intézetben végzett jelölt génasszociációs vizsgálatokba bekapcsolódva alkalmaztam a disszertációban ismertetett bayesi módszereket, a *leukémia* [4], továbbá a *rheumatoid arthritis* [7], [14], [15], [16] kutatási területein.

- *Orálbiológiai Tanszék, Semmelweis Egyetem.* A bayesi relevancia analízis alkalmazására az Orálbiológiai Tanszék által vezetett két jelölt génasszociációs vizsgálatban került sor, melyek a hypodontia [17] és a parodontitis kórképek genetikai hátterének magyar populáción történő vizsgálatára irányultak.

- *MTA-SE Neuropszichofarmakológiai és Neurokémiai Kutatócsoport (MTA-SE-NNRG).* A MTA-SE-NNRG célkitűzése a depresszióhoz kötődő releváns genetikai és környezeti faktorok kutatása. A BME-MIT kutatócsoportjával együttműködésben több vizsgálat zajlott ezen a területen [1], [5], [6]. A bayesi relevancia analízis alkalmazásával azonosítottunk egy - a depresszió szempontjából releváns - genetikai tényezőcsoportot, a Galanin génrendszert [5], amely egy új gyógyszercélpontként szolgálhat a betegség kezelésében.

  A depresszió kialakulásához kapcsolódó komplex mechanizmusok vizsgálata tette szükségessé a kontextuális megközelítést a statisztikai vizsgálatok kapcsán. Ez szolgált a 2. tézishez kapcsolódó kutatás kiindulási alapjául. Az eredmények alkalmazására egy depresszióhoz kapcsolódó vizsgálatban került sor [5], [10].

Az 1. tézisben tárgyalt eredményeket a következő génasszociációs vizsgálatokban alkalmaztam: az akut limfoid leukémia [4] és a rheumatoid arthritis [7] genetikai hátterének vizsgálatában (SE-DGCI által vezetett kutatás), továbbá a depressziót befolyásoló genetikai tényezők vizsgálatában [6] és egy impulzivitást vizsgáló esettanulmányban [1] (MTA-SE-NNRG által vezetett kutatás).

Mindezek alapján irányelveket javasoltam a bayesi relevancia analízis jelölt génasszociációs vizsgálatokban történő alkalmazásához, melyek az alkalmazást érintő gyakorlati megfontolásokat rendszerezték [7].

6.2. Bioinformatikai célú statisztikai elemzőeszközök fejlesztése

Az 1. és a 3. tézisek eredményeinek alkalmazására az alábbi projektekben került sor:

- *GENAGRID projekt.* A GENAGRID projekt egyik fő célja bioinformatikai alkalmazások részére történő elemzési eszközök és módszerek fejlesztése volt. A BME-MIT és a SE-DGCI kutatócsoportjai együttműködésének keretében a bayesi relevancia analízist és annak új kiterjesztéseit alkalmaztam a projekthez kötődő vizsgálatokban [3], [4], [25], [11].

- *KOBAK projekt.* A KOBAK projekt célja a bioinformatika és biotechnológia oktatásának fejlesztése volt tananyagok, tankönyvek készítése [20] [21] [22] [23] [24], valamint szoftvereszközök fejlesztése által. A szoftverek egy része különböző genetikai mérések, mint például génasszociációs vizsgálatok kiértékelését és elemzését tették lehetővé. A projekt keretén belül a bayesi relevancia analízis és kiterjesztései felhasználására került sor.

Köszönetnyilvánítás

Legelőször szeretném kifejezni köszönetem Dr. Antal Péternek az elmúlt évek alatt nyújtott útmutatásért, bátorításért és kutatásom támogatásáért. Köszönettel tartozom Dr. Strausz Györgynek a támogatásért, illetve hogy hozzájárult egy stabil környezet kialakításához, amely lehetővé tette a kutatómunkámat. Szeretném megköszönni Dr. Bagdy György professzor úrnak, az MTA-SE Neuropszichofarmakológiai és Neurokémiai Kutatócsoport vezetőjének a támogatását, és hogy hozzájárulását adta kutatócsoportja értékes genetikai adathalmazai felhasználásához az új módszerek tesztelésére. További köszönettel tartozom Dr. Juhász Gabriellának munkám támogatásáért. Szintén köszönet illeti prof. Dr. Buzás Editet, Dr. Pál Zsuzsannát, prof. Dr. Szalai Csabát és számos kollégájukat a korábbi közös kutatási együttműködésekért. Köszönöm a COMBINE munkacsoportbeli kollégáimnak a kutatást stimuláló beszélgetéseket és a jó munkahelyi légkört. Végül pedig szeretnék köszönetet mondani családomnak és barátaimnak a végtelen türelmükért, és a disszertáció írása alatt nyújtott támogatásukért.

Ezen kutatómunkát támogatta a Nemzeti Agykutatási Program a KTIA_13_NAP-A-II/14. nyilvántartási számú (ny.sz.) szerződés (KTIA_NAP_13-1-2013-0001) keretében, továbbá a Magyar Tudományos Akadémia és a Nemzeti Agykutatási Program MTA-SE-NAP B Genetikai Agyi Képalkotó Migrén Kutató Csoportja (KTIA_NAP_13-2-2015-0001), és a MTA-SE Neuropszichofarmakológiai és Neurokémiai Kutatócsoportja (Magyar Tudományos Akadémia, Semmelweis Egyetem).

7. Publikációs lista

Lektorált (peer-reviewed) publikációk száma:	30
Független hivatkozások száma:	38

7.1. A tézisekhez kapcsolódó publikációk

	1	2	3	4	5	6	7	8	9	10	11	12
Thesis 1:	•	•		•		•	•				•	
Thesis 2:					•					•		•
Thesis 3:		•	•				•	•	•			

Folyóiratcikkek

[1] G. Hullam, G. Juhasz, G. Bagdy, and P. Antal. "Beyond Structural Equation Modeling: model properties and effect size from a Bayesian viewpoint. An example of complex phenotype - genotype associations in depression". In: *Neuropsychopharmacologia Hungarica* 14.4 (2012), pp. 273–284. DOI: 10.5706/nph201212009

[2] G. Hullam and P. Antal. "The effect of parameter priors on Bayesian relevance and effect size measures". In: *Periodica Polytechnica - Electrical Engineering* 57.2 (2013), pp. 35–48. DOI: 10.3311/PPee.2088

[3] I. Ungvari, G. Hullam, P. Antal, P. Kiszel, A. Gezsi, E. Hadadi, V. Virag, G. Hajos, A. Millinghoffer, A. Nagy, A. Kiss, A. Semsei, G. Temesi, B. Melegh, P. Kisfali, M. Szell, A. Bikov, G. Galffy, L. Tamasi, A. Falus, and C. Szalai. "Evaluation of a Partial Genome Screening of Two Asthma Susceptibility Regions Using Bayesian Network Based Bayesian Multilevel Analysis of Relevance". In: *PLOS ONE* 7.2 (2012), 1–14, e33573. DOI: 10.1371/journal.pone.0033573

Ez a folyóiratcikk egy asztmához kapcsolódó génasszociációs vizsgálat eredményeit mutatja be, melyben a bayesi relevancia analízis központi szerepet kapott és a módszer egyik első ilyen célú alkalmazása volt. A kutatás az SE-DGCI és a BME-MIT kutatócsoportjainak együttműködéseként jött létre. A cikk osztott első szerzőjeként (co-first author) a bayesi relevancia analízissel és az eredmények statisztikai értelmezésével járultam hozzá cikkhez. Továbbá részt vettem a bayesi metodológia leírásában, diszkussziójában.

[4] O. Lautner-Csorba, A. Gezsi, D. Erdelyi, G. Hullam, P. Antal, A. Semsei, N. Kutszegi, G. Kovacs, A. Falus, and C. Szalai. "Roles of Genetic Polymorphisms in the Folate Pathway in Childhood Acute Lymphoblastic Leukemia Evaluated by Bayesian Relevance and Effect Size Analysis". In: *PLOS ONE* 8.8 (2013), 1–13, e69843. DOI: 10.1371/journal.pone.0069843

Ebben a SE-DGCI-vel közös kutatásban alkalmaztam a struktúra függő bayesi hatáserősség (structure conditional Bayesian odds ratio) mértéket a leukémiához kapcsolódó releváns faktorok hatáserősségének részletes jellemzésére.

[5] G. Juhasz, G. Hullam, N. Eszlari, X. Gonda, P. Antal, I. Anderson, T. Hokfelt, J. Deakin, and G. Bagdy. "Brain galanin system genes interact with life stresses in depression-related phenotypes". In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111.16 (2014), E1666–73. DOI: 10.1073/pnas.1403649111

Ez a publikáció jelentős eredményeket mutat be a Galanin génrendszer depresszióban betöltött szerepéről. Az én hozzájárulásom a bayesi relevancia analízis és további többváltozós statisztikai elemzések végrehajtása, illetve ezek eredményeinek bemutatása volt.

[6] G. Juhasz, X. Gonda, G. Hullam, N. Eszlari, D. Kovacs, J. Lazary, D. Pap, P. Petschner, R. Elliott, J. Deakin, I. Anderson, P. Antal, K. Lesch, and G. Bagdy. "Variability in the Effect of 5-HTTLPR on Depression in a Large European Population: The Role of Age, Symptom Profile, Type and Intensity of Life Stressors". In: *PLOS ONE* 10.3 (2015), e0116316 15p. DOI: 10.1371/journal.pone.0116316

Ebben a cikkben az 5-HTTLPR (szerotonin transzporter) genetikai variánsainak a szerepét vizsgáltuk különböző depresszió fenotípusok szempontjából. Az 5-HTTLPR relevanciájának és a környezeti tényezőkkel való interakciójának vizsgálatára alkalmaztam a bayesi relevancia analízist, illetve a Markov-takaró alapú bayesi hatáserősség mértéket (MBG-based Bayesian odds ratios).

Könyvfejezet szerkesztett kötetben

[7] G. Hullam, A. Gezsi, A. Millinghoffer, P. Sarkozy, B. Bolgar, S. Srivastava, Z. Pal, E. Buzas, and P. Antal. "Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis". In: *Arthritis Research: Methods and Protocols*. Ed. by S. Shiozawa. Methods in Molecular Biology, vol. 1142. New York: Springer, 2014, pp. 143–176. DOI: 10.1007/978-1-4939-0404-4_14

Ez a könyvfejezet a bayesi relevancia analízis és egy ahhoz kapcsolódó génprioritizáló módszer alkalmazását mutatja be génasszociációs vizsgálatoknál. Az általam készített bayesi relevancia analízis alkalmazási útmutatója, beleértve a szükséges előfeldolgozási lépések ismertetését, továbbá a javasolt beállítások és a lehetséges elemzési, utófeldolgozási opciók leírása alkotta a publikáció lényegi részét.

[8] P. Antal, A. Millinghoffer, G. Hullam, G. Hajos, P. Sarkozy, C. Szalai, and A. Falus. "Bayesian, Systems-based, Multilevel Analysis of Biomarkers of Complex Phenotypes: From Interpretation to Decisions". In: *Probabilistic Graphical Models for Genetics, Genomics and Postgenomics*. Ed. by C. Sinoquet and R. Mourad. New York: Oxford University Press, 2014, pp. 318–360

A könyvfejezet fő célja a bayesi relevancia analízis metodológiájának és alapvető fogalmainak bemutatása volt. Az én hozzájárulásom az erős relevanciát (strong relevance) reprezentáló strukturális jegyek, illetve kapcsolódó strukturális relevancia típusok felhasználása és az ezen alapuló eredmények értelmezésének bemutatása volt.

[9] P. Antal, A. Millinghoffer, G. Hullam, G.Hajos, C. Szalai, and A. Falus. "A bioinformatic platform for a Bayesian, multiphased, multilevel analysis in immunogenomics". In: *Bioinformatics for Immunomics*. Ed. by M. Davies, S. Ranganathan, and D. Flower. Springer, 2010, pp. 157–185. DOI: 10.1007/978-1-4419-0540-6_11

A bayesi relevancia analízis alkalmazását és az eredmények értelmezését leíró szekció volt az én hozzájárulásom a könyvfejezethez.

Nemzetközi konferencia kiadványában megjelent cikk

[10] G. Hullam and P. Antal. "Towards a Bayesian Decision Theoretic Analysis of Contextual Effect Modifiers". In: *Proceedings of the 7th European Workshop on Probabilistic Graphical*

Models. Ed. by L. van der Gaag and A. Feelders. LNAI. Utrecht, The Netherlands: Springer, 2014, pp. 222–237

[11] G. Hullam and P. Antal. "Estimation of effect size posterior using model averaging over Bayesian network structures and parameters". In: *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*. Ed. by A. Cano, M. Gomez-Olmedo, and T. Nielsen. Granada, Spain: DECSAI, University of Granada, 2012, pp. 147–154

[12] P. Antal, A. Millinghoffer, G. Hullam, C. Szalai, and A. Falus. "A Bayesian View of Challenges in Feature Selection: Feature Aggregation, Multiple Targets, Redundancy and Interaction". In: *Workshop on New challenges for feature selection in data mining and knowledge discovery (FSDM 2008) at The 19th European Conference on Machine Learning (ECML 2008), Journal of Machine Learning Research - Workshop and Conference Proceedings: FSDM 2008*. Ed. by Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. V. de Peer. Vol. 4. 2008, pp. 74–89

Ebben a cikkben a bayesi relevancia analízishez kapcsolódó új fogalmak, úgymint releváns részhalmazok és interakciós score-ok bemutatására kerül sor. A demonstrációs célokat szolgáló asztma adathalmaz elemzésével és az eredmények diszkutálásával járultam hozzá a munkához.

7.2. További publikációk

Folyóiratcikk

[13] A. Millinghoffer, G. Hullam, and P. Antal. "Statisztikai adat- és szövegelemzés Bayes-hálókkal: a valószínűségektol a függetlenségi és oksági viszonyokig". In: *Híradástechnika* 60.10 (2005), pp. 40–49

[14] Z. Pal, P. Antal, A. Millinghoffer, G. Hullam, K. Paloczi, S. Toth, H. Gabius, M. Molnar, A. Falus, and E. Buzas. "A novel galectin-1 and interleukin 2 receptor haplotype is associated with autoimmune myasthenia gravis". In: *Journal of Neuroimmunology* 229.1-2 (2010), pp. 107–111. DOI: 10.1016/j.jneuroim.2010.07.015

[15] S. Srivastava, P. Antal, J. Gal, G. Hullam, A. F. Semsei, G. Nagy, A. Falus, and E. I. Buzas. "Lack of evidence for association of two functional SNPs of CHI3L1 gene(HC-gp39) with rheumatoid arthritis." In: *Rheumatology International* 31.8 (2010), pp. 1003–1007. DOI: 10.1007/s00296-010-1396-3

[16] Z. Pal, P. Antal, S. Srivastava, G. Hullam, A. F. Semsei, J. Gal, M. Svebis, G. Soos, C. Szalai, S. Andre, E. Gordeeva, G. Nagy, H. Kaltner, N. Bovin, M. Molnar, A. Falus, H. Gabius, and E. I. Buzas. "Non-synonymous single nucleotide polymorphisms in genes for immunoregulatory galectins: association of galectin-8 (F19Y) occurrence with autoimmune diseases in a Caucasian population". In: *Biochimica et Biophysica Acta-general Subjects* 1820.10 (2012), pp. 1512–1518. DOI: 10.1016/j.bbagen.2012.05.015

[17] G. Jobbagy-Ovari, C. Paska, P. Stiedl, B. Trimmel, D. Hontvari, B. Soos, P. Hermann, Z. Toth, B. Kerekes-Mathe, D. Nagy, I. Szanto, A. Nagy, M. Martonosi, K. Nagy, E. Hadadi, C. Szalai, G. Hullam, G. Temesi, P. Antal, G. Varga, and I. Tarjan. "Complex analysis of multiple single nucleotide polymorphisms as putative risk factors of tooth agenesis in the Hungarian population". In: *Acta Odontologica Scandinavica* 72.3 (2013), pp. 216–227. DOI: 10.3109/00016357.2013.822547

[18] G. Temesi, V. Virág, E. Hadadi, I. Ungvari, L. Fodor, A. Bikov, A. Nagy, G. Galffy, L. T. L, I. Horvath, A. Kiss, G. Hullam, A. Gezsi, P. Sarkozy, P. Antal, E. Buzás, and C. Szalai. "Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations". In: *Allergy Asthma and Immunology Research* 6.6 (2014), pp. 496–503. DOI: 10.4168/aair.2014.6.6.496.

[19] A. Gezsi, O. Lautner-Csorba, D. Erdelyi, G. Hullam, P. Antal, A. Semsei, N. Kutszegi, M. Hegyi, K. Csordas, G. Kovacs, and C. Szalai. "In interaction with gender a common CYP3A4 polymorphism may influence the survival rate of chemotherapy for childhood acute lymphoblastic leukemia". In: *Pharmacogenomics Journal* 15.3 (2015), pp. 241–247. DOI: 10.1038/tpj.2014.60

Könyvfejezet

[20] G. Hullam. "Tudásmérnökség, biasok és heurisztikák becsléseknél és döntéseknél". In: *Valószínűségi döntéstámogató rendszerek*. Ed. by A. Antos, P. Antal, G. Hullam, A. Millinghoffer, and G. Hajos. Budapest: Typotex, 2014, pp. 65–85

[21] G. Hullam. "Orvosi döntéstámogatás". In: *Valószínűségi döntéstámogató rendszerek*. Ed. by A. Antos, P. Antal, G. Hullam, A. Millinghoffer, and G. Hajos. Budapest: Typotex, 2014, pp. 128–162

[22] G. Hullam. "Hiányos Adatok". In: *Intelligens adatelemzés*. Ed. by P. Antal, A. Antos, G. Horvath, G. Hullam, I. K. Imre, P. Marx, A. Millinghoffer, A. Pataricza, and A. Salanki. Budapest: Typotex, 2014, pp. 48–57

[23] G. Hullam. "Genetikai asszociációs vizsgálatok standard elemzése". In: *Bioinformatika: molekuláris méréstechnikától az orvosi döntéstámogatásig*. Ed. by P. Antal, G. Hullam, A. Millinghoffer, G. Hajos, P. Marx, A. Arany, B. Bolgar, A. Gezsi, P. Sarkozy, and L. Poppe. Budapest: Typotex, 2014, pp. 90–106

[24] G. Hullam. "Standard analysis of genetic association studies". In: *Bioinformatics*. Ed. by P. Antal, G. Hullam, A. Millinghoffer, G. Hajos, P. Marx, A. Arany, B. Bolgar, A. Gezsi, P. Sarkozy, and L. Poppe. Budapest: Typotex, 2014, pp. 84–100

Nemzetközi konferencia kiadványában megjelent cikk

[25] G. Hullam, P. Antal, C. Szalai, and A. Falus. "Evaluation of a Bayesian model-based approach in G[enetic]A[ssociation] studies". In: *Machine Learning in Systems Biology (MLSB2009), Journal of Machine Learning Research - Workshop and Conference Proceedings: MLSB 2009*. Ed. by S. Dzeroski, P. Geurts, and J. Rousu. Vol. 8. 2010, pp. 30–43

[26] A. Millinghoffer, G. Hullam, and P. Antal. "On inferring the most probable sentences in Bayesian logic". In: *Workshop notes on Intelligent Data Analysis in Biomedicine and Pharmacology (IDAMAP-2007), Artificial Intelligence in Medicine (AIME2007)*. Ed. by C. Combi and A. Tucker. Amsterdam, The Netherlands, 2007, pp. 13–18

[27] P. Antal, G. Hullam, A. Gezsi, and A. Millinghoffer. "Learning complex bayesian network features for classification". In: *Proc. of Third European Workshop on Probabilistic Graphical Models (PGM06)*. Ed. by M. Studeny and J. Vomlel. Prague, Czech Republic, 2006, pp. 9–16

Helyi szervezésű konferencia kiadványában megjelent cikk

[28] G. Hullam. “Discovery of causal relationships in presence of hidden variables”. In: *13th PhD Mini-Symposium*. Ed. by B. Pataki. Budapest University of Technology and Economics. Budapest, Hungary: Department of Measurement and Information Systems, Feb. 2006, pp. 50–51

[29] G. Hullam. “A first-order Bayesian logic for heterogeneous data sets”. In: *14th PhD Mini-Symposium*. Ed. by B. Pataki. Budapest University of Technology and Economics. Budapest, Hungary: Department of Measurement and Information Systems, Feb. 2007, pp. 42–45

[30] G. Hullam. “Bayesian analysis of relevance in presence of missing and erroneous data”. In: *15th PhD Mini-Symposium*. Ed. by B. Pataki. Budapest University of Technology and Economics. Budapest, Hungary: Department of Measurement and Information Systems, Feb. 2008, pp. 30–33

Hivatkozások

[AC08] N. Angelopoulos and J. Cussens. "Bayesian learning of Bayesian networks with informative priors". In: *Annals of Mathematics and Artificial Intelligence* 54.1-3 (2008), pp. 53–98.

[Agr02] A. Agresti. *Categorical data analysis*. Wiley & Sons, 2002.

[Ant07] P. Antal. *Integrative Analysis of Data, Literature, and Expert Knowledge*. Ph.D. dissertation, K.U.Leuven, D/2007/7515/99, 2007.

[ATS03] C. Aliferis, I. Tsamardinos, and A. Statnikov. "Causal Explorer: A Probabilistic Network Learning Toolkit for Biomedical Discovery". In: *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03)*. 2003, pp. 371–376.

[Bal07] D. J. Balding. *Handbook of Statistical Genetics*. Wiley & Sons, 2007.

[Bar12] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[Bel+96] R. Bellazzi, C. Larizza, A. Riva, A. Mira, S. Fiocchi, and M. Stefanelli. "Distributed intelligent data analysis in diabetic patient management". In: *Proc. AMIA Annual Fall Symposium 1996*. 1996, pp. 194–198.

[Ber95] J. M. Bernardo. *Bayesian Theory*. Wiley & Sons, 1995.

[BH95] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *J. R. Stat. Soc.* 57.1 (1995), pp. 289–300.

[BL97] A. Blum and P. Langley. "Selection of Relevant Features and Examples in Machine Learning". In: *Artificial Intelligence* 97.1-2 (1997), pp. 245–271. DOI: 10.1016/S0004-3702(97)00063-5.

[Bor98] A. Borovkov. *Mathematical Statistics*. Gordon and Breach, 1998.

[Bou+96] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. "Context-Specific Independence in Bayesian Networks". In: *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*. 1996, pp. 115–123.

[Bun91] W. L. Buntine. "Theory Refinement of Bayesian Networks". In: *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*. 1991, pp. 52–60.

[BW00] D. Bell and H. Wang. "A formalism for relevance and its application in feature subset selection". In: *Machine learning* 41.2 (2000), pp. 175–195.

[BZ08] R. Bellazzi and B. Zupan. "Predictive data mining in clinical medicine: Current issues and guidelines". In: *International Journal of Medical Informatics* 77.2 (2008), pp. 81–97.

[CH92] G. F. Cooper and E. Herskovits. "A Bayesian Method for the Induction of Probabilistic Networks from Data". In: *Machine Learning* 9 (1992), pp. 309–347.

[CHM04] D. M. Chickering, D. Heckerman, and C. Meek. "Large-sample learning of Bayesian networks is NP-hard". In: *The Journal of Machine Learning Research* 5 (2004), pp. 1287–1330.

[Com+00] D. Comings, R. Gade-Andavolu, N. Gonzaleza, S. Wu, D. Muhleman, H. Blake, F. Chiu, E. Wang, K. Farwell, S. Darakjy, R. Baker, G. Dietz, G. Saucier, and J. MacMurray. "Multivariate analysis of associations of 42 genes in ADHD, ODD and conduct disorder". In: *Clin Genet* 58.1 (2000), pp. 31–40.

[Coo90] G. F. Cooper. "The computational complexity of probabilistic inference using Bayesian belief network". In: *Artificial Intelligence* 42 (1990), pp. 393–405.

[CY08] Y. Chen and Y. Yao. "A multiview approach for intelligent data analysis based on data operators". In: *Information Sciences* 178.1 (2008), pp. 1–20.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[Dun61] O. Dunn. "Multiple Comparisons Among Means." In: *Journal of the American Statistical Association* 56.293 (1961), pp. 52–64.

[EN14] H. Ehrenreich and K.-A. Nave. "Phenotype-based genetic association studies (PGAS)towards understanding the contribution of common genetic variants to schizophrenia subphenotypes". In: *Genes* 5.1 (2014), pp. 97–105.

[Esb+02] K. H. Esbensen, D. Guyot, F. Westad, and L. P. Houmoller. *Multivariate data analysis-in practice: an introduction to multivariate data analysis and experimental design*. 2002.

[FGW99] N. Friedman, M. Goldszmidt, and A. Wyner. "On the Application of the Bootstrap for Computing Confidence Measures on Features of Induced Bayesian networks". In: *AI&STAT VII*. 1999.

[FK03] N. Friedman and D. Koller. "Being Bayesian about Network Structure". In: *Machine Learning* 50 (2003), pp. 95–125.

[FY96] N. Friedman and Z. Yakhini. "On the Sample Complexity of Learning Bayesian Networks". In: *Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*. 1996, pp. 274–282.

[GE03] I. Guyon and A. Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[Gel+95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 1995.

[GH95] D. Geiger and D. Heckerman. "A Characterization of the Dirichlet Distribution with Application to Learning Bayesian Networks". In: *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*. 1995, pp. 196–207.

[GHY12] A. Gelman, J. Hill, and M. Yajima. "Why We (Usually) Dont Have to Worry About Multiple Comparisons". In: *Journal of Research on Educational Effectiveness* 5 (2012), pp. 189–211.

[GSM08] X. Gao, J. Starmer, and E. Martin. "A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms". In: *Genetic Epidemiology* 32.4 (2008), pp. 361–369. DOI: 10.1002/gepi.20310.

[Hec99] D. Heckerman. "Learning in graphical models". In: MIT Press, 1999. Chap. A Tutorial on Learning with Bayesian Networks.

[HGC95] D. Heckerman, D. Geiger, and D. Chickering. "Learning Bayesian networks: The Combination of Knowledge and Statistical Data". In: *Machine Learning* 20 (1995), pp. 197–243.

[Hir+02] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn. "A comprehensive review of genetic association studies". In: *Genetics in Medicine* 4.2 (2002), pp. 45–61.

[HMC97] D. Heckermann, C. Meek, and G. Cooper. *A Bayesian Aproach to Causal Discovery*. Technical Report, MSR-TR-97-05. 1997.

[Hoe+99] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. “Bayesian Model Averaging: A Tutorial”. In: *Statistical Science* 14.4 (1999), pp. 382–417.

[HS97] M. Hall and L. Smith. “Feature Subset Selection: A Correlation Based Filter Approach”. In: *International Conference on Neural Information Processing and Intelligent Information Systems (1997)*. 1997, pp. 855–858.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data minig inference and prediction*. Springer-Verlag, 2001.

[Inz+00] I. Inza, P. Larrañaga, R. Etxeberria, and B. Sierra. “Feature subset selection by Bayesian network-based optimization”. In: *Artificial intelligence* 123.1 (2000), pp. 157–184.

[KJ97] R. Kohavi and G. H. John. “Wrappers for feature subset selection”. In: *Artificial Intelligence* 97 (1997), pp. 273–324.

[Kon+98] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. *A comparison of non-informative priors for Bayesian networks*. 1998.

[Kon01] I. Kononenko. “Machine learning for medical diagnosis: history, state of the art and perspective”. In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89–109.

[KS96] D. Koller and M. Sahami. *Toward optimal feature selection*. Technical Report SIDL-WP-1996-0032. 1996.

[Lew02] C. Lewis. “Genetic association studies: design, analysis and interpretation”. In: *Briefings in bioinformatics* 3.2 (2002), pp. 146–153.

[LKZ00] N. Lavrac, E. Keravnou, and B. Zupan. *Intelligent Data Analysis in Medicine*. Technical Report. 2000.

[LT06] P. Lisboa and A. Taktak. “The use of artificial neural networks in decision support in cancer: A systematic review”. In: *Neural Networks* 19.4 (2006), pp. 408–415.

[Lun+00] D. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. “WinBUGS A Bayesian modelling framework: Concepts, structure, and extensibility”. In: *Statistics and Computing* 10.4 (2000), pp. 325–337.

[LY05] H. Liu and L. Yu. “Toward integrating feature selection algorithms for classification and clustering”. In: *Knowledge and Data Engineering, IEEE Transactions on* 17.4 (2005), pp. 491–502.

[Mad+96] D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. “Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs”. In: *Comm.Statist. Theory Methods* 25 (1996), pp. 2493–2520.

[Man+09] T. Manolio et al. “Finding the missing heritability of complex diseases”. In: *Nature* 461.7265 (2009), pp. 747–753.

[Mao04] K. Mao. “Feature Subset Selection for Support Vector Machines Through Discriminative Function Pruning Analysis”. In: *IEEE TRANS ON SYSTEMS, MAN, AND CYBERNETICSPART B: CYBERNETICS* 34.1 (2004), pp. 60–67.

[Mik+95] S. Miksch, W. Horn, C. Popow, and F. Paky. “Therapy Planning Using Qualitative Trend Descriptions”. In: *Proceedings of 5th Conference on Artificial Intelligence in Medicine Europe (AIME 95), June 25 - 28, 1995, Pavia, Italy*. 1995, pp. 197–208.

[Mit07] A. Mittal. *Bayesian Network Technologies: Applications and Graphical Models: Applications and Graphical Models*. IGI Global, 2007.

[MSL12] R. Mourad, C. Sinoquet, and P. Leray. "Probabilistic graphical models for genetic association studies". In: *Briefings in bioinformatics* 13.1 (2012), pp. 20–33.

[NC06] A. Neath and J. Cavanaugh. "A Bayesian approach to the multiple comparisons problem". In: *Journal of Data Science* 4.2 (2006), pp. 131–146.

[Nea04] R. Neapolitan. *Learning bayesian networks*. Vol. 38. Prentice Hall Upper Saddle River, 2004.

[Nob09] W. Noble. "How does multiple testing correction work?" In: *Nature biotechnology* 27.12 (2009), pp. 1135–1137.

[OLD01] A. Onisko, P. Lucas, and M. Druzdzel. "Comparison of Rule-Based and Bayesian Network Approaches in Medical Diagnostic Systems". In: *S. Quaglini, P. Barahona, and S. Andreassen (Eds.): AIME 2001, Lecture Notes in Artificial Intelligence 2101*. 2001, pp. 283–292.

[PCB06] K. Preacher, P. Curran, and D. Bauer. "Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis". In: *Journal of Educational and Behavioral Statistics* 31.4 (2006), pp. 437–448.

[PCB13] R. Patnala, J. Clements, and J. Batra. "Candidate gene association studies: a comprehensive guide to useful in silico tools". In: *BMC genetics* 14.1 (2013), p. 39.

[Pea00] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[Pen+07] J. Pena, R. Nilsson, J. Bjrkegren, and J. Tegnér. "Towards Scalable and Data Efficient Learning of Markov Boundaries". In: *International Journal of Approximate Reasoning* 45 (2007), pp. 211–232.

[PS12] B. Pei and D. Shin. "Reconstruction of biological networks by incorporating prior knowledge into Bayesian network models". In: *Journal of Computational Biology* 19.12 (2012), pp. 1324–1334.

[RN10] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[SB09] M. Stephens and D. Balding. "Bayesian statistical methods for genetic association studies". In: *Nature Review Genetics* 10(10) (2009), pp. 681–690.

[SGS01] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2001.

[SP78] P. Szolovits and S. Pauker. "Categorical and Probabilistic Reasoning in Medical Diagnosis". In: *Artificial Intelligence* 11.1–2 (1978), pp. 115–144.

[Ste09] D. A. Stephens. "Complexity in Systems Level Biology and Genetics: Statistical Perspectives". In: *Encyclopedia of Complexity and Systems Science*. 2009, pp. 1226–1244.

[Sto+04] J. Stoehlmacher, D. Park, W. Zhang, D. Yang, S. Groshen, S. Zahedy, and H.-J. Lenz. "A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer". In: *British Journal of Cancer* 91.2 (2004), pp. 344–354. DOI: 10.1038/sj.bjc.6601975.

[Sto02] J. Storey. "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[TA03] I. Tsamardinos and C. Aliferis. “Towards Principled Feature Selection: Relevancy, Filters, and Wrappers”. In: *Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics*. 2003, pp. 334–342.

[UC14] G. Upton and I. Cook. *A Dictionary of Statistics 2 rev. ed.* Oxford university press, 2014.

[WB10] S. Wong and C. Butz. “A Comparative Study of Noncontextual and Contextual Dependencies”. In: *Foundations of Intelligent Systems*. Vol. 1932. Lecture Notes in Computer Science. 2010, pp. 247–255. DOI: 10.1007/3-540-39963-1_26.

[Wei+78] S. Weiss, C. Kulikowski, S. Amarel, and A. Safir. “A Model-Based Method for Computer-Aided Medical Decision-Making”. In: *Artificial Intelligence* 11.1–2 (1978), pp. 145–172.