A multikollinearitás téves kezelése, avagy információvesztést generáló változókizárás a hasonlóságelemzésben

(Wrong interpretation of the multicollinearity or even the exclusion of the affected variables leads to lack of information)

Pitlik László, Pitlik Marcell (MYX-team)

Kivonat: Az egymással „lineáris” (együttmozgó) kapcsolatot mutató változók modellezésben betöltött szerepe teljesen mintázatmentes, vagyis ezek kizárása tilos, felesleges. A kizárás lenne maga a modellezési anomália – szemben a kánonnal. A kijelentések alapján az adja, hogy bizonyítéknak már az is elegendő, ha bármely nem várt együttállásból már csak egyetlen egy eset is előfordul.

Kulcsszavak: kánon, KNUTH, hermeneutika

Abstract: The input variables having a strong parallelism between them may not be excluded from the modelling processes with subjective human decisions - although the canon tries to enforce this solution. These specific variables produce no pattern concerning the modelling success. The exclusion of these variable will only lead to lack of information. The evidence is trivial: if only one occurrence of a not expected situation can be detected, then we may not speak about a rule/recommendation.

Keywords: canon, KNUTH, hermeneutics

# Előzmények

Forrás: <https://hu.wikipedia.org/wiki/T%C3%B6bbsz%C3%B6r%C3%B6s_line%C3%A1ris_regresszi%C3%B3> – vö. „Multikollinearitás kizárása: két független változó közti korrelációs együttható nem haladhatja meg a 0,7-es, míg a determinációs együttható: 0,5-ös értéket. Ha ilyen mégis előfordul, ki kell hagyni a modellépítésből, mert torzít**hat**ja az eredményeket. Ha mégis benne hagyjuk a modellben, **nem leszünk képesek tisztán elkülöníteni a magyarázó változók egyenkénti hatását**. Ilyen természetű változónak tekinthető adott modellbe bevont testsúly, testmagasság és életkor.”

A modellezés kapcsán, amikor is pl. termelési függvényeket állítunk fel (vö. idézett példa független változói esetén táppénzes napok számának modellezése a dolgozók fiziológiai paraméterei függvényében, vagyis a testsúly, testmagasság, életkor alapján), akkor számos modellező, akit már megfertőzött a statisztikai matematika néhány alapvetése, gyanakodva vizsgálja a fenti idézet alapján input-változói és az ezek közötti korrelációs értékeket.

Fontos kiemelni, hogy a torzít**hat**ja szó nem egyértelmű semmilyen tekintetben: sem azt nem mondja meg, mit, mikor és miért torzít az a hatás, sem azt nem mondja meg, mit is jelent maga a torzítás. Vagyis ez egy tipikus szómágia, mely tetszőlegesen értelmezhető – vagyis nem felel meg a knuth-i elvnek, mely esetében tudás az, ami forráskódba átírható. Ha lenne olyan context-free tudás, mely alapján egy regressziós számításból ki kell zárni egy/több változót multikollinearitás okán, akkor annak részének kellene lenni ezen algoritmusoknak, ill. ki nem hagyva ezeket triviálisan belátható torzulások kellenek, hogy létrejöjjenek minden olyan esetben, ahol a kizárást az algoritmus felvállalta a megfelelő feltételek teljesülése miatt.

A változók hatásának egymástól való elkülönítése pedig eleve egy hermeneutikai illúzió: a világ elvileg mérhető, s különösen a pl. csak kérdőíves válaszadók szubjektivitásán alapulóan „felmérhető” jelenségei (pl. a pirézekről, mint nem létező népcsoportról alkotott vélemény 1<10-es skálán) néhány nagyon alapvető (fizikai) jelenségen túl, melyek hatásmechanizmusai kapcsán valóban vélelmezhető egy fajta kihagyás- és átfedés-mentességre törekvés a modellező által, a többi absztrakciós szinten nincs értelme ilyen idealisztikus elvárásoknak: pl. az iskolai testsúlymérés esetén zömmel az sem világos, ruhástól, vagy sem? Evés előtt vagy után? Sportolás előtt, vagy után? Vagyis lényegében csak valami nagyon homályos dolog kerül mérésre általában. S ha sikerül is a fenti mérési elveket előre rögzíteni a testtömeg kapcsán, akkor is még illett volna gondolni arra, miként változhat a gravitáció, stb. Ha a testtömeg és a testmagasság korrelációja magas adott adathalmaz esetén, az semmit nem von le annak a tetszőleges arányú esetnek az információértékéből, amikor valakik alacsonyak, de kövérek, stb. Mivel a korreláció magas értéke nem törvényszerűségre, csak gyakori együttmozgásra utal, így lemondani valamiféle benchmarktól való eltérés információértékéről, eleve egy modellezés-filozófiai nonszensz. Az természetesen igaz lesz, hogy a magas korrelációt mutató változók közül általában valamelyik képviseli a többit, így nem beszélhetünk elkülöníthető hatásokról, de nem azért, mert a modellbe került minden magas korrelációt mutató változó, hanem azért, mert ha bármelyiket is kihagyjuk, akkor a másik kerül be a modellbe, vagyis ezek hatásmechanizmusa quasi egy közös eredő hatásként jelenik meg. Más megfogalmazásban: tulajdonképpen melyik változót is kell kihagyni, ha pl. 3 változó között erős a korreláció? Miért pont A-t, B-t, vagy C-t? Mi történik a modellel, ha csak A, vagy csak B, vagy csak C marad benn? Vagy esetleg csak AB? AC? BC?

Változót tehát soha nem a modellezőnek kell kihagynia, hanem a modellezést végző módszertant (algoritmust) kell oly mértékig finomhangolni, hogy az elvégezzen minden olyan lépést, mely legitim. A modellezőnek nem felelőssége információvesztésről döntéseket hozni – neki az a dolga, hogy minél több nyers adatot állítson rendelkezésre, vagyis hogy érdemi lépéseket tegyen a bigdata felé. Mert z egyetlen egy értelmes kiindulása alap csak az lehet a modellezés kapcsán, ha minden! Rendelkezésre álló adat feldolgozásra kerül.

Ez a probléma már visszaköszön a modellek ún. klasszikus tesztelése során is, ahol nyersadatok elkülönítésre kerülnek feldolgozás helyett. Az ideális modellezés minden nyersadatot feldolgoz, s a legjobb modellt a részeredmények konzisztenciája (egymást erősítő és egymást gyengítő mechanizmusai) alapján választja ki – automatizáltan.

További előzmények: <http://miau.my-x.hu/miau/157/faktoranalizis.docx> (2011) - vö. „[1] = Sajtos László, Mitev Ariel, 2007: SPSS kutatási és adatelemzési kézikönyv, Alinea Kiadó, Budapest

Bevezetés: [1] 272. oldal faktorelemzés: „…ne fogadjuk el az első felmerülő megoldást…” Ne kerteljünk: Amikor ilyen mondattal/gondolattal találkozik a felhasználó, nem árt megrettenni! Miért? Mert azon módszertanok, melyek teljesen lemondanak arról, hogy a számítási eredmények értelmezését zárt rendszerben oldják meg, egyben utat nyitnak a vonnegutizmusoknak, vagyis az önkényes, művészi értekkel kecsegtető belemagyarázásnak [vö. 3].

A feladat: [1] 246. oldal faktorelemzés: „…a változók számának redukálása a cél…” Ha ez a cél, tegyük ezt. Tegyük ezt úgy, hogy egy-egy változó megtartása vagy elvetése context free módon, az elemzői átértelmezés esélye és veszélye nélkül történjen. Hiszen akkor, ha az elemző azt és úgy fogad el, amit akar, akkor lényegében már kár is elemző módszerhez folyamodnia. Válasszon egyszerűen intuitív alapon: főleg akkor, ha ettől nem is függ semmi…(?!)”

Mint az SPSS kézikönyvre való több, mint kritikus utalások jelzik: a knuth-i elv semmibe vétele a hermeneutikai tudás forráskódba forgatása kapcsán a tudománynak remélt munkák zömét sodorja az értéktelenség lejtőjén egyre lejjebb és lejjebb. Szerencsére a kutatói intuíció nem cska kockázatok, hanem valódi ráérzések forrása is, de itt és most a cél az elemző robot létezhetősége szempontjából vizsgálni a kutatói/elemzői gyakorlatot. S ez a vizsgálat arra mutat rá, hogy alapvető stratégia tévedés áll fenn generációk óta, mely tévedés a bigdata és a mesterséges intelligencia fogalmak megjelenésével, s főleg a knuth-i elv egyetlen prioritásként való elfogadásával a jövőben fokozatos fel fog oldódni.

# Feladat

Részletek: <https://miau.my-x.hu/miau/270/korrel_korrel.xlsx>

A megadott XLSX-ben reprodukálhatóan nyomon követhető módon bemutatásra kerül egy minimalista demo, ahol

* 10 modell kerül felépítésre
* 50 objektum (O1,…,O50)
* 14 (konstans, de véletlenszám-generátorral előállított) input-attribútum (X1,…,X14, ill. A1,…,A14)
* és 1 db 10 verzióban véletlenszámokból álló output (X15=A15=Y) alapján,

ahol minden fenti paraméter- és számérték véletlenszerűen került kiválasztásra.

A 10 OAM alapján kiszámításra kerül

* az összes változó közötti korreláció
* a mindenkori Xi-k és az Y közötti korrelációk maximuma
* minimuma
* létrehozásra kerül OAM-onként 1 lépcsős függvény alapú (hasonlóságelemzési modell)
* s kiszámításra kerül ennek becslés-tény korrelációja
* valamint a magas korrelációt mutató inputváltozók összes modell-hatásmechanizmusának és a változó csoportok maximális hatásmechanizmusainak különbsége (a modellekhez kényszerűen mindenkor felkínált összes változó hatásmechanizmusának fennmaradó mértékét reprezentálandó)
* kiszámításra kerül az input-változónkénti átlagos korrelációk és a modellek változónkénti hatásmértékeinek korrelációja is…

Csak a leggyengébb becslés-tény-korrelációjú modell és a legjobb becslés-tény korrelációjú modell esetén kiszámításra kerül egy-egy inverz és egy-egy dupla-attribútumkészletű (hasonlóság-alapú) modell (COCO STD) is.

A 10 alapmodell kapcsán felállításra kerül egy új OAM, melyben értelemszerűen

* 10 objektum és
* 4+1 attribútum szerepel
* melyek közül az Y a magas korrelációúj input-változók fennmaradó hatásmértéke
* az X1-X2-X3-X4 pedig
  + az átlagos korrelációk és a hatásmértékek közötti korreláció
  + a maximális input-output korreláció
  + a minimális input-output korreláció
  + a modell tény-becslés korrelációja…

A legjobb és legrosszabb modellek direkt, inverz és dupla-attribútumkészletű paraméterei alapján levezetésre kerül

* a magas korrelációjú inputok ceteris paribus viselkedése
* a tény-becslés korrelációk alakulása…

# Célok

A fenti számítási részfeladat lényege az, hogy

* demonstrálja már szimpla statisztikai értelmezések, ill. a puszta előfordulások tényei alapján, hogy a magas korrelációjú változócsoportok egymáshoz képesti viselkedése a véletlen outputok függvényében tetszőleges, azaz semmilyen törvényszerűség felismerését nem engedi vélelmezni - szemben az előzmények kapcsán ismertetett elvekkel,
* a dupla-attribútumkészletű modellek ceteris paribus alakzatai önkényesen alakulhatnak összevetve a változók közötti korrelációk mértékével és irányában, vagyis ismét csak semmilyen összefüggés és/vagy anomáliagyanú nem mutatható ki - szemben az előzmények kapcsán ismertetett elvekkel…
* legalább egy olyan előfordulást lehessen felmutatni véletlen beállításokat felhasználva, melyek szembe mennek a kánon által sugalmazott szabályszerűséggel/ajánlással, hiszen már egyetlen előfordulás is érv arra, hogy nincs univerzális mintázat és csak és kizárólag a mindenkori változók közötti számmisztikus kapcsolatok azok, mely vezérelni képesek a változók információtartalmának hasznosulását, amennyiben olyan feldolgozási módszerek kerülnek bevonásra a modellezésbe, melyek kellően rugalmasak, ill. eleve konzisztenciára törekvők…
* a humán döntéshozó/modellező szerepe csak a konzisztencia-alakzatok értelmezésekor érdemi, módszertani alapvetésekbe beleszólnia tilos…

# Alapadatok



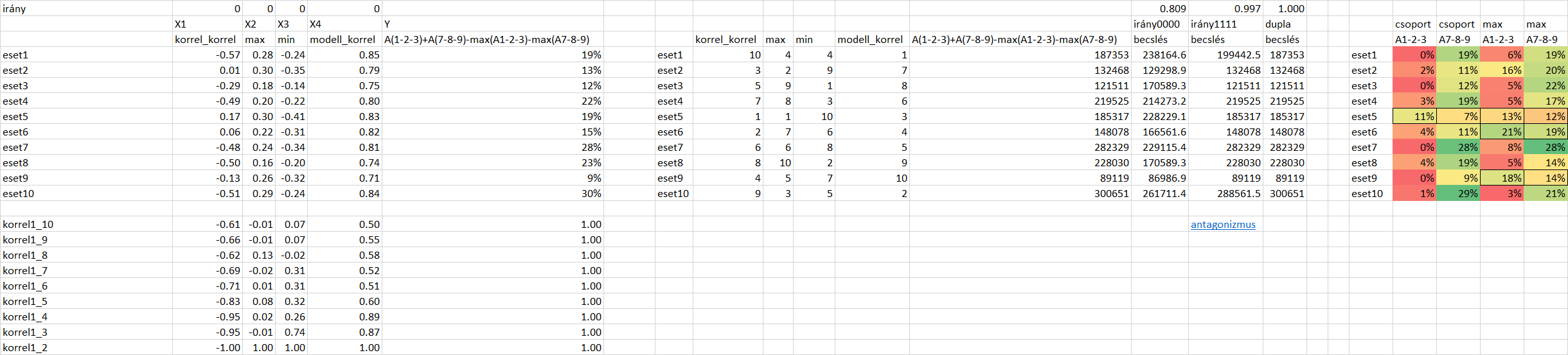
1. ábra: A nyers OAM, a sorszámozott OAM, a korrelációs mátrix az első modell esetében (forrás: saját ábrázolás)

Az 1. ábra alapján látható, hogy

* az A1-A2-A3 attribútumok mutatnak egymással magas korrelációt, ill.
* az A7-A8-A9 változók is rel. magas korrelációt mutatnak (vö. 0.7-es küszöbérték) – emellett ezen változótriplet korrelációnak negatív előjele lehet negatív is
* a teljesen véletlenszám-alapú példa kapcsán a hasonlóságelemzés által kialakított monoton lépcsősfüggvények eredőjeként értelmezett becslés és a tények (Y) közötti korreláció (0.85) magas
* lényegesen magasabb, mint bármely input-attribútum és az Y közötti korreláció abszolút értéke (0.28, ill. 0.24)
* a hatásmértékek attribútumonként 0-23% között változnak,
* s ezek összege 100%
* az attribútumok többi attribútummal (inkl. saját magukkal) mutatott átlagos korrelációja és a hatásmértékek közötti korreláció tetszőleges lehet (vö. -0.57 – s lásd OAM2 később)
* az A7-A8-A9 csoport modellbeli hatása a maximális 19% levonása után is magas (4+15%),
* az A1-A2-A3 csoport modellbeli hatása a maximális 6% levonása után 0% (vagyis itt akár még igaz is lehetne, hogy a nagyobb korrelációk ténylegesen feleslegessé tesznek változókat – de ekkor sem manuális önkénnyel kizárva, hanem a többi változóhoz mért hasznosságát context free algoritmussal vizsgálva – hiszen itt az attribútumok jelentése bármi lehet, mégis lehet rájuk vonatkozóan algoritmikus döntéseket hozni a hasonlóságelemzés mögött álló egyenletrendszer-megoldó online solver alapján) modellezői ráhatások nélkül…

A fenti ábrához hasonló további 9 eset kerül még kialakításra – véletlenszerű A15=Y értékekkel.

# OAM2

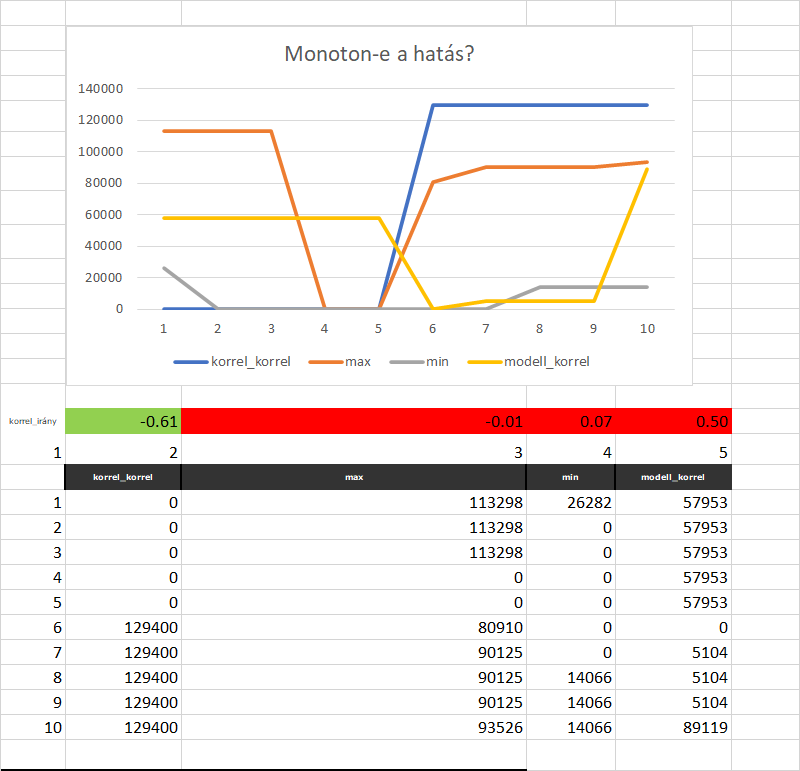


1. ábra: Az esetek kiemelt mutatói (forrás: saját ábrázolás)

A 2. ábra alapján látható, hogy

* az X1=A1 (korrel\_korrel) értékek (vagyis az átlagos korrelációk és a hatásmértékek közötti kapcsolat) korrelációja -0.57 és +0.17 között, tehát irányt váltó módon ingadozik, ami kizárja bármiféle monoton kapcsolat létét a nyers inputok közötti korrelációk és a változók modellezés szempontú hasznossága között…
* a max és min (X2, X3) oszlopok önmagukon belüli korrelációi azonos előjelűek,
* de egymáshoz képest eltérő előjelet mutatnak minden esetben…
* a hasonlóságelemzéssel nyert modellek korrelációi 0.71 és 0.85 között ingadoznak (így majd eset9 és eset1 lesznek azok, melyek kapcsán inverz és dupla-modellek is készülnek)
* a max, min értékek abszolút nagysága és egymáshoz való viszonya és a modellkorrelációk között nincs semmilyen együttmozgás, hiszen
* a korrel1\_10—korrel1\_2 táblázatrész korrelációs értékei az X1-2-3-4 és az Y között semmilyen kapcsolatot nem mutatnak, ill. az esetszámmal pl. a max-oszlopban még előjelváltás is történik (vö. -0.02 vs. +0.13)
* a hasonlóságelemzés 0000-iránnyal vagyis az X1-2-3-4 esetén is minél nagyobb, annál jobb elvet kikényszerítve tény-becslés korreláció 0.809,
* míg ennek inverze (kényszerű 1111-irány) 0.997,
* s a dupla modell esetén 1.000
* ott, ahol a korrel1\_10 sorban két negatív és két pozitív korrelációs érték szerepel, vagyis sem az 1111, sem a 0000 paramétersor nem felel meg a korrelációk 1100-s alapértelmezésének
* a Y oszlop, vagyis a maximumhatásoktól mentesített magas kiindulási input-korrelációs csoportok (A1-2-3 és A7-8-9) modell-hatásmértékei 9-30% között ingadoznak, ami olyan jelentős, hogy már ez alapján is kizárt lemondani a magas korrelációt mutató változók bármelyikéről
* a A1-2-3 tiszta hatása (jobb oldal felül -színkódokkal és rácsozással kiemelt táblázat) ugyan valóban sokszor 0%, vagyis egy változó képes reprezentálni a csoportot, de nem mindig (60%-ban nem!)
* az A7-8-9 tiszta hatása soha nem 0%, vagyis az alacsonyabb korrelációs csoport nagyobb hatással bír, de
* az eset5 alapján az sem igaz, hogy mindenkor magasabb lenne a hatásmértéke az A7-8-9 csoportnak, mint az A1-2-3-csoportnak (vö. 11%>7%)
* az A7-8-9-csoport maximum-hatásai általában nagyobbak, de itt sem igaz, hogy mindenkor nagyobbak lennének, mint az A1-2-3-csoport maximum hatásmértékei (vö. eset5-6-9)

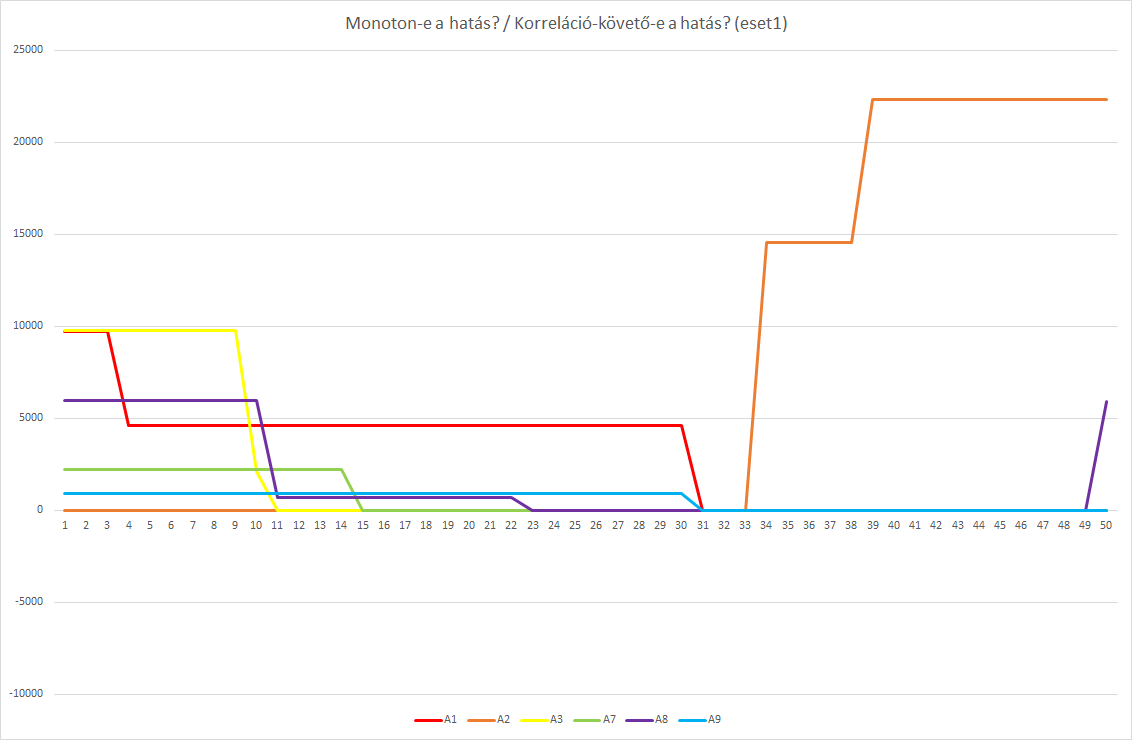
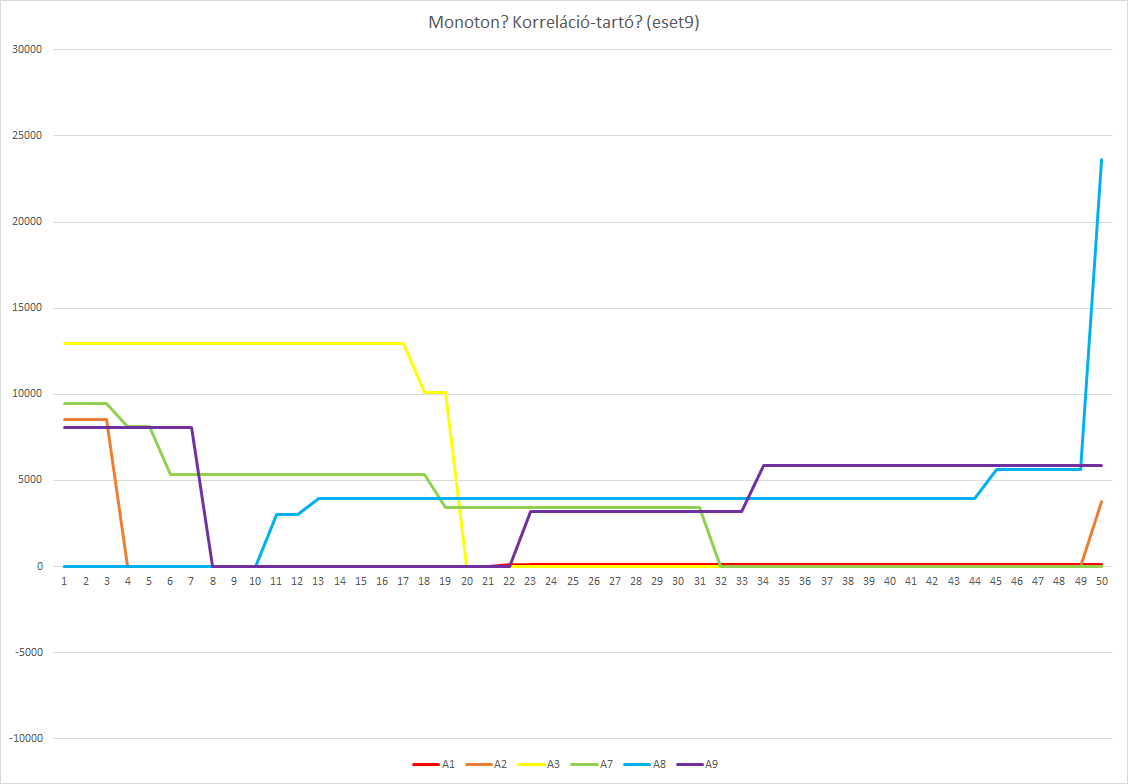
Monoton-e egy-egy input hatása az outputra?



1. ábra: Az OAM2 dupla-modelljének lépcsős-függvényei (forrás: saját ábrázolás)

A 3. ábra alapján látható, hogy:

* a korrel\_korrel változó monoton maradt és fordított arányú, mint ahogy azt a 2. ábra korrel1\_10 sorának -0.61-es értéke is előrevetítette
* a többi változó azonban nem maradt monoton (noha pl. modell\_korrel esetében a korrel1\_10 érték +0.5 volt – s itt az utolsó okkersárga pont kapcsán még a minél nagyobb (vö. első sorszám jobb, mint 10. sorszám) annál jobb elv sem érvényesült
* vagyis monoton kapcsolatról alapvetően nem lehet beszélni…

1. ábra: Nyers inputok komplex hatása az Y-ra (eset1&9) – forrás: saját ábrázolás

A 4. ábra bal oldala (eset1) alapján látható, hogy

* az A1-2-3-csoport, melyek egymással magas pozitív korrelációban álltak, így az Y-nal is azonos hatásmechanizmust várhatnánk el tőlük (ami azonban nem igaz, mert az A1vsY kapcsolat korrelációja pozitív, míg az A2&3vsY kapcsolat negatív: 0.05, -0.16, -0.15), az A2 (narancssárga) görbe esetén tipikusan ellentétes lefutást mutat, mint a citromsárga és a piros görbe a nyers OAM vagyis az 50\*(14+1) adat alapján – tehát a monotonitás adott, de az iránytartás nem, mert az A2-nek az A3-mal azonos módon és az A1-hez mérten ellentétesen kellene hatnia, de itt a kölcsönhatások között semmilyen mintázat nem fedezhető fel…
* az A7-8-9-es csoport esetén a lila görbe a monotonitást sérti meg és a balról jobbra való ereszkedés (kék és zöld görbe esetén) pedig a nyers adatoknál tapasztalt A7-8-9vsY nem minden esetben azonos korrelációkat (vö. -0.17, 0.05, 0.09), hiszen A7 (zöld) esetén a lefutásnak ellentétesnek illett volna lennie…

A 4. ábra jobb oldala (eset9) alapján látható, hogy

* míg az összes nyers korreláció pozitív (vö. A1-2-3: 0.14, 0.21, 0.26 – ill. A7-8-9: 0.04, 0.00, 0.02)
* addig az A1 (piros) görbe ezzel ellentétesen fut le, de monoton
* az A2 (narancssárga) görbe nem monoton
* az A3 (citromsárga) görbe monoton és elvárást követő, ill.
* az A7 (zöld) görbe szintén monoton és elvárát követő,
* az A8 (kék) görbe monoton, de ellentétes lefutással
* az A9 (lila) görbe is monoton és ellentétes lefutású…
* vagyis mintázatról ismét csak nem lehet beszélni!

# Konklúziók

Az egymással „lineáris” (együttmozgó) kapcsolatot mutató változók modellezésben betöltött szerepe teljesen mintázatmentes, vagyis ezek kizárása tilos, felesleges. A kizárás lenne maga a modellezési anomália – szemben a kánonnal. A kijelentések alapján az adja, hogy bizonyítéknak már az is elegendő, ha bármely nem várt együttállásból már csak egyetlen egy eset is előfordul.

# Referenciák

…szövegközben…