

Lektori vélemény

Kapelner Tamás, Madarász László, Ferenci Tamás: A független komponens analízis és empirikus vizsgálata c. kéziratáról.

A független komponens analízis (Independent Component Analysis, ICA) egy relatíve új eljáráscsalád gyűjtőneve, mely egy nagyon fontos, a gyakorlatban igen sok helyen felmerülő probléma megoldását célozza: hogyan lehet bizonyos feltételeknek eleget tevő, statisztikailag független jelforrások különböző módon (lineárisan) összekevert érzékelt jeleiből pusztán ezen érzékelt jeleknek az ismeretében visszaállítani az eredeti jelforrások jeleit. A fontosabb, általánosan ismert, független komponens analízisre vezető feladatok az ún. koktélparti probléma vagy a vak forrás szeparálás, de számos szakterületen – köztük pénzügyi, gazdasági folyamatok leírásánál, vizsgálatánál is – találunk független komponens analízis feladatént megfogalmazható problémákat. Az ICA fontosságához mérten talán kevésbé ismert, ezért a cikk fontos szerepet tölthet be az eljáráscsalád szélesebbkörű megismertetésében.

A cikk rövid bevezetés után tömören összefoglalja azokat a matematikai módszereket, melyek a független komponens analízis alapját képezik, majd az ICA-t egy másik, az ICA-nál jóval ismertebb és elterjedtebben használt eljárással, a főkomponens analízissel (Principal Component Analysis, PCA) veti össze. A két eljárást demonstratív példákon keresztül is bemutatja, rámutatva, hogy – bár mindkét eljárás komplex jelek komponensekre bontására szolgál – a kétféle eljárás célja jelentősen különbözik. A cikk végül néhány alkalmazási példát mutat be.

A cikk tehát fontos témakörrel foglalkozik, elfogadását – bizonyos átdolgozás, módosítás után – javaslom. Az átdolgozás célja a cikk pontosítása, az ICA alapját képező matematikai eljárások egyes esetekben történő precízebb megfogalmazása, illetve a PCA-val való viszony pontosabb bemutatása lehet. Minthogy az ICA alapját matematikai (statisztikai) eljárások adják, ezek bemutatása kellő precizitást igényel. E téren néhány esetben a cikk korrekcióra szorul. A cikk könnyebb követhetőségét néhány kisebb felépítésbeli módosítás is segítené.

A tételes megjegyzéseket a cikk felépítését követve adom meg. A tételes megjegyzésekben kisebb pontatlan megfogalmazások, illetve komolyabb hiányosságok említése együtt szerepelnek.

Kivonat: ebben a szakaszban megemlítsre kerül a PCA-val való összehasonlítás, azonban míg az ICA magyar megnevezése szerepel, a PCA megnevezése (főkomponens analízis) hiányzik.

1. Bevezetés:

Az első bekezdés utolsó mondata: „...*mely képes egymástól független adatok egy keverékéből visszaállítani az eredeti, független adatokat.*”. egy keverékből több jel visszaállítása nem lehetséges. A mondatba helyesen az *egy keverékéből* helyett *lineárisan független keverékeiből* írandó.

2. A független komponens analízis (ICA)

3. bekezdés „*statisztikailag teljesen független*” helyesen *statisztikailag független*

3. bekezdés: az N megfigyelt valószínűségi változó, időfüggvény, stb nem kell, hogy azonos eloszlású legyen.

2.2 szakasz: a fehérités kérdésének felvetése ezen a ponton még korai. Mivel a fehérités nem lényegi eleme a független komponens analízisnek, célszerű lenne ennek a lehetőségét és szerepét később

felvetni. Ugyanakkor itt hiányzik a függetlenség olyan definíciója, mely alapját képezi az egyes eljárásoknak.

4. oldal pontatlan fogalmazás: „*így nyert összeg inkább fog hasonlítani egy normális eloszlásra*” az *így nyert összeg eloszlása egyre inkább hasonlítani fog a normális eloszlásra*.

4. oldal pontatlan fogalmazás: „*...az csak azért lehet, mert az Y összegben kevesebb S_i konvolválódik.*” Az összegben nem konvolválnak a forrás jelek. Legfeljebb az összeg sűrűségfüggvénye a komponensek sűrűségfüggvényeinek konvolúciója.

6. oldal első mondata: nem szerencsés megfogalmazás. „*Az ICA becsléséhez használható további módszer a maximum likelihood (ML) becslés.*” Nem az ICA-t becsüljük: Szerencsésebb lenne: *A független komponens analízis feladat megfogalmazható maximum likelihood becslési feladatként is.*

A 6. és a 7. oldalon a szövegben több helyen szerepel *a valószínűségi változó sűrűsége*. Helyesebb lenne *sűrűségfüggvényről* beszélni.

8. oldal: a kumulánst és a negyedrendű kereszt-kumulánst definiálni kellene.

9. oldal: „*... a kevert jelek számánál közvetlenül kevesebb komponenst állítsunk elő.*” Mit jelent a közvetlenül kevesebb?

9. oldal utolsó mondat: pongyola megfogalmazás „*...úgy, hogy az új ábrázolás bizonyos szempontból jobb legyen, mint a régi.*”

10. oldal: „*A PCA először standardizálja az adatokat...*” A standardizálás általában nulla várhatóértékre és egységnyi varianciára (egységmátrix kovarianciamátrixra) való transzformációt jelent. Egységmátrix kovarianciamátrix mellett viszont a PCA határozatlan eredményt ad, hiszen ekkor minden sajátérték egyforma és egységnyi. Ezért itt is pontosabb fogalmazásra van szükség.

10. oldal: az oldalon szerepel információ-vesztés, optimális reprezentáció, azonban nem szerepel, hogy ezt milyen értelemben értjük. Később egy lábjegyzetben történik erre utalás, de ez sem elég precíz. Itt a lineáris transzformációk körében maradva igaz a megállapítás, és célszerű lenne nem lábjegyzetben, hanem a szövegben szerepeltetni, hiszen ez kiindulópontja a PCA-nak.

10. oldal: „*Belátható, hogy az új tengelyek irányai az eredeti adatbázis korrelációs mátrixának sajátvektoraival lesznek kapcsolatban...*” Ennél pontosabban kellene fogalmazni: az új tengelyek irányai a sajátvektorok irányaival egyeznek meg.

11. oldal: „*Tehát az NLPCA elvégzése tulajdonképpen egyenértékű az ICA probléma megoldásával...*” Ez a megállapítás a konkrét összefüggésre igaz, de az NLPCA feladat számos egyéb módon is definiálható, és általánosságban a kérdéses mondat így már nem igaz.

3. Empirikus vizsgálat:

12. oldal: „*ahol a f tetszőleges függvény*” Sem f , sem g nem lehet tetszőleges függvény. A függvénnyel kapcsolatos követelmények több, a cikk szerzői által idézett forrásban is megtalálhatók.

16. oldal: Nem szerencsés megfogalmazás: „*A cél tehát nem az adathalmaz lehető legnagyobb pontossággal történő leírása, hanem annak felderítése...*” Sem a PCA sem az ICA nem az adatok legnagyobb pontossággal történő leírására szolgálnak. A mondatból arra lehet következtetni, hogy az ICA-val szemben a PCA erre szolgál.

17. oldal: A következő monda nem világos: „*A számítások során a legfontosabb mérőszámunk az ICA előtti – eredeti – és ICA-val visszaállított komponensek közötti korreláció lesz.*” Ez a keverés előtti

eredeti vagy az ICA előtti, tehát már kevert jel? Nem egyértelmű a fogalmazás. A félreérthetőség elkerülhető lenne, ha a mérőszámok pontos matematikai összefüggései szerepelnének a cikkben.

4. Alkalmazási példák

23. oldal: „*A módszer különösen nagyfrekvenciás idősorok esetén alkalmazható jó eredménnyel...*” Nem világos mit jelent a nagyfrekvenciás idősor. Pontosítani kellene, hogy mit jelent a nagyfrekvenciás jelző és azt is hogy mi a makroidősor. Az is magyarázatra szorul, hogy miért van – ha valóban van – a kétféle jelnél a hatékonyságban különbség.

23. oldal: „*A külföldi folyóiratokban fellelt cikkek...*” Szükséges lenne ide néhány releváns hivatkozást betenni.

24. oldal: Itt több alkalmazás szerepel hivatkozott irodalommal, azonban nem látszik ezen alkalmazásoknál, hogy mi volt az a feladat, melyet független komponens analízissel lehetett megoldani. Egy ilyen, részben a témára a figyelmet felhívó cikkben nagyon hasznos lenne röviden magát a megoldandó feladatot is bemutatni és azt, hogy mért jelent ezen feladatoknál hasznos eszközt vagy megoldást az ICA.

A dolgozat nyelvi átnézését is javaslom, szükség lenne néhány helyen kisebb megfogalmazásbeli javításra.

Összességében: a cikk fontos terület bemutatására vállalkozik. Minthogy azonban itt matematikai eljárások bemutatásáról van szó, a precíz fogalmazás fontos. E tekintetben a cikk feltétlenül javítandó. További általános megjegyzés, hogy az ICA és a PCA céljai alapvetően eltérőek – annak ellenére, hogy mindkét eljárás statisztikai módszer. Ez talán hangsúlyosabban kellene, hogy a dolgozatban szerepeljen. Az alkalmazások részben pedig a néhány alkalmazási eredményre történő hivatkozás helyett vagy mellett talán hasznosabb lenne olyan gazdasági problémákat bemutatni, melyek ICA feladatra vezethetők vissza. A korrekciók, kiegészítések után a cikk elfogadását javaslom.

Budapest, 2012. nov. 21.

Dr. Horváth Gábor