A független komponens analízis és empirikus vizsgálata*

Kapelner Tamás,
okleveles villamosmérnök,
egészségügyi mérnök
E-mail: kapimail@zoho.com

Madarász László,
okleveles közgazdász,
kockázatkezelő
E-mail: lamadarasz@gmail.com

Ferenci Tamás,
a Budapesti Corvinus Egyetem
óraadó tanára
E-mail: tamas.ferenci@medstat.hu

A tanulmány a gazdasági adatok elemzésében egyre elterjedtebb módszer, a független komponens analízis (ICA) elméleti hátterét és empirikus vizsgálatát mutatja be. Az ICA képes több, egymással korreláló adatsort olyan komponensekre szétválasztani, melyek egymástól a lehető legnagyobb mértékben függetlenek, és melyek lineáris kombinációjaként felírható az eredeti adatsorok mindegyike. A módszer tehát lehetőséget nyújt az adatsorok alakulását befolyásoló rejtett komponensek elkülönítésére. A szerzők az elméleti háttér bemutatása után először néhány összehasonlító vizsgálatot végeznek az ICA és a nála gyakrabban használt főkomponens analízis (PCA) között, majd részletesebben vizsgálják az ICA tulajdonságait a rendelkezésre álló adatok száma, dimenziója és függőségi viszonyai tekintetében. Végül néhány példát mutatnak be a módszer alkalmazási lehetőségei közül.

TÁRGYSZÓ:
Független komponens analízis.
Főkomponens analízis.

* A szerzők köszönetet mondanak *Prof. Hunyadi Lászlónak* és a tanulmány bírálójának az értékes észrevételeikért. A dolgozatban előforduló esetlegeses hibákért kizárólag a szerzőket terheli felelősség.

Tegyük fel, hogy egy olyan koktélpartin vagyunk, ahol minden résztvevőnek saját mikrofonja van, mely felvesz minden beszélgetést az este folyamán. Hogyan állítanánk elő a rögzített hangfelvételekből olyanokat, melyek mindegyikén csak egy-egy résztvevő szavait hallani? A probléma megoldásának kulcsgondolata, hogy mivel a felvételeken hallható hangzavarhoz az egyes beszélők egymástól függetlenül járulnak hozzá, célunk eléréséhez egy olyan módszerre van szükség, mely képes egymástól független adatok lineárisan független keverékeiből visszaállítani az eredeti, független adatokat.

A feladat megoldását szolgáltató egyik módszer a független komponens analízis (independent component analysis – ICA). Az ICA alapvetően különböző adatok – legyenek azok valószínűségi változók, időfüggvények, de akár tetszőleges adatstruktúrák is – látens komponensekre bontására képes statisztikai módszer. A felbontás az eredményül kapott komponensek függetlenségét célozza, vagyis a módszer lényege a bemeneti adatok független komponensekre való dekompozíciója.

Célunk ennek a módszernek a részletes elméleti és gyakorlati bemutatása. Az első fejezetben a módszer elméleti hátterét, illetve a független komponensek előállítását biztosító statisztikai megközelítéseket mutatjuk be. A második fejezetben néhány empirikus vizsgálat segítségével szemléltetjük, hogy valóban független komponenseket hoz létre az ICA, továbbá összehasonlítjuk egy másik, szintén elterjedt módszerrel, a főkomponens analízissel (principal component analysis – PCA). Ezek után az ICA alkalmazhatóságának feltételeit, illetve azok nem teljesülésének hatását vizsgáljuk szimulációs eszközökkel. A tanulmány lezárásaként a módszer főbb (mérnöki, pénzügyi) alkalmazási területeit mutatjuk be néhány példán keresztül.

1. A független komponens analízis (ICA)

Ebben a fejezetben bevezetjük a független komponens analízis módszerét és tárgyaljuk a legalapvetőbb módszertani kérdéseit. Látni fogjuk, hogy milyen kihívások merülnek fel a megoldandó probléma kapcsán, illetve, hogy ezekre milyen válaszok adhatók.

1.1. A koktélparti-probléma

Az ICA alkalmazásának leggyakoribb példája a vak forrásszétválasztás (blind source separation – BSS) problémája. A megoldandó feladat több, rendelkezésre álló

időfüggvény független komponenseinek meghatározása, pusztán az időfüggvényekből nyerhető adatok alapján.

Szemléletes példa erre az említett koktélparti, ahol egyszerre több beszélgetést is hallani, és az egyes beszélőket akarjuk egymástól elkülöníteni. Ehhez mikrofonokat helyezünk el, és az azok által felvett jelek – melyek a partin zajló beszélgetésekből egyszerre többet is tartalmaznak, a távolság és a beesési szög függvényében különbözően súlyozva – dekompozíciójával különítjük el az egyes beszélgetéseket. Végeredményül olyan időfüggvényeket kapunk, melyek már nem több beszélgetés – illetve zaj – keverékét tartalmazzák, hanem csak valamilyen beszélgetést vagy zajt.

Formálisan megfogalmazva ugyanezt a problémát: keresett N darab független valószínűségi változó, $S_1, S_2, \ldots, S_N$, melyek a koktélparti-probléma beszélőit reprezentálják. (A tanulmányban függetlenség alatt mindenhol teljes (tehát nem páronkénti) függetlenséget értünk, azaz az N változóból bármely k darabot kiválasztva, az együttes sűrűségfüggvény az egyes vetületi sűrűségfüggvények szorzata kell legyen minden $2 \le k \le N$-re.) Legyen adott – az egyszerűség kedvéért – szintén N darab megfigyelt valószínűségi változó: $X_1, X_2, \ldots X_N$, amelyek a mikrofonok által felvett jeleket jelentik. Amint tehát látható, megfigyeléseink nem feltétlenül „időfüggvények", a valószínűségi változók bármilyen FAE (független azonos eloszlású, tehát egymástól független, azonos eloszlásból származó) realizációi is lehetnek. Amennyiben a megfigyelések időfüggvények (idősorok), az az előbbinek olyan speciális esete, melyre igaz, hogy a mintákon definiáltunk egy rendezést, mégpedig egyszerűen aszerint, milyen sorrendben – mely időpillanatokban – történt a mintavétel.

A zajmentes ICA-modell szerinti feltételezésünk tehát a következő:

$$X_i = \sum_{j=1}^{N} a_{ij} S_j, \quad i = 1, 2, ..., N, \qquad /1/$$

azaz a kevert jelek a független komponensek valamilyen lineáris kombinációjaként állnak elő. Mátrixos írásmódot használva:

$$\underline{X} = \mathbf{A}\underline{S}, \qquad /2/$$

ahol $\underline{X} = \left[X_1, X_2, \ldots, X_N\right]^T$, $\underline{S} = \left[S_1, S_2, \ldots, S_N\right]^T$, $\mathbf{A}$ pedig az ún. keverőmátrix, elemei az a_{ij} konstans koefficiensek, melyek ebben az esetben azt fejezik ki, hogy a mikrofonok milyen mértékben hallják az egyes beszélőket. A független komponens analízis tehát legegyszerűbb esetben (négyzetes és nemszinguláris $\mathbf{A}$ esetén, az $\mathbf{A}^{-1} =: \mathbf{B}$ jelölést használva) az

$$\underline{S} = \mathbf{A}^{-1}\underline{X} = \mathbf{B}\underline{X} \qquad /3/$$

probléma megoldását jelenti. További megfontolásokat igényel, ha a korábbi feltételezéseink nem teljesülnek, azaz ha például **A** szinguláris, különböző számú megfigyelt jelünk és rejtett komponensünk van, kapcsolatuk nemlineáris, a méréseket zaj terheli stb. Az esetek ismertetése meghaladná e tanulmány kereteit, következményeiket a szakirodalom bőségesen tárgyalja (*Hyvärinen–Oja* [2000]).

Az alapprobléma nehézsége tehát, hogy nem csak $\underline{S}$, de **A** – és így **B** – szintén ismeretlen. A gyakorlatban ezért általában nem is **A** meghatározásával oldható meg a probléma, hanem olyan S_i-k keresésével, melyek a lehető legnagyobb mértékben függetlenek egymástól.[1] A probléma kulcskérdése tehát, hogy hogyan lehet a valószínűségi változók függetlenségét ellenőrizni, illetve biztosítani.

1.2. A függetlenség eldöntésének lehetőségei

A következőkben ismertetjük azon főbb megközelítéseket és módszereket, melyek biztosítják az eredményül kapott változók függetlenségét.

1.2.1. Nemnormalitás

Alakítsuk át a /2/ egyenletet a következőképpen:

$$Y = \mathbf{b}^T \underline{X} = \mathbf{b}^T \mathbf{A} \underline{S} = \mathbf{q}^T \underline{S}, \qquad /4/$$

Ebből a felírásból kitűnik, hogy ha **b** értékét meg tudnánk választani úgy, hogy éppen $\mathbf{B} = \mathbf{A}^{-1}$ egy sorának feleljen meg, akkor egy kivételével **q** minden eleme nulla értékű lenne, azaz $\underline{S}$-ből éppen egy S_i független komponenst választanánk ki. A kérdés tehát: hogyan válasszuk meg **b**-t?

A centrális határeloszlás tétel klasszikus alakja szerint azonos eloszlású, egymástól független valószínűségi változók standardizált összege – elég általános feltételek mellett – normális eloszláshoz tart (*Rényi* [1973]), sőt a tétel Ljapunov- vagy Lindeberg-féle alakja az azonos eloszlásra vonatkozó kitételt nem követeli meg (*Billingsley* [1995]). Ez tulajdonképpen azt jelenti, hogy elég általános feltételek mellett, ha független valószínűségi változók összegéhez olyan valószínűségi változót adunk, mely az előbbiektől független, akkor az így nyert összeg eloszlása egyre inkább hasonlítani fog a normális eloszlásra. Eszerint, ha a $\mathbf{q}^T \underline{S}$ lineáris kombinációban **q** elemeit (legyenek ezek a súlyok) megváltoztatjuk, és az S_i-k függetlenek, akkor az összeg annál kevésbé fog hasonlítani a normális eloszlásra, minél inkább csak

[1] Az egyértelműség kérdésére a 1.4. pontban még visszatérünk.

egyetlen S_i határozza meg az összeg értékét – feltéve, hogy az S_i-k nem normális eloszlásúak. Ha ugyanis az összeg kevésbé hasonlít a normális eloszlásra, az csak azért lehet, mert az Y összeg sűrűségfüggvényében kevesebb konvolválódik az S_i változók sűrűségfüggvényei közül. Célunk pedig éppen az, hogy Y minél inkább hasonlítson az egyik S_i komponensre, azaz **q** változásának hatására – melyet nyilván **b** változtatásával érhetünk el – minél kevésbé legyen normális eloszlású.

A további S_i-k meghatározásához természetesen más-más **b** vektorok meghatározására van szükség. A keresést megkönnyíti, ha fehérített adatokat használunk (lásd az 1.3. pontot), hiszen ebben az esetben maguk a keresett **b** vektorok is ortonormáltak lesznek, tehát elegendő a már megtalált **b**-re merőleges altérben keresni a következő megoldást.

E gondolatmenet fontos következménye, hogy az S_i-k között legfeljebb egy lehet normális eloszlású, hiszen két normális eloszlású komponenst az összeg normális eloszláshoz való hasonlósága alapján nyilván nem tudunk megkülönböztetni.

Tehát, hogy maximalizáljuk $\mathbf{b}^T \underline{X}$ nemnormalitását, olyan mérőszámra van szükségünk, mely információt ad arról, hogy a valószínűségi változó eloszlása mennyire hasonlít a normális eloszlásra.

Csúcsosság

Ilyen mérőszám a csúcsosság (kurtosis). Egy μ várhatóértékű, σ szórású Y valószínűségi változó csúcsossága:

$$\operatorname{kurt}(Y) = \frac{\mathbb{E}(Y-\mu)^4}{\sigma^4}. \qquad /5/$$

A csúcsosság előnyös tulajdonsága, hogy normális eloszlás esetén értéke három, annál csúcsosabb eloszlások esetén nagyobb, míg ellenkező esetben kisebb.[2]

A probléma megoldása mindezek ismeretében már lehetséges valamilyen, erre a mérőszámra alapozott optimalizációs eljárást (például gradiens-módszerrel történő megoldást) vagy fixpont-algoritmust alkalmazva (*Li–Adali* [2008]).

Negentrópia

A csúcsossággal, mint a nemnormalitás mérőszámával kapcsolatban azt a gyakorlati megfigyelést kell azonban tennünk, hogy nagyon érzékeny az outlierekre, nem

[2] Annak érdekében, hogy az összefüggés szemléletesebb legyen, szokásos az ún. excess kurtosis használata, melynek értéke a fenti definícióval kurt(Y)–3. Így a normális eloszlásnál csúcsosabb eloszlások csúcsossága pozitív, míg a kevésbé csúcsosaké negatív lesz.

robusztus mérőszáma a nemnormalitásnak. Egy hasonló, ám kedvezőbb statisztikai tulajdonságokkal rendelkező mérőszám a differenciális entrópia:

$$H(Y) = -\int_{supp(f_Y)} f_Y(y) \log f_Y(y) \mathrm{d}y, \qquad /6/$$

ahol f_Y az Y valószínűségi változó sűrűségfüggvénye. Érdemesebb azonban egy olyan entrópia alapú mérőszámot használni, melyen keresztül a valószínűségi változó nemnormalitása közvetlenül jelenik meg. Ilyen mérőszám a negatív normalizált differenciális entrópia, azaz a negentrópia:[3]

$$J(Y) = H\left(Y_{norm}\right) - H(Y), \qquad /7/$$

ahol $H(Y_{norm})$ az Y-nal azonos várható értékű és szórású normális eloszlás entrópiája. Ez a jellemző mindig pozitív,[4] azaz az összes eloszlás közül a normális eloszlás negentrópiája a legkisebb (nulla). A nemnormalitás mérőszámának tehát ez is kiválóan megfelel. Számítása azonban nehézkes, mert a sűrűségfüggvény pontos ismerete kellene hozzá. Optimalizációs algoritmus megvalósításakor emiatt Y sűrűségfüggvényének valamilyen közelítésére van szükség $J(Y)$ becsléséhez (*Prasad–Saruwatari–Shikano* [2005]).

1.2.2. Maximum likelihood becslés

Az ICA megfogalmazható maximum likelihood becslési feladatként is. A likelihoodok kiszámítása ICA-modellre a lineáris transzformáció sűrűségfüggvényének meghatározásán alapul.

Továbbra is adott a /2/ egyenletnek megfelelő összefüggés, ahol **A** a keverőmátrixot jelenti. Ekkor a transzformált valószínűségi változó sűrűségfüggvénye a következő formulával írható le (*Barbakh–Wu–Fyfe* [2009]):

$$f_{\underline{X}}(\mathbf{x}) = \left|\frac{1}{\det \mathbf{A}}\right| f_{\underline{S}}\left(\mathbf{A}^{-1}\mathbf{x}\right) = \left|\det \mathbf{B}\right| f_{\underline{S}}(\mathbf{s}) = \left|\det \mathbf{B}\right| \prod_i f_i\left(s_i\right), \qquad /8/$$

[3] Megjegyezzük, hogy negentrópiának néha az entrópia szokásos információelméleti (általunk differenciális entrópiának nevezett) tartalmát hívják. Mi most nem követjük ezt a szokást, és – az angolszász irodalmakkal összhangban – az itt definiált normalizált differenciális entrópiát nevezzük negentrópiának.

[4] Adott szórású és várható értékű, valós értékű eloszlások közül mindig a normális eloszlás a legnagyobb entrópiájú azon eloszlások körében, melyek tartója az egész számegyenes (*Park–Bera* [2009]).

ahol **x** és **s** egy-egy $\underline{X}$-re, illetve $\underline{S}$-re vonatkozó megfigyelés, és f_i az *i*-edik független komponens sűrűségfüggvénye. A /8/ egyenlőség kifejezhető $\mathbf{B} = [\mathbf{b}_1,...,\mathbf{b}_n]^T$ és **x** függvényeként – felhasználva, hogy $\mathbf{s} = \mathbf{Bx}$ – a következő egyenlőséggel:

$$f_{\underline{X}}(\mathbf{x}) = |\det \mathbf{B}| \prod_i f_i\left(\mathbf{b}_i^T \mathbf{x}\right). \qquad /9/$$

Ha T számú FAE megfigyelésünk van $\underline{X}$-re, amit jelöljön $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)$, akkor az $L(\mathbf{B})$ ún. likelihood-függvény a sűrűségfüggvények szorzataként áll elő, tehát

$$L(\mathbf{B}) = \prod_{t=1}^{T} |\det \mathbf{B}| \prod_{i=1}^{n} f_i\left(\mathbf{b}_i^T \mathbf{x}(t)\right). \qquad /10/$$

Ez a függvény tehát annak a likelihoodját mutatja meg, hogy adott **B** mellett a T darab minta éppen $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)$ lesz. Algebrailag egyszerűbb a log-likelihooddal számolni, ami a következő formában adott:

$$\log L(\mathbf{B}) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i\left(\mathbf{b}_i^T \mathbf{x}(t)\right) + T \log |\det \mathbf{B}|. \qquad /11/$$

Mindkét oldalt *T*-vel osztva a következő összefüggést kapjuk:

$$\frac{1}{T} \log L(\mathbf{B}) = \mathbb{E}\left(\sum_{i=1}^{n} \log f_i(\mathbf{b}_i^T \mathbf{x}) \right) + \log |\det \mathbf{B}|. \qquad /12/$$

Itt $\mathbb{E}$ nem az elméleti várható értéket jelöli, hanem a mintából számított átlagot. Az azonnal látható, hogy a második tag **B** ortogonalitása miatt mindig nulla. Emellett megmutatható (*Hyvärinen–Karhunen–Oja* [2001]), hogy adott f_i-k esetén a /12/ egyenlőség jobb oldalának első tagja éppen akkor maximális, ha $\mathbf{y} = \mathbf{Bx}$ egyenlőség teljesül. Ebben az esetben **y** éppen a független komponensek megfigyelt értékeit adja.

Ezt a megközelítést alkalmazva tehát az egyetlen fennmaradó probléma az f_i sűrűségfüggvények meghatározása. Amennyiben ezekről nincs sejtésünk, akkor meghatározásukhoz nemparaméteres becslésre, vagy valamilyen eloszláscsalád kiválasztására és paraméteres becslésre van szükség. Ezekre több megoldási lehetőség is ismert (*Hyvärinen–Oja* [2000]).

1.2.3. Kölcsönös információ

Valószínűségi változók függetlenségének egy másik kiváló mérőszáma lehet a kölcsönös információ. A /6/ egyenlet jelöléseit használva *n* darab valószínűségi változó kölcsönös információja:

$$I\left(Y_1, Y_2, ..., Y_n\right) = \left[\sum_{i=1}^{n} H\left(Y_i\right)\right] - H\left(\underline{Y}\right), \qquad /13/$$

ahol $\underline{Y}$ az összes Y_i-t tartalmazó vektor, $H\left(\underline{Y}\right)$ pedig $\underline{Y}$ együttes eloszlásának entrópiája, definíció szerint:

$$H\left(\underline{Y}\right) = -\iint\limits_{supp\left(f_{\underline{Y}}\right)} \cdots \int f_{\underline{Y}}\left(y_1, y_2, \ldots, y_n\right) \log f_{\underline{Y}}\left(y_1, y_2, \ldots, y_n\right) \mathrm{d}y_1 \mathrm{d}y_2 \cdots \mathrm{d}y_n. \qquad /14/$$

Látható, hogy függetlenség esetén ez a mérőszám nulla, hiszen ekkor $H\left(\underline{Y}\right) = \sum_{i=1}^{n} H\left(Y_i\right)$.

Ennek segítségével kifejezhető a /3/ egyenlettel adott transzformációval transzformált Y_i valószínűségi változók kölcsönös információja:

$$I\left(Y_1, Y_2, ..., Y_n\right) = \left[\sum_{i=1}^{n} H\left(\mathbf{b}_i^T \underline{X}\right)\right] - H\left(\underline{X}\right) - \log\left|\det \mathbf{B}\right|. \qquad /15/$$

Ez az egyenlet szintén használható egy optimalizációs eljárás költségfüggvényeként, amennyiben $H\left(Y_i\right) = H\left(\mathbf{b}_i^T \underline{X}\right)$ és $H\left(\underline{X}\right)$ valamilyen becslése rendelkezésünkre áll, ahogy az a negentrópia meghatározásakor is szükséges volt.

Megjegyzendő, hogy a megközelítés a negentrópián alapuló módszerrel egyenértékű egyenletekre vezet, sőt az ML-módszerrel való rokonsága is egyszerűen kimutatható (*Hyvärinen–Karhunen–Oja* [2001]). Ha ugyanis a /12/ egyenlet jobb oldalán az f_i ismeretlen sűrűségfüggvények éppen a megfelelő $\mathbf{b}_i^T\mathbf{x}$-ek sűrűségfüggvényével lennének egyenlők, akkor az egyenlet a következő alakot öltené:

$$\frac{1}{T} \log L(\mathbf{B}) = -\sum_{i=1}^{n} H\left(\mathbf{b}_i^T \mathbf{x}\right) + \log\left|\det \mathbf{B}\right|, \qquad /16/$$

ennek jobb oldala pedig láthatóan csak egy konstansban különbözik /15/ jobb oldalától. Az ehhez szükséges feltevés pedig egyáltalán nem légből kapott, hiszen mivel az

f_i-k ismeretlenek, ezeket általában $\mathbf{b}_i^T\mathbf{x}$ segítségével becsüljük, vagyis az ekvivalencia a gyakorlatban valóban fennáll.

1.2.4. Kumuláns tenzor

Ahogy azt a csúcsosság esetében is láthattuk, a valószínűségi változók függetlenségének vizsgálatakor a negyedrendű statisztikák nagy segítséget nyújthatnak. Nem meglepő tehát, hogy negyedrendű kumulánsok (*Kendall–Stuart–Ord* [1983]) vizsgálatával a független komponensekre való felbontás szintén elvégezhető.

Az S_i, S_j, S_k, S_l valószínűségi változók negyedrendű keresztkumulánsa definíció szerint:

$$\begin{aligned}\operatorname{cum}\left(S_i,S_j,S_k,S_l\right) = \mathbb{E}\left(S_iS_jS_kS_l\right) - \mathbb{E}\left(S_iS_j\right)\mathbb{E}\left(S_kS_l\right) - \\ -\mathbb{E}\left(S_iS_k\right)\mathbb{E}\left(S_jS_l\right) - \mathbb{E}\left(S_iS_l\right)\mathbb{E}\left(S_jS_k\right).\end{aligned} \qquad /17/$$

Definiáljuk a negyedrendű kumuláns tenzort, mint lineáris operátort az $n\times n$ méretű mátrixok terében, az S_i, $i=1,2,\ldots,n$ valószínűségi változók negyedrendű keresztkumulánsai segítségével:

$$\mathbf{F}_{\underline{S}}\left(\mathbf{M}\right)_{ij} = \sum_{kl} m_{kl}\cdot\operatorname{cum}\left(S_i,S_j,S_k,S_l\right), \qquad /18/$$

ahol $\mathbf{F}_{\underline{S}}\left(\mathbf{M}\right)_{ij}$ a tenzor általi transzformáció eredményének *ij*-edik eleme, m_{kl} pedig a transzformált **M** mátrix *kl*-edik eleme. Ez a négydimenziós tenzor nyilván szimmetrikus,[5] tehát diagonalizálható, azaz létezik olyan **K** sajátmátrix és λ sajátérték (*Praszolov* [2005]), hogy:

$$\mathbf{F}(\mathbf{K}) = \lambda\mathbf{K}. \qquad /19/$$

Megmutatható, hogy a tenzornak *n* nemnulla sajátértéke van, melyek éppen az S_i valószínűségi változók csúcsosságaival egyenlők (*Hyvärinen–Karhunen–Oja* [2001]). Állítsuk elő az $\underline{X}$ bemeneti adatokból a $\underline{V}$ fehérített adatokat, és képezzük ezekből az $\mathbf{F}_{\underline{V}}$ kumuláns tenzort. Megmutatható az is, hogy ekkor a **K** sajátmátrixok mindegyike $\mathbf{K}_i = \mathbf{w}_i^T\mathbf{w}_i$ alakú, azaz a szétválasztómátrix egy $\mathbf{w}_i$ oszlopának önma-

[5] $\operatorname{cum}\left(S_i,S_j,S_k,S_l\right)$ értéke nem függ *i*, *j*, *k* és *l* sorrendjétől.

gával vett diadikus szorzataként áll elő. Ennek megfelelően a kumuláns tenzor sajátmátrixainak sajátvektorai a szétválasztómátrix egy-egy oszlopát adják.

Megjegyzendő a módszerrel kapcsolatban, hogy ebben a formában sok számítást és a nagyméretű tenzorok miatt sok memóriát is igényel, így jellemzően csak kisdimenziós esetekben használják.

1.3. Az adatok fehérítése

A koktélparti-probléma vizsgálatakor természetesen merül fel az ötlet, hogy megoldja-e a problémát az $\underline{X}$ megfigyelések fehérítése. Fehérítésnek nevezünk egy transzformációt, ha a transzformált valószínűségi változók mindegyikének várható értéke nulla, korrelációs mátrixuk pedig az egységmátrix lesz. Egy ilyen, „fehér" változókat előállító transzformáció például $\underline{X}$ szorzása korrelációs mátrixának $-1/2$-edik hatványával, az egyes X_i-k centrálása után.

Egyértelmű, hogy az ICA bemeneti adatainak fehérítésével előállított $\underline{V}$ változók, bár korrelálatlanok lesznek, de nem feltétlenül függetlenek. Mivel $\underline{V}$ bármely ortogonális transzformációja szintén fehér,[6] ezért csupán ez a feltétel nem elegendő annak eldöntésére, hogy $\underline{V}$ a valódi független komponenseket tartalmazza-e vagy csak korrelálatlanokat.

Gyakorlati megfontolásként azonban megemlítendő, hogy bár nem nyújt közvetlen megoldást a problémára, mégis érdemes fehérített adatokat használni az ICA számításakor. Ennek megértéséhez írjuk fel a /3/ egyenletet a fehérített adatokra. Jelölje **W** az ún. szétválasztómátrixot, amely előállítja a független komponenseket a fehérített adatokból, melyeket a **V** mátrix által reprezentált lineáris transzformációval állítunk elő.[7] Ezekkel a jelölésekkel a kapott egyenlet:

$$\underline{S} = \mathbf{W}\underline{V} = \mathbf{WV}\underline{X} = \mathbf{WVA}\underline{S}, \qquad /20/$$

Az $\underline{S}$ független komponensek korrelációs mátrixa biztosan az egységmátrix, tehát ha felírjuk az

$$\mathbf{I} = \mathbb{E}\left(\underline{S}\underline{S}^T\right) = \mathbb{E}\left(\mathbf{W}\underline{V}\underline{V}^T\mathbf{W}^T\right) = \mathbf{W}^T\mathbf{W} \qquad /21/$$

egyenletet, láthatjuk, hogy a szétválasztómátrix ortogonális lesz. Ha tehát fehérített adatokon dolgozunk, a /2/ egyenlet a következő formát ölti:

[6] $\underline{Z} = \mathbf{U}\underline{V}$ esetén $\mathbb{E}(\underline{Z}\underline{Z}^T) = \mathbb{E}(\mathbf{U}\underline{V}\underline{V}^T\mathbf{U}^T) = \mathbf{UIU}^T = \mathbf{I}$, ha **U** ortogonális, és $\underline{V}$ fehér; $\mathbb{E}\underline{Z}$ pedig nulla marad, hiszen erre **U** nincs hatással.

[7] Amennyiben $\underline{X}$ várható értéke nem nulla, ezt természetesen $\mathbb{E}\underline{X}$ kivonásával korrigálnunk kell.

$$\underline{V} = \mathbf{V}\underline{X} = \mathbf{VA}\underline{S} = \mathbf{W}^T\underline{S}, \qquad /22/$$

ahol a $\mathbf{W}^T$ ortogonális mátrix az adatokat fehérítő keverőmátrix, mely egyben a szétválasztómátrix inverze. Numerikus szempontból tehát mindenképpen kedvező a fehérített adatok használata az egyszerű inverzszámítás miatt, de a fehérített bemenet feltételezése az elméleti megfontolásokat is egyszerűsíti.

1.4. A módszer korlátai

Az eddig elmondottak alapján tehát az ICA által használt modell két előfeltevéssel él: az egyik, hogy a szétválasztandó S_i komponensek függetlenek; a másik pedig, hogy a komponensek közül csak legfeljebb egy lehet normális eloszlású. A két feltétel közül az utóbbi az erősebb; az ICA képes kismértékben korreláló komponenseket is szétválasztani két olyan összetevőre, melyek többitől való függetlensége – a korábbiakban leírt mérőszámok és módszerek alapján – maximális.

Meg kell emellett említenünk a módszer 1.2. pontban már érintett néhány tulajdonságát, melyek nagyban befolyásolják az ICA alkalmazhatóságát: nem tudjuk meghatározni a független komponensek számát, sorrendjét és varianciáját. Ezek a nehézségek abból adódnak, hogy mind $\underline{S}$, mind **A** ismeretlenek, így a probléma aluldeterminált.

Egyrészt az 1.1. pontban ismertetett probléma feltételezi, hogy a független komponensek és a keverékek száma azonos. Az nyilvánvaló, hogy a probléma aluldetermináltsága már nem kezelhető abban az esetben, ha több független komponensre van szükségünk, mint ahány kevert jel rendelkezésünkre áll. Mi a helyzet azonban akkor, ha több kevert jel áll rendelkezésünkre annál, mint ahány komponens keverékei ezek a jelek? Ekkor ugyan az alapfeltételezésünk nem áll fenn, a probléma azonban kezelhető. A leggyakoribb megoldás az, ha a függetlenítés előtt főkomponens analízist (PCA) használva adunk becslést a dimenzióra (erről a következő pontban részletesebben lesz szó). Emellett néhány algoritmus esetén arra is van lehetőség, hogy közvetlenül a kevert jelek számánál kevesebb komponenst állítsunk elő. Ezt a megközelítést alkalmazhatjuk, ha a komponensek száma valahonnan – például elméletileg – ismert, vagy ha félő, hogy a PCA segítségével végzett dimenzióredukció során értékes adatot veszítünk.[8]

Másrészt probléma, hogy egy konstans szorzó bármely eredeti komponensben eliminálható az **A** mátrix megfelelő $\mathbf{a}_i$ oszlopának az adott konstanssal való osztásával. Ilyen módon változtatható bármely komponens varianciája anélkül, hogy a mo-

[8] Ebben az esetben viszont az ICA-algoritmusok nem adnak becslést a komponensek számára vonatkozóan, így legtöbbször az egyetlen használható módszer a próbálgatás marad.

dellel ellentmondásba kerülnénk. Emiatt érdemes azzal a feltevéssel élni, hogy a komponensek varianciája 1, így csak az előjel okozhat problémát, hiszen a komponensek még így is szorozhatók (–1)-gyel anélkül, hogy ez befolyásolná a modellt.

Harmadrészt, a komponensek sorrendjének meghatározása szintén önkényes, hiszen a **W** mátrix sorainak felcserélésével a komponensek nem változnak, csak azok sorrendje.

Érdekes azonban, hogy ezen három megfontolástól eltekintve a /2/ egyenlet szerinti ICA-modell megoldása egyértelmű, azaz a komponensek egyértelműen állnak elő a kevert jelekből, amennyiben a keverőmátrix invertálható (*Comon* [1994]).

1.5. Egy hasonló módszer: a főkomponens analízis (PCA)

Az előzőkben leírt módszer lényegének megértéséhez érdemes összehasonlítani azt egy másik hasonló célú, pénzügyi adatok elemzéséhez gyakrabban használt eljárással, a PCA-val.

A PCA célja többféleképp is megragadható. A legkézenfekvőbb felfogás szerint ez egy lineáris transzformáción alapuló dimenzióredukciós módszer: ha adott egy n dimenziós adatbázis, akkor a PCA azt egy másik, adott esetben kevesebb dimenziós koordinátarendszerben ábrázolja lineáris transzformáció segítségével úgy, hogy a megőrzött információ mennyisége – mérve ezt azzal a várható négyzetes hibával amit a kevesebb dimenzión történő ábrázolás miatt vétünk – a lehető legkisebb legyen adott dimenzióra az összes lehetséges lineáris transzformáció körében.

A PCA először standardizálja az adatokat, majd megkeresi azt a tengelyt, amelyre vetítve az adatbázist, a legnagyobb lesz annak varianciája. Ez lesz az első főkomponens. Belátható, hogy ha csak egyetlen dimenziót használhatunk az adatbázis ábrázolására, akkor ezt érdemes használni ahhoz, hogy az információveszteséget minimalizáljuk. (Már ebből is látható, hogy itt bizonyos értelemben a variancia mutatja meg egy adott tengely által hordozott információt.) Ezt követően megkeresi azt a tengelyt, mely az előbbire merőleges tengelyek közül a legtöbb információt őrzi meg (azaz rávetítve legnagyobb a variancia) és így tovább.

Az új koordinátarendszerről tehát elmondható, hogy a tengelyei csökkenő ,,fontossági" sorrendbe lesznek állítva, aszerint, hogy a tengelyekre vett vetületek vagy komponensek, mennyire járulnak hozzá az eredeti adatok visszaállításához. A ,,dimenzióredukció" kifejezés azért is jogos, mert bebizonyítható, hogy ez a konstrukció az, ami egy eredetileg n dimenziós adatbázist optimálisan reprezentál $m \le n$ dimenzióban.[9] Belátható, hogy az új tengelyek irányai az eredeti adatbázis korreláci-

[9] Optimális alatt azt értve, hogy adott dimenziószám mellett a reprezentáció hibája a lehető legkisebb lesz a lineáris transzformációval elérhető reprezentációk között; a hibát most négyzetes értelemben mérve.

ós mátrixának sajátvektorainak irányaival fognak egyezni, és a megtalált komponensek korrelálatlanok lesznek egymással (*Jolliffe* [2010]).

Az ICA ezzel szemben nemcsak a korrelálatlanságot, de a függetlenséget is előírja az egyes komponensek számára. Amint az tehát sejthető, az ICA a PCA-val rokon módszer, hiszen ahhoz hasonlóan az adatok – ezek az ICA esetén jellemzően idősorok – egy speciális reprezentációját keresi. Ez viszont nem jelenti azt, hogy az ICA helyettesíthetné ezt a módszert, mert ahogy azt később is látni fogjuk: a két eszközt különböző problémák megoldására használhatjuk, és korlátaik is különbözők.

A két módszer rokonságának megértése érdekében érdemes a PCA formális matematikai leírásán keresztülhaladnunk. Ehhez használjuk fel a következő definíciókat: legyen $\underline{X} = \left[X_1, X_2, \ldots, X_n\right]^T$ egy n dimenziós valószínűségi vektorváltozó, $\underline{Y} = \left[Y_1, Y_2, \ldots, Y_n\right]^T$ az $\underline{X}$ transzformációja után kapott valószínűségi vektorváltozó, $w_i = \left[w_{1i}, w_{2i}, \ldots, w_{ni}\right]^T$ pedig a tér bázisvektorai közül egy, melyek együttesen a $\mathbf{W}$ transzformációs mátrix oszlopterét feszítik ki,[10] és melyeket ortonormáltra választunk meg. A PCA probléma alapjául szolgáló egyenlet pedig legyen:

$$\underline{Y} = \mathbf{W}^T \underline{X}. \qquad /23/$$

Keressük azt a $\mathbf{W}\, n \times n$ méretű ortonormált transzformációs mátrixot, melyre igaz, hogy $\underline{X}$-et oszlopterének bármely $m < n$ dimenziós alterére merőlegesen vetítve $\underline{X}$ legkisebb négyzetes hibájú becslését kapjuk, azaz minimalizáljuk a következő költségfüggvényt minden m-re:

$$C_{PCA}\left(\mathbf{W}_{n\times m}\right) = \mathbb{E}\left\|\underline{X} - \sum_{i=1}^{m}\left(\mathbf{w}_i^T \underline{X}\right)\mathbf{w}_i\right\|^2 = \mathbb{E}\left\|\underline{X} - \mathbf{W}_{n\times m}\mathbf{W}_{n\times m}^T \underline{X}\right\|^2. \qquad /24/$$

Felmerül a kérdés, hogy kiterjeszthető-e a PCA olyan esetekre, amikor a keresett főkomponensek a bemeneti adatok valamilyen nemlineáris transzformációja által adottak (nemlineáris PCA – NLPCA). A /24/ lineáris kritériumot ilyen esetekben a főkomponensekre alkalmazott g_i nemlineáris függvényekkel kell módosítanunk a következőképpen:

$$C_{NLPCA}\left(\mathbf{W}_{n\times m}\right) = \mathbb{E}\left\|\underline{X} - \sum_{i=1}^{m} g_i\left(\mathbf{w}_i^T \underline{X}\right)\mathbf{w}_i\right\|^2 = \mathbb{E}\left\|\underline{X} - \mathbf{W}_{n\times m}\mathbf{g}\left(\mathbf{W}_{n\times m}^T \underline{X}\right)\right\|^2, \qquad /25/$$

[10] A mátrix oszloptere az oszlopai által kifeszített altér, mely a mátrix által reprezentált transzformáció képtere.

ahol $\mathbf{g}\left(\mathbf{W}^T \underline{X}\right)$ egy oszlopvektor, *i*-edik eleme $g_i\left(\mathbf{w}_i^T \underline{X}\right)$.

Az 1.3. pontban említett okokból állítsuk elő az $\underline{X}$-ből a $\underline{V}$ fehérített változókat, és tegyük fel, hogy ezen vektor és $\underline{Y}$ dimenziója megegyezik, azaz $m = n$. Ekkor /23/ egyenlet jelöléseit használva igaz a következő összefüggés:

$$\begin{aligned} C_{NLPCA}(W) &= \mathbb{E}\left\|\underline{V} - \mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right\|^2 = \mathbb{E}\left[\left(\underline{V} - \mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right)^T \left(\underline{V} - \mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right)\right] = \\ &= \mathbb{E}\left[\left(\underline{V} - \mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right)^T \mathbf{W}\mathbf{W}^T \left(\underline{V} - \mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right)\right] = \\ &= \mathbb{E}\left[\left(\mathbf{W}^T \underline{V} - \mathbf{W}^T\mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right)^T \left(\mathbf{W}^T \underline{V} - \mathbf{W}^T\mathbf{W}\mathbf{g}\left(\mathbf{W}^T \underline{V}\right)\right)\right] = \\ &= \mathbb{E}\left\|\underline{Y} - g\left(\underline{Y}\right)\right\|^2 = \sum_{i=1}^{n} \mathbb{E}\left[Y_i - g_i\left(Y_i\right)\right]^2. \end{aligned} \qquad /26/$$

Legyen most minden *i*-re igaz, hogy

$$g_i(y) = y^2 + y. \qquad /27/$$

Látható, hogy ekkor a /26/ egyenlet éppen a fehérített adatok esetén érvényes /5/ egyenlettel lesz ekvivalens:

$$C_{NLPCA}\left(\mathbf{W}_{m\times n}\right) = \sum_{i=1}^{m} \mathbb{E} Y_i^4. \qquad /28/$$

A konkrét feladat vonatkozásában tehát az ICA-probléma megoldása tulajdonképpen egy NLPCA-feladat megoldásával egyenértékű, ha a megfelelő műveleteket fehérített adatokon végezzük.

Az elméleti modellek hasonlósága ellenére azonban az ICA és a PCA más-más típusú problémák megoldására hivatott, ahogy azt a következő fejezetben szereplő példáink is szemléltetik majd. Ugyanakkor a két módszer kiválóan képes kiegészíteni egymást, a PCA ugyanis alkalmas az ICA egyik hiányosságának – a komponensek számának meghatározásának – pótlására, hiszen rávilágít arra, hogy hány dimenzióban tudjuk az adatainkat megfelelően kis hibával ábrázolni. A PCA tehát utal arra, hogy az ICA-algoritmus legfeljebb hány független komponenst találhat az adott mintákat használva, emellett maga a felbontás is egyszerűsödik, hiszen amennyiben a főkomponenseket tekintjük kevert jeleknek, a keverőmátrix biztosan ortogonális lesz, és így invertálható is.

2. Empirikus vizsgálat

A gyakorlati alkalmazások elemzése során először azt mutatjuk be, hogy az előálló komponensek valóban függetlenek. Ezt követően az ICA és a PCA összehasonlítására kerül sor, annak érdekében, hogy az elméleti megfontolások szemléltetése mellett rávilágítsunk a két módszer közötti lényeges gyakorlati különbségekre is. Végezetül az ICA néhány gyakorlati sajátosságát elemezzük szimulált adatok segítségével három dimenzió – a minta elemszáma, a komponensszám és az adatok eloszlása – alapján.

Az ICA egyik leggyakoribb implementálási módja a fixpont-megközelítésen alapuló FastICA-algoritmus (*Hyvärinen–Oja* [2000]). Az algoritmus fehérített adatokon dolgozik, és feltételes optimalizációt végez. A $C_{FastICA} = \mathbb{E}\left(f\left(\mathbf{w}^T\mathbf{x}\right)\right)$ költségfüggvény minimumát keresi $\|\mathbf{w}\|^2 = 1$ feltétel mellett,[11] ahol *f* kevés kivételtől eltekintve tetszőleges nemkvadratikus függvény lehet, deriváltját *g*-vel jelöljük. Az algoritmus alapváltozata külön-külön határozza meg a független komponenseket, lépései egy független komponens meghatározásához (a korábbi jelöléseket használva):

1. Kezdeti – véletlen – w_0 súlyvektor kiválasztása.

2. $\mathbf{w}_{k+1} = \mathbb{E}\left(\mathbf{x}g\left(\mathbf{w}_k^T\mathbf{x}\right)\right) - \mathbb{E}\left(g'(\mathbf{w}_k^T\mathbf{x})\right)\mathbf{w}_k$.

3. $\mathbf{w}_{k+1} = \dfrac{\mathbf{w}_k + 1}{\|\mathbf{w}_k + 1\|}$.

4. Ismétlés 2-től a konvergencia eléréséig. A kiszámított független komponens ekkor $\mathbf{w}_k^T\mathbf{x}$.

A várható értékeket a számítások során az algoritmus mintaátlaggal becsli, így teljesítménye szempontjából az idősorok elemszáma fontos tényező. További fontos tulajdonsága a módszernek, hogy *f* megválasztása az algoritmus eredménye szempontjából nem közömbös, bizonyos függvények esetén a kurtózis jobb közelítését kaphatjuk. Az algoritmus láthatóan nem igényli a keresés lépésnagyságát befolyásoló ún. bátorsági tényezők hangolását, konvergenciatulajdonságai pedig kiválók.

Ahogy azt említettük, ez az egyik legelterjedtebben használt ICA-algoritmus. Számos változatát dolgozták ki, és tulajdonságai, alkalmazási lehetőségei is széles körben ismertek. A részletekbe menő ismertetés ennél fogva meghaladná a dolgozat

[11] Megmutatható, hogy az így leírt algoritmus a kurtózison, mint a nemnormalitás mérőszámán alapul. Az algoritmusnak létezik negentrópiát használó változata is (*Hyvärinen–Karhunen–Oja* [2001]).

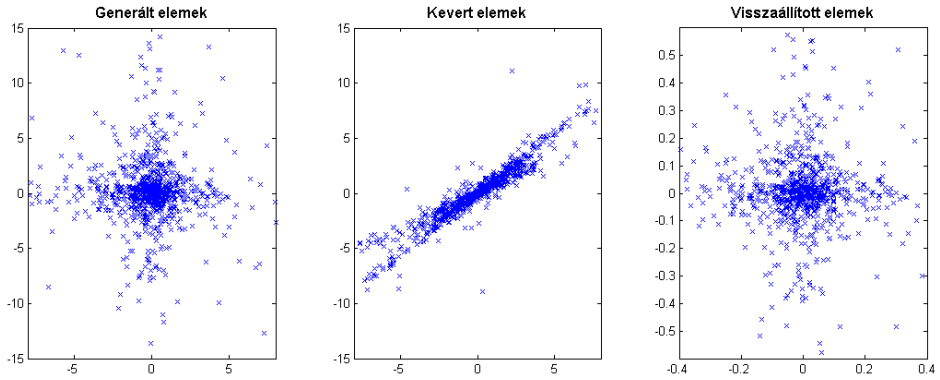
kereteit, ehhez lásd például *Horváth et al*. [2006], *Hyvärinen–Oja* [2000], *Hyvärinen–Karhunen–Oja* [2001].

2.1. Független komponensek

Tegyük tehát először próbára a független komponens analízist. Példánkban előállítottunk egy kétváltozós adathalmazt két, egymástól független 1 szabadságfokú *t*-eloszlásból. Ezek után az adatokat egy véletlenszerűen megválasztott értékekkel rendelkező keverőmátrix által reprezentált lineáris transzformációval képeztük le. A kevert adatokon a FastICA algoritmust lefuttatva azt tapasztaltuk, hogy a visszaállított és az eredeti komponensek majdnem teljesen megegyeztek.

Az eltérés annak tulajdonítható, hogy míg az eredetileg generált jelek között volt alacsony szintű összefüggés – a korrelációs együttható értéke 0,012 – addig a kevert jelekből visszaállított adatok közötti korreláció mértéke 10^{-15} nagyságrendű. Az 1. ábrán látható, hogy a generált és a visszaállított elemek közötti eltérés minimális, feltűnhet azonban, hogy a visszakapott adathalmaz a kiindulási mínusz egyszerese. Ez annak köszönhető, hogy a komponensek skálázása tetszőleges lehet (lásd az 1.4. pontot).

1. ábra. Generált, kevert és visszaállított adathalmazok



Mivel azonban a korrelációs együtthatóval a függetlenség nem mérhető, így a következőkben az együttes eloszlás, valamint a peremeloszlások segítségével vizsgáljuk ezt a kérdést. Az együttes- és perem-sűrűségfüggvényeket magfüggvényes sűrűségbecslés[12] segítségével állítottuk elő, és a függetlenség vizsgálatához az együttes sűrűségfüggvény értékéből levontuk a perem-sűrűségfüggvények szorzatát, majd a különbség négyzetes integrálját vettük. Az így kapott mutatót, az integrált négyzetes hibát (integrated squared error – ISE) a következő egyenlet mutatja:

[12] Lásd *Scott* [1992] vagy *Terrell–Scott* [1992].

$$ISE = \iint_{\mathbb{R}^2} \left[\left(\hat{f}_{xy} \right) - \left(\hat{f}_x \right) \left(\hat{f}_y \right) \right]^2 \mathrm{d}x\mathrm{d}y. \qquad /29/$$

Az 1. táblázat mutatja az ISE-értékeket a generált, kevert és visszaállított adatok esetén.

1. táblázat

Együttes eloszlás és a peremeloszlások szorzatának különbsége

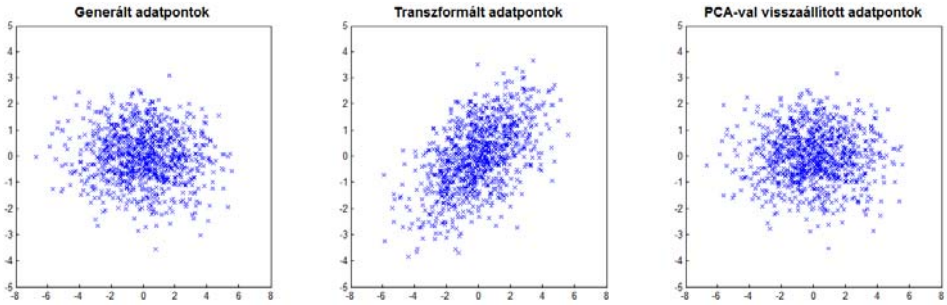
Jel	Generált	Kevert	Visszaállított
ISE	0,1353	0,4916	0,0068

Látható, hogy az eredeti adatok között is volt kismértékű összefüggés, azonban a visszakapott adatokra a függetlenítés eredményeképp ez az érték jelentősen csökkent. A keverés után előállt adatokra a különbség – és így az egymástól való függés – jelentősnek mondható, továbbá a másik két értéknél legalább egy nagyságrenddel nagyobb.

2.2. Főkomponensek

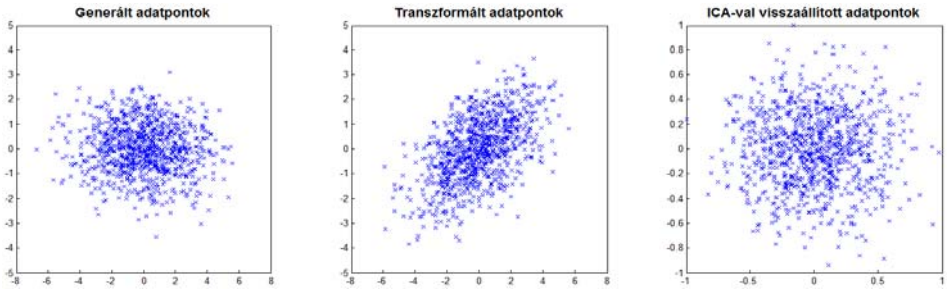
Nézzük, miben különböznek az előbbiektől a PCA segítségével előállt főkomponensek. A vizsgálathoz normális eloszlású mintákat generálunk, a minták szórása az egyik valószínűségi változó esetén 1, a másiknál 2, a várható érték mindkettőnél 0. Az így kapott adatokon lineáris transzformációt, egy 45 fokos forgatást alkalmaztunk. Ezek után a kapott adatok főkomponenseinek meghatározása következett, az eredményeket a 2. ábra szemlélteti.

2 ábra. A PCA eredményeinek szemléltetése normális eloszlású adathalmazokon



Amint látható, a PCA kitűnően visszaállítja az eredeti komponenseket, a visszaállítás átlagos négyzetes hibája 0,005. Az ICA-nak ugyanez a feladat leküzdhetetlen problémát jelent. Egyrészt a skálázási invariancia miatt az eredményt a [–1,1] tartományba normálva kell megjelenítenünk, másrészt a visszaállítás sem megfelelő, az átlagos négyzetes hiba még úgy is 0,2, ha az eredeti adatokat is ugyanebbe a tartományba skálázzuk.

3. ábra. Az ICA eredményeinek szemléltetése normális eloszlású adathalmazokon



Ez az eredmény az elméleti megfontolások alapján várható is volt, hiszen a nemnormalitást használó ICA-algoritmusok képtelenek szétválasztani a független komponenseket akkor, ha azok között egynél több normális eloszlású található (lásd az 1.4. pontot).

2.3. Az ICA és a PCA összehasonlítása

A két módszer természetesen nemcsak adathalmazok, hanem rendezett adatok, idősorok esetén is használható, különbség csak a megjelenítés módjában van. Tekintve, hogy több mint két dimenzió esetén az ábrázolás egyébként is nehézkes lenne, illetve mivel az ICA-t jellemzően idősorok elemzésére használják, a két módszer összehasonlítását idősorok segítségével végeztük. Az idősorokat vagy – az ICA alkalmazási területén gyakrabban használt megnevezésükkel – jeleket úgy választottuk meg, hogy azok szabad szemmel is jól elkülöníthetők legyenek. Az alkalmazott jelek:

1. Chirp-jel: a jelfeldolgozási gyakorlatban használatos, időben változó frekvenciájú jel, időfüggvénye $x(t) = 10\sin\left(2\pi\left[\frac{3}{2}t^2 + \frac{\pi}{2}\right]\right)$.

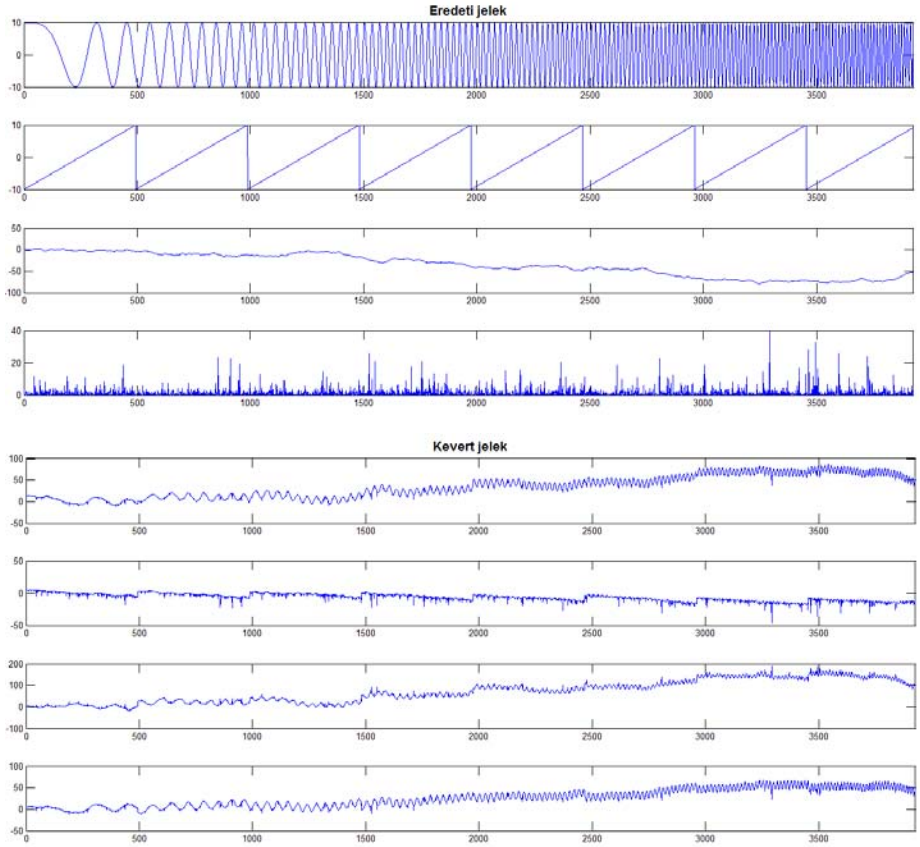
2. Fűrészjel: egyszerűsége miatt gyakran használt jelfeldolgozási mintapélda, időfüggvénye $x(t) = 20\left(\frac{t}{500} - \left\lfloor \frac{t}{500} - \frac{1}{2} \right\rfloor\right)$.

3. Brown-mozgás: nullából induló függvény, melyre igaz, hogy nem átfedő intervallumokra minden növekménye független, és minden $x(t+s)-x(s)$ $N(0,t)$-ből származik.

4. Weibull-eloszlásból származó FAE-minták: az eloszlás sűrűségfüggvénye $f(x)=(2x)^{-\frac{1}{2}}\,\mathrm{e}^{-(2x)^{\frac{1}{2}}}$.

A jelekből azok létrehozása után egy véletlenszerűen választott elemekből álló keverőmátrix segítségével keverékeket állítottunk elő. A jeleket, valamint azok keverékeit a 4. ábra mutatja egy konkrét esetben.

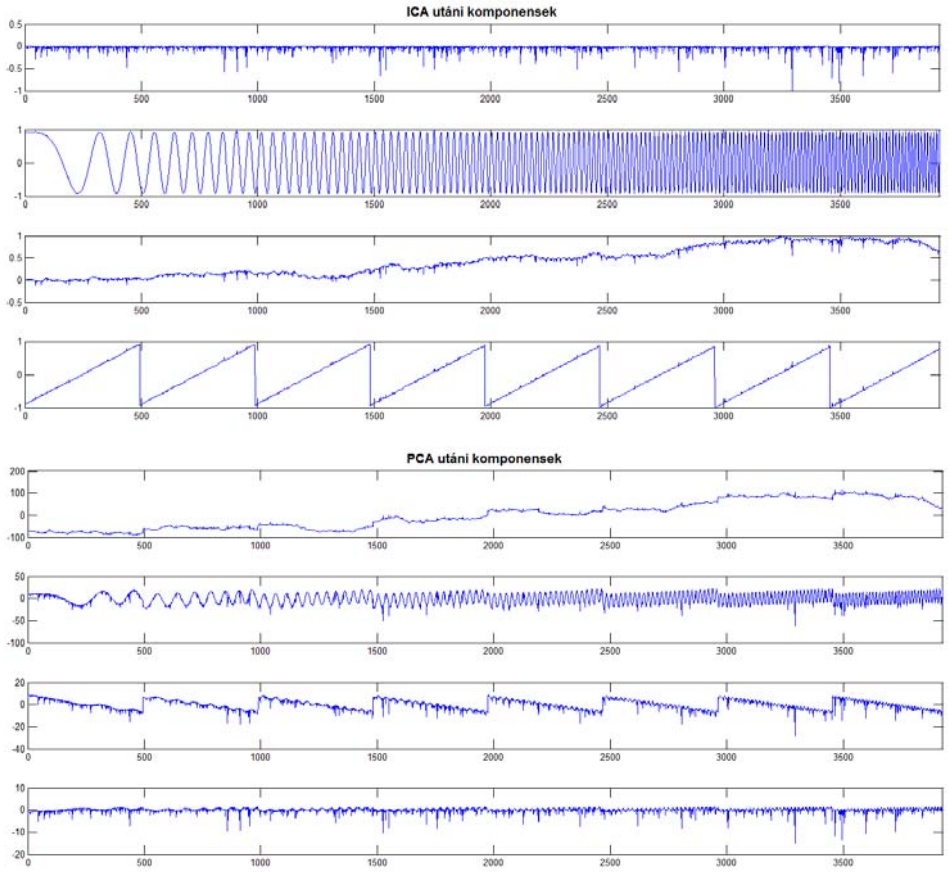
4. ábra. Eredeti jelek és egy véletlen keverőmátrix segítségével előállt keverékeik



A kevert jelekre ezután lefuttattuk a FastICA-algoritmust, valamint a szinguláris értékekre való felbontáson alapuló PCA-algoritmust. A szétválasztások eredményei.

az 5. ábrán láthatók. Megfigyelhető, hogy a PCA közel sem állítja vissza a kiindulási jeleket. Igaz ugyan, hogy a PCA utáni komponensek korrelálatlanok, de a jelalakok korántsem ismerhetők fel, több eredeti jel keveredik bennük. Ezzel szemben az ICA szinte tökéletes eredményt nyújt, minimális torzulás mellett állnak vissza az eredeti jelalakok, természetesen a skálázástól és a sorrendjüktől eltekintve, ahogy azt az 1.4. pontban is említettük.

5. ábra. A keverékek ICA, illetve PCA után adódó komponensei



E példán keresztül tehát a gyakorlatban is láthattuk, mekkora a különbség a két, több szempontból rokonnak minősíthető módszer között, és mennyire más célú alkalmazások esetén indokolt a használatuk. A PCA az 1.5. pontban ismertetett módon, a legnagyobb varianciájú irányokban képes az adathalmazt ábrázolni, és ezzel akár dimenzióredukcióra is lehetőséget ad, hiszen elképzelhető, hogy találunk olyan kevésbé fontos főkomponenseket, melyek elhagyása nem okoz lényeges hibát az

adathalmazra vonatkozóan. Nem alkalmas viszont az adathalmaz mögött megbújó látens struktúrák teljeskörű vizsgálatára.

Az ICA ezzel szemben nem nyújt információt a komponensek számára vonatkozóan, képes viszont az adatokban olyan strukturális összefüggések felismerésére és elkülönítésére, melyek az adatok mögött rejlő, egymástól független hatásokra világítanak rá. Egy példával szemléltetve: a PCA képes például keresztmetszeti adatokon kimutatni, hogy adott koordináta mennyire lényeges az adathalmaz egészét tekintve, és mely más paraméterekkel alkot egyetlen, az adathalmazt a legkisebb hibával leíró komponenst. Az ICA ezzel szemben arra világít rá, hogy az egyes változók milyen más változókkal függnek össze oly módon, hogy azok együttesen a többitől – a lehető legnagyobb mértékben – független hatást reprezentálnak. A cél tehát nem az adathalmaz dimenzionalitásának csökkentése, hanem annak felderítése, hogy mely paraméterek függetlenek egymástól, nem törődve azzal, hogy ezek elhagyása mekkora hibát okoz.

Az eddigiekben leírtak áttekintésére a 2. táblázat ad lehetőséget.

2. táblázat

Az ICA és a PCA összehasonlítása

Előfeltevés	PCA	ICA
Bemeneti adatok eloszlása	Tetszőleges	Nemnormális
Komponensek közötti kapcsolat	Korrelálatlanság	Függetlenség
Alkalmazás célja	Tömörítés, lényegkiemelés	Független hatások elemzése

2.4. Az ICA alkalmazása

Ebben a fejezetben három dimenzió mentén vizsgáljuk az ICA-t mint módszert. Mint azt az előző pontban is láttuk, ha a megfigyelt adatok független komponensekből állnak össze, a komponensek visszaállítása – és ezzel a mögöttes információ kiemelése – az ICA esetén jobb minőségű, mint a PCA-nál, ha a generált adatok nem normális eloszlásból származnak.

A következőkben az elemszám tekintetében végzünk teszteket, a módszer hatékonyságát górcső alá véve, majd a komponensek számának kérdését tárgyaljuk, melynek során rávilágítunk, hogy a komponensszám növelésével hogyan változik a módszer teljesítménye. Ezt követően a generált adatok eloszlásának hatását vizsgáljuk.

A számítások során a legfontosabb mérőszámunk a keverés előtti – eredeti – és a kevert jeleken végzett ICA-val visszaállított komponensek közötti korreláció lesz.

Ennek oka, hogy a keresztkorreláció a nemfüggetlenség, vagyis az ICA esetében a hiba mérőszámának tekinthető. Emellett egyszerűen számítható, és invariáns a lineáris skálázásra, azaz mintegy automatikusan kezeli az ICA végrehajtásakor felmerülő többértelműséget, miszerint a komponensek varianciája tetszőleges lehet. Emellett a komponensek sorrendjének változásából adódó problémákat is megoldja, hiszen a komponensek felcserélése a keresztkorrelációs mátrixnak csak az oszlopait cseréli fel, egyéb tekintetben nem változtatja azokat. Ennek megfelelően mérőszámaink a $\overline{C_{cross}}$ átlagos keresztkorreláció – azaz a keresztkorrelációs mátrix elemeinek átlagos értéke, minden oszlopnál eltekintve annak legnagyobb elemétől – és a $\overline{C_{max}}$ átlagos maximális korreláció, azaz a mátrixoszlopok legnagyobb elemeinek az átlagos értéke.

A tesztek során a következő szcenáriókat elemeztük, minden esetben ezer ismétlést végezve:

– Elemszám: 100, 1 000, 5 000, 10 000.
– Komponensszám: 2, 5, 10, 50, 100.
– Eloszlások: *t*-eloszlás $(\upsilon = 1)$, lognormális $(m = 0,\ s = 1)$ és exponenciális eloszlás $(\lambda = 1)$.

Első kérdésként az vetődött fel, hogy a keverőmátrix elemeinek eloszlása befolyásolja-e a metrikák értékeit. A korábbiakban felvázolt esetekben nem volt jelentős különbség e tekintetben. A tesztek során ugyanazokat az idősorokat kevertük össze különböző keverőmátrixokkal, és azt találtuk, hogy a keresztkorrelációs átlagok a különböző esetekben 0,0002 és 0,0025 között, míg a hozzájuk tartozó szórások 0,0005 és 0,0056 között változtak.

Ettől lényegesen eltérő eredményeket kaptunk, amikor az idősorokat is újrageneráltuk, vagyis amikor a keverőmátrix elemeinek megválasztása után több különböző idősort kevertünk össze ugyanazzal a keverőmátrixszal, és az átlagos kereszt-, valamint maximális korrelációk átlagát és szórását kalkuláltuk. Így fontosabbnak tartottuk, hogy mind az idősorokat, mind a keverőmátrix elemeit többször szimuláljuk, és azokból számoljuk az alkalmazott mérőszámok értékeit.

A számítások eredményei a 3–5. táblázatokban láthatók. A tesztek során az említett eloszlásokból generáltunk adott elemszámú idősorokat, amelyeket véletlen számokból előállított keverőmátrix segítségével kevertük össze. Az így nyert kevert jelekre futtattuk le az ICA-t, és kiszámítottuk az előbbiekben leírt korrelációs mérőszámokat az eredeti és a visszaállított komponenseket használva. A táblázatokban az első oszlop mutatja, hogy idősoronként hány véletlen elemet generáltunk az adott eloszlásból, míg az első sor azt, hogy hány idősort és ezzel együtt komponenst szimuláltunk. Az adott elemszámhoz tartozó értékeknél az első sorban az átlagos keresztkorrelációk és átlagos maximális korrelációk átlagai, míg a második sorban azok szórásai szerepelnek, ezer szimulációból számítva.

2.4.1. A komponensszám és az elemszám hatása

Amint a 3. táblázatban látható, 50-nél kevesebb *t*-eloszlású komponens esetén a komponensszám növelése alig befolyásolja az eredményt, míg az elemszám növelésének jelentős hatása van a keresztkorrelációkra. 5 000 elemű idősorok esetén viszont akár 100 független komponens is elegendően kis hibával különíthető el egymástól.

3. táblázat

*Komponenesek átlagos keresztkorrelációs és átlagos maximális korrelációs értékeinek átlagai és szórásai**

Elemszám	Komponensszám				
	2	5	10	50	100
	$\overline{C_{cross}}$				
100	0,0854	0,0845	0,0847	0,0748	0,0558
	(0,0930)	(0,0327)	(0,0171)	(0,0033)	(0,0012)
1 000	0,0253	0,0259	0,0255	0,0261	0,0268
	(0,0327)	(0,0122)	(0,0061)	(0,0019)	(0,0014)
5 000	0,0115	0,0116	0,0117	0,0115	0,0116
	(0,0136)	(0,0079)	(0,0036)	(0,0008)	(0,0005)
10 000	0,0083	0,0081	0,0083	0,0082	0,0082
	(0,0109)	(0,0036)	(0,0027)	(0,0007)	(0,0004)
	$\overline{C_{max}}$				
100	0,9906	0,9639	0,9190	0,6181	0,5098
	(0,0279)	(0,0347)	(0,0372)	(0,0264)	(0,0139)
1 000	0,9990	0,9956	0,9909	0,9474	0,8881
	(0,0083)	(0,0105)	(0,0096)	(0,0016)	(0,0141)
5 000	0,9998	0,9989	0,9974	0,9885	0,9764
	(0,0008)	(0,0064)	(0,0057)	(0,0044)	(0,0048)
10 000	0,9999	0,9996	0,9988	0,9937	0,9874
	(0,0007)	(0,0015)	(0,0033)	(0,0034)	(0,0034)

* Ezer ismétlés esetén, 1 szabadságfokú *t*-eloszlással generált idősorokkal.

2.4.2. Különbségek különböző eloszlások esetén

A 4. és az 5. táblázatból kitűnik, hogy bár a tendenciák exponenciális és lognormális eloszlásra is érvényesek, az algoritmus teljesítménye azonban alulmarad

a *t*-eloszlás esetén tapasztaltakhoz képest. Ez azt mutatja, hogy a szeparálhatóság annak függvénye is, hogy a komponensek milyen eloszlásúak. A FastICA-algoritmusról elmondottak alapján ennek oka érthető: a különbség abban rejlik, hogy az eloszlások közül csúcsosságuk alapján az exponenciális hasonlít leginkább a normális eloszlásra: a használt exponenciális eloszlás kurtózisa 9, a lognormálisé $e^4 + 2e^3 + 3e^2 - 3 \approx 113,94$, míg a *t*-eloszlásé nem meghatározott (az integrál a végtelenhez divergál).

4. táblázat

*Komponensek átlagos keresztkorrelációs és átlagos maximális korrelációs értékeinek átlagai és szórásai**

Elemszám	Komponensszám				
	2	5	10	50	100
			$\overline{C_{cross}}$		
100	0,1393	0,1425	0,1398	0,0968	0,0717
	(0,1249)	(0,0425)	(0,0198)	(0,0014)	(0,0005)
1 000	0,0526	0,0543	0,0555	0,0733	0,0684
	(0,0504)	(0,0176)	(0,0092)	(0,0037)	(0,0006)
5 000	0,0250	0,0253	0,0252	0,0277	0,0378
	(0,0189)	(0,0068)	(0,0035)	(0,0013)	(0,0022)
10 000	0,0184	0,0183	0,0182	0,0189	0,0208
	(0,0147)	(0,0057)	(0,0023)	(0,0006)	(0,0009)
			$\overline{C_{max}}$		
100	0,9804	0,9198	0,8222	0,4739	0,3961
	0,0424	0,0526	0,0526	0,0142	0,0079
1 000	0,9972	0,9878	0,9711	0,6943	0,4309
	0,0098	0,0140	0,0165	0,0347	0,0120
5 000	0,9995	0,9977	0,9950	0,9637	0,8290
	0,0009	0,0020	0,0024	0,0073	0,0242
10 000	0,9997	0,9987	0,9975	0,9848	0,9575
	0,0006	0,0041	0,0012	0,0021	0,0081

* Ezer ismétlés esetén, $\lambda = 1$ paraméterű exponenciális eloszlással generált idősorokkal.

5. táblázat

*Komponensek átlagos keresztkorrelációs és átlagos maximális korrelációs értékeinek átlagai és szórásai**

Elemszám	Komponensszám				
	2	5	10	50	100
	$\overline{C_{cross}}$				
100	0,0821	0,0894	0,0881	0,0782	0,0603
	(0,0786)	(0,0303)	(0,0157)	(0,0026)	(0,0009)
1 000	0,0343	0,0331	0,0332	0,0345	0,0370
	(0,0450)	(0,0125)	(0,0068)	(0,0018)	(0,0015)
5 000	0,0184	0,0178	(0,0177)	0,0179	0,0184
	(0,0265)	(0,0075)	(0,0038)	(0,0009)	(0,0005)
10 000	0,0134	0,0136	(0,0138)	0,0137	0,0140
	(0,0132)	(0,0060)	(0,0029)	(0,0006)	(0,0003)
	$\overline{C_{max}}$				
100	0,9923	0,9623	0,9174	0,6100	0,5032
	(0,0240)	(0,0354)	(0,0378)	(0,0225)	(0,0117)
1 000	0,9982	0,9944	0,9867	0,9202	0,8125
	(0,0119)	(0,0095)	(0,0120)	(0,0141)	(0,0193)
5 000	0,9994	0,9983	0,9962	0,9793	0,9547
	(0,0061)	(0,0053)	(0,0050)	(0,0053)	(0,0060)
10 000	0,9998	0,9990	0,9977	0,9881	0,9745
	(0,0006)	(0,0030)	(0,0039)	(0,0036)	(0,0038)

* Ezer ismétlés esetén, $m=0$, $s=1$ paraméterű lognormális eloszlással generált idősorokkal.

*

Általánosságban elmondható tehát, hogy a komponensszám növelése adott elemszám mellett rontja a szeparáció hatékonyságát. A hatékonyság azonban különösen alacsonyabb elemszám és nagyobb komponensszám esetén függ az eloszlástól. Míg lognormális és *t*-eloszlású generált adatoknál hasonló, addig exponenciális eloszlásból vett minták esetén alacsonyabb a maximális korreláció és magasabb a keresztkorreláció értéke. Az elemszámok jelentősebb növelésekor azonban ez a különbség lényegesen csökken.

Ugyanakkor megfigyelhető, hogy az átlagos keresztkorrelációk értékére – adott elemszám mellett – kevésbé van hatással a komponensszám növelése, mint az átlagos maximális korrelációk értékére. Ennek oka, hogy a FastICA-algoritmusnak szüksége van a várható érték becslésére, így az eredeti komponensek visszaállításához

növekvő komponensszám esetén nagyobb elemszámra van szükség. Az alacsony keresztkorrelációs értékek azt mutatják, hogy a visszaállított és az eredeti jelek közötti lineáris összefüggőség alacsony marad, miközben a maximális korrelációk értékében bekövetkező csökkenés arra utal, hogy bár a szeparáció jó, csak épp nem az eredeti komponenseket kapjuk vissza.

Az eloszlásbeli különbségek vizsgálatához más paraméterű *t*- és exponenciális eloszlású változókra is elvégeztük a teszteket. Az elméleti várakozásokkal összhangban azt találtuk, hogy az exponenciális eloszlásból vett mintákat tekintve a korrelációs mérőszámok nem mutatnak összefüggést a paraméterrel. Ez várható is volt, hiszen az eloszlás csúcsossága és ferdesége konstans. Ezzel ellentétben a *t*-eloszlású generált elemek esetén a keresztkorrelációs értékek a szabadságfok növelésével nőnek, míg a maximális korreláció értékei csökkennek. Ez ismét csak érthető, hiszen a szabadságfok növelésével az eloszlás a normális eloszláshoz tart, így a szeparáció minősége romlik (lásd az 1.4. pontot).

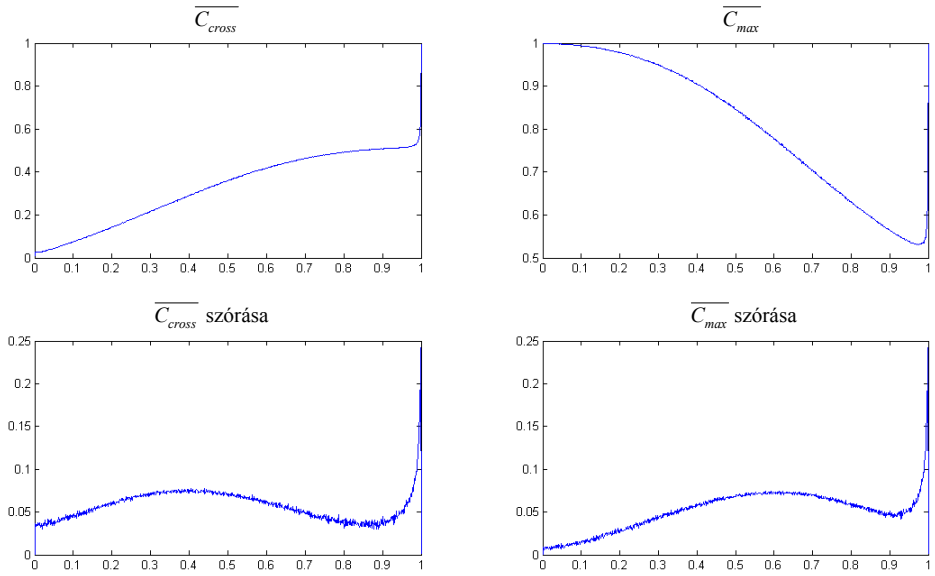
2.4.3. A komponensek összefüggőségének hatása

A generált komponenseket eddig úgy választottuk meg, hogy azok függetlenek legyenek egymástól. Ebben a részben azt vizsgáljuk, hogy milyen hatása van a komponensek közötti különböző mértékű lineáris függőségnek, azaz az ICA alapfeltevését sértő feltételeknek. A vizsgálathoz generáltunk két 1 szabadságfokú *t*-eloszlásból származó véletlen vektort, melyek között a korrelációs együtthatót 0–1 között változtattuk. Minden korrelációs együttható esetén 5 000 szimuláció átlagát vizsgáltuk, ahol az egyes idősorok 1 000 elem hosszúságúak voltak. Az eredményeket a 6. ábra szemlélteti. A bal oldali képen az átlagos keresztkorreláció, alatta annak szórása, míg a jobb oldalon az átlagos maximális korreláció és a szórások értékei figyelhetők meg a korreláció függvényében.

Az átlagos keresztkorrelációs értékek az összefüggőség növelésével egy ideig lineárisan nőnek, 0,7 körüli korreláció után a keresztkorrelációs értékre gyakorolt hatás csökken. A maximális korrelációk értékei parabola jelleget mutatnak, vagyis az összefüggés mértékének növelése egy idő után nagyobb hatással van a korrelációs értékek alakulására, mint a kisebb mértékű összefüggőség esetén. Ez a parabola jelleg a korreláció növekedésével eltűnik, és lineárisan csökken a 0,52-es korrelációs szintig.

1-hez közeli korreláció esetén mind a kereszt, mind a maximális korrelációs értékek az egyhez tartanak. Ennek oka, hogy ilyen mértékű összefüggőség esetén az adathalmaz már egy dimenzióban is ábrázolható. Az előzetes PCA-val való komponensszám-meghatározás tehát egy komponenst talál, így a kereszt- és a maximális korrelációs értékek megegyeznek, valamint mindkét mutató értéke egy. Ezzel egyidejűleg a szórások értékei nullára csökkennek.

6. ábra. Átlagos keresztkorreláció és maximális korreláció az idősorok közötti összefüggés függvényében



A kereszt- és maximális korrelációs értékek alakulásának magyarázata – vagyis, hogy miért 0,5 körüli szintig nőnek, illetve csökkennek – az, hogy a generált idősorok közötti lineáris összefüggőség növelésével egyre inkább ugyanazon értékek jelennek meg a második idősor adatai között, mint az első idősorban. A visszaállított komponensek viszont függetlenek lesznek, így az eredeti jelek és a visszaállított komponensek korrelációs mátrixában az egyik komponens és a generált idősorok közötti lineáris kapcsolat magas, míg a másik komponenst tekintve a korrelációs értékek elhanyagolható nagyságúak lesznek, tehát két komponens és két idősor esetén mind a maximális korreláció minimum értéke, mind a keresztkorreláció maximum értéke 0,5 lesz. Amint azonban az összefüggőség az 1-hez közeli tartományba esik, csak egy komponenst kapunk vissza, amely esetben pedig az algoritmus 1 értékű korrelációt ad vissza az eredeti idősorokkal.

3. Alkalmazási példák

Ebben a fejezetben áttekintjük az ICA pár jellemző (és néhány innovatív) alkalmazási területét. Egyaránt hozunk példákat a – klasszikusabb – mérnöki és az újabb közgazdasági alkalmazásokra.

3.1. Mérnöki alkalmazások

Számos olyan probléma merül fel a mérnöki gyakorlatban, melynek során az adatok függetlenítésére, illetve a különböző komponensek elkülönítésére van szükség. Ennek megfelelően az ICA az évek során számos esetben bizonyította alkalmazhatóságát. A téma bőséges irodalma is a módszer elterjedtségére utal.

Legjellemzőbb felhasználási módjai a különféle dekompozíciós feladatok. Ezek közül is a legegyszerűbb eset az, ha az /1/ egyenlet változóinak száma egy, az *a* koefficiens pedig a jelekhez hasonlóan időfüggő. Ez az ún. vak dekonvolúciós probléma (blind deconvolution – BD) esete, amikor is a megfigyelt jel az eredeti jel valamilyen szűrt változata. Az eredeti jel visszaállítása például a *Kaplan* és *Ulrych* [2003] által leírt módszerrel lehetséges.

A többdimenziós dekompozíciós feladatok nagy része a korábban említett BSS-probléma megoldását igényli, mivel számos gyakorlati kérdésfeltevés megfogalmazható formálisan az /3/ egyenlet szerinti klasszikus alakban. A BSS-probléma tipikus példája – a koktélparti probléma mintájára – a hang-, beszédfelismerés, illetve a több mikrofon által rögzített jelek szétválasztása későbbi feldolgozás céljából. A probléma megoldására számos gyakorlati alkalmazás született, többek között olyan is, mely több független komponens szétválasztására is képes, mint ahány megfigyelt jel rendelkezésre áll (*Lee et al*. [1999]).

További alkalmazási lehetőség bármilyen többdimenziós mért adat dekompozíciója. *Liu et al.* [2007] például agyi fMRI-felvételek[13] feldolgozására használták a módszert. A szerzők a felvételeken tapasztalható elváltozások és az egypontos nukleotid-polimorfizmus (single nucleotide polymorphism – SNP) kapcsolatát vizsgálták. Az SNP a DNS egy bázispárjának mutációja, mely a populáció legalább 1 százalékában azonos helyen jelenik meg. Az SNP-k az emberi genetikus variációk nagyjából 90 százalékát teszik ki, hatásuk feltárása a genetika egyik intenzíven kutatott területe (Genomic Science Program 2013). A vizsgálathoz a mért adatok dekompozícióját egy ún. párhuzamos ICA- (parallel ICA-, pICA-) architektúrával végezték, mely a független komponensek meghatározásával párhuzamosan képes az egyes modalitások – azaz az fMRI- és az SNP-adatok – közötti kapcsolatok elemzésére. Ezáltal a módszer betekintést nyújt a különböző agyi régiók aktivitása és a genotípus lehetséges kölcsönhatásaiba.

A BSS-problémákon túl az ICA használható például zajszűrésre is (*Gruber et al*. [2004]), mivel a módszer a különböző mért jelek és a zaj elkülönítésében igen robusztus eszköznek bizonyult. További, nem triviális alkalmazása lehet a módszernek

[13] A funkcionális mágneses rezonanciás képalkotás (functional magnetic resonance imaging) olyan orvosi képalkotási eljárás, mely képes a test valamely szövetének, jellemzően az agyi régiók vérfelhasználásának változásán keresztül ábrázolni annak aktivitását.

különböző digitális képek vízjelezése is. Digitális csatornák védelmére gyakran alkalmaznak ún. vízjelezést, például dokumentumok hitelesítése esetén. Ekkor az adatokhoz valamilyen érzékeny információt adnak hozzá, mely az adatok sérülése esetén – például jogtalan módosítás miatt – jelzi a sérülés jelenlétét a fogadó fél számára. Lehetséges a vízjel hozzáadása az adatok független komponenseihez is (*Bounkong et al*. [2003]), mely további védelmet nyújt azok sérülése ellen.

3.2. Pénzügyi alkalmazások

A módszer népszerűsége a gazdasági alkalmazások területén is egyre nő, hiszen ezen adatok elemzésekor is sok, a felsoroltakhoz hasonló jellegű probléma merül fel. Ahogy említettük, az ICA a PCA-hoz hasonlóan keresztmetszeti adatok feldolgozására is alkalmas, a közgazdaságtan területén azonban a módszert elsősorban pénzügyi adatok és modellek vizsgálatára alkalmazzák. A módszer különösen nagyfrekvenciás – napi vagy napon belüli adatokat tartalmazó – idősorok esetén alkalmazható jó eredménnyel. Más, például makroidősorok esetén, ahol jellemzően havi, negyedéves vagy éves adatok állnak rendelkezésre, kevésbé hatékony, azok alacsonyabb száma miatt. A magyar gazdasági szakirodalom ennek ellenére eddig nem foglalkozott a módszer használatával. A következőkben ismertetett, nemzetközi irodalomban fellelt cikkek alapvetően a PCA és az ICA összehasonlítására helyezik a hangsúlyt különböző pénzügyi modellek vizsgálata során.

Az ICA-t pénzügyi adatok vizsgálatával kapcsolatban alapvetően három nagyobb területen használják. Egyrészt pénzügyi modellek, módszerek vizsgálatára, másrészt pénzügyi ökonometriai modellek alkalmazásában, harmadrészt pedig egyéb adatelemzési eszközök tesztelése esetén. A következőkben rövid áttekintést adunk a pénzügyi alkalmazásokkal kapcsolatos szakirodalomról.

A portfoliókezelés során a szakemberek egyik feladata a kamatláb-érzékenység kiküszöbölése, amit az immunizációs stratégia segítségével érhetnek el. Ehhez az ún. átlagidőt veszik alapul, amely az értékpapírok kamatláb-érzékenységének egy mértéke. A portfoliót immunizáltnak tekintjük, ha az átlagideje nulla, vagyis a kamatláb kicsiny változása nem befolyásolja a portfolió értékét. Ahhoz azonban, hogy az érzékenységet kiküszöböljék, a gyakorlatban nagyszámú kötvény átlagidejét kellene kiszámítani, tehát az immunizáció hatékonysága és a számításigény között erős tradeoff van. A számításigény csökkentésére általában főkomponenseket használnak, amelyek egy adott elemű kötvényportfolió varianciájának a lehető legnagyobb hányadát magyarázzák, és ezek segítségével „állítják be" az átlagidőt nullára. *Gonzalez* és *Nave* [2010] cikkükben PCA és ICA megközelítést használva elemezték, hogy melyik módszer szerint alakítható ki megbízhatóbb immunizációs stratégia. A tesztek során két különböző periódust vizsgáltak, és mindkettőben az ICA teljesített jobban. Hasonló eredményekre jutott *Bellini* és *Salinelli* [2003] is.

Egy másik gyakran használt pénzügyi modell az arbitrázs értékelési elmélet (arbitrage pricing theory – APT) (lásd *Ross* [1976]), amely a faktormodellek körébe sorolható. A modell segítségével lehetőség van egy adott értékpapír hozamának faktorokra való felbontására. *Cha* és *Chan* [2000] cikkükben részvényhozamokra alkalmazták a független komponens analízist, hogy felállítsanak egy többfaktoros modellt. Szintén részvényhozamok dekompozíciójára használta az ICA-t *Back* és *Weigend* [1997], akik elemzésük során összehasonlították eredményeiket a PCA által szolgáltatott eredményekkel. Következtetésük szerint a független komponensek jobban visszaadják a hozamok mögött rejlő látens változókat, mint a főkomponensek, sőt a sokkokra sokkal robusztusabban reagálnak az ICA által generált komponensek.

A pénzügyi ökonometriában gyakran használt GARCH-modellek (generalized autoregressive conditional heteroskedasticity) (lásd *Bollerslev* [1986]) a volatilitás folyamatát modellezik. *Wu* és *Yu* [2005] GARCH-modellekkel kötötte össze az ICA használatát. A szerzők azt vizsgálták, hogy a hozamokra illesztett VAR-modellek reziduumainak PCA és ICA által dekomponált változókra illesztett volatilitásmodellek közül melyik illeszkedik jobban az idősorokra. Eredményeik szerint az ICA-GARCH kevésbé volatilis reziduumokat eredményezett, mint a PCA-GARCH, illetve a független komponenseken alapuló modellek pontosabban követték az idősort, mint a főkomponenseken alapulók.

Szintén ezen a területen, de az előrejelzések megbízhatóságát vizsgálta *Lu* [2009] a Nikkei225- és a TAIEX-indexek részvényeinek hozamaira. A vizsgálat során a független komponens elemzést a zaj idősorból való kiszűrésére használták. A korábbiakhoz hasonlóan a PCA alapú megközelítéssel vetették össze a módszert. A zaj kiszűrése után a komponensekre szupport vektor regressziót (support vector regression – SVR) alkalmaztak az előrejelző modell becsléséhez. A tesztek során több mutatót is vizsgáltak az előrejelzések megbízhatóságának mérésére. Az előrejelzési hibát (mean absolute deviation – MAD), RMSE (root mean squared error) és NMSE (normalized mean squared error), míg az előrejelzés adekvátságát DS (directional symmetry), CP (correct up trend) és CD (correct down trend) mutatókkal mérték. Eredményeik alapján minden mutató esetén a többinél jobb teljesítményt nyújtott az ICA-SVR módszer.

Az ICA-t pénzügyi területen előrejelzésre is gyakran alkalmazzák. *Liu* és *Wang* [2011] az ICA és a PCA által előállított zajtól mentesített komponenseket használva készítettek többrétegű perceptron (multi-layer perceptron – MLP) neurális hálót a részvényárfolyamok előrejelzéséhez. Az MLP-t hiba-visszaterjesztéses algoritmussal (backpropagation – BP) tanították. Az előrejelzési hiba becslésére MAE (mean absolute error) és RMSE mutatókat számítottak, mindkét mutató esetén az ICA-BP modell teljesített jobban.

Látható tehát, hogy az ICA több tudományterületet átölelő, sokrétű alkalmazási lehetőségeket rejt magában. Az ICA legfőbb előnye a PCA-val szemben, hogy az

adatok mögötti rejtett függőségi viszonyokra világít rá, kiemelve ezzel az adatok és azok változása mögötti struktúrát. A leghatékonyabban viszont a két módszer együttesen használható oly módon, hogy mindkét alkalmazás a másik egy kevésbé előnyös tulajdonságát kompenzálja: az ICA a PCA-nál jobb információkiemelést végez, míg a PCA rávilágít az adatok dimenzionalitására, melyről az ICA nem ad információt.

Irodalom

BACK, A. D. – ANDREAS, S. W. [1997]: A First Application of Independent Component Analysis to Extracting Structure from Stock Returns. *International Journal of Neural Systems*. Vol. 8. No. 4. pp. 473–484.

BARBAKH, W. A. – YING, W. – FYFE, C. [2009]: *Non-Standard Parameter Adaptation for Exploratory Data Analysis.* Springer. Berlin.

BELLINI, F. – SALINELLI, E. [2003]: Independent Component Analysis and Immunization: An Explanatory Study. *International Journal of Theoretical and Applied Finance*. Vol. 6. No. 7. pp. 721–738.

BILLINGSLEY, P. [1995]. *Probability and Measure*. John Wiley & Sons. New York.

BOLLERSLEV, T. [1986]: Generalized Autoregressive Conditional Heteroskedastisity. *Journal of Econometrics*. Vol. 31. No. 3. pp. 307–327.

BOUNKONG, S. – TOCH, B. –SAAD, D. – LOWE, D. [2004]: ICA for Watermarking Digital Images. *Journal of Machine Learning Research*. Vol. 4. No. 7-8. pp. 1471–1498.

CHA, S.-M. – CHAN, L.-W. [2000]: *Applying Independent Component Analysis to Factor Model in Finance*. Springer. Berlin.

CHIU, C.-C. – LEE, T.-S. – LU, C.-J. [2009]: Financial Time Series Forecasting Using Independent Component Analysis and Support Vector Regression. *Elsevier, Decision Support Systems*. Vol. 47. No. 2. pp. 115–125.

COMON, P. [1994]: Independent Component Analysis, a New Concept? *Signal Processing*. Vol. 36. No. 3. pp. 287–314.

GENOMIC SCIENCE PROGRAM [2013]: *HumanGenome Project Information*. http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

GONZALEZ, M. – NAVE, J. M. [2010]: Portfolio Immunization Using Independent Component Analysis. *Revista De Economica Financeria*. Vol. 21. pp. 37–46.

GRUBER, P. – THEIS, F. J. – STADLTHANNER, K. – LANG, E. W. – TOME, A. M. – TEIXEIRA, A. R. [2004]: Denoising Using Local ICA and Kernel-PCA. *IEEE International Joint Conference on Neural Networks*. Vol. 3. No. 3. pp. 2071–2076.

HORVÁTH G. (szerk.) [2006]: *Neurális hálózatok*. Panem. Budapest.

HYVÄRINEN, A. – KARHUNEN, J. – OJA, E. [2001]: *Independent Component Analysis*. John Wiley & Sons. New York.

HYVÄRINEN, A. – OJA, E. [2000]: Independent Component Analysis: Algorithms and Applications. *Neural Networks.* Vol. 13. No. 4–5. pp. 411–430.

JOLLIFFE, I. T. [2010]: *Principal Component Analysis*. Springer. New York.

KAPLAN, S. T. – ULRYCH, T. J. [2003]: *Blind Deconvolution and ICA with a Banded MixingMatrix*. 4th International Symposium on Independent Component Analysis and Blind Signal Separation. pp. 223–228.

KENDALL, S. M. – STUART, A. – ORD, J. K. [1983]: The Advanced Theory of Statistics. Griffin. London.

LEE, T.-W. – LEWICKI, M. S. – GIROLAMI, M. – SENOWSKI, T. J. [1999]: Blind Source Separation of More Sources than Mixtures Using Overcomplete Representations. *IEEE Signal Processing Letters*. Vol. 6. No. 4. pp. 87–90.

LI, H. – ADALI, T. [2008]: A Class of Complex ICA Algorithms Based on the Kurtosis Cost Function. *IEEE Transactions on Neural Networks*. Vol. 19. No. 3. pp. 408 –420.

LIU, H. – WANG, J. [2011]: Integrating Independent Component Analysis and Principal Component Analysis with Neural Networks to Predict Chinese Stock Market. *Mathematical Problems in Financial Engineering*. pp. 1–15.

LIU, J. – PEARLSON, G. – WINDEMUTH, A. – RUANO, G. – PERRONE-BIZZOZENO, N. I. – CALHOUN, V. [2007]: Combining fMRI and SNP Data to Investigate Connections Between Brain Function and Genetics Using Parallel ICA. *Human Brain Mapping*. Vol. 30. No. 1. pp. 241–255.

PARK, Y. S. – BERA, K. A. [2009]: Maximum Entropy Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*. Vol. 150. No. 2. pp. 219–230.

PRASAD, R. – SARUWATARI, H. – SHIKANO, K. [2005]: Blind Separation of Speech by Fixed-Point ICA with Source Adaptive Negentropy Approximation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences.* Vol. E88-A. No. 7. pp. 1683–1692.

PRASZOLOV, V. V. [2005]: *Lineáris algebra*. Typotex. Budapest.

RÉNYI A. [1973]: *Valószínűségszámítás*. Tankönyvkiadó. Budapest.

ROSS, S. [1976]: The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*. Vol. 13. No. 3. pp. 341–360.

SCOTT, D. W. [1992]: *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley-Interscience. New York.

TERRELL, G. R. – SCOTT, D. W. [1992]: Variable Kernel Density Estimation. *The Annals of Statistics*. Vol. 20. No. 3. pp. 1236–1265.

WU, E. H. C. – YU, P. L. H. [2005]: *Volatility Modelling of Multivariate Financial Time Series*. Springer. Berlin.

Summary

In this study we introduce the theoretical background and empirical analysis of the independent components analysis (ICA), a method that is increasingly popular in terms of economic data analysis. It is capable to decompose correlating data to independent components, which are as independent from each other as possible, and from the linear combination of which the original data is expressible. Thus the method provides an opportunity to distinguish the hidden factors responsible for the dynamics of the data. After reviewing the theoretical background, we compare the ICA to the more commonly used principal component analysis (PCA), after that we study the properties of the ICA in detail, along the following dimensions: number and dimensionality of the data, and their dependency relations. In the end, we introduce a few of the method's application possibilities.