

OTDK dolgozat

Kapelner Tamás

Madarász László

2013

Független komponens analízis és empirikus tesztjei kötvényhozamok
felhasználásával

Independent component analysis and its empirical tests using bond returns

Kézirat lezárása: 2012 március 26.

Tartalomjegyzék

1. Bevezetés (Kapelner Tamás)	1
2. A független komponens analízis (Kapelner Tamás)	3
2.1. A kocktélparti-probléma	3
2.2. A függetlenség eldöntésének lehetőségei	4
2.2.1. Csúcsosság	5
2.2.2. Negentrópia	6
2.2.3. Maximum likelihood becslés	6
2.2.4. Kölcsönös információ	8
2.2.5. Kumuláns tenzor	8
2.3. ICA és PCA	9
2.4. ICA megvalósítása nemlineáris PCA-hálóval	12
2.5. FastICA algoritmus	14
2.6. Alkalmazási példák	15
2.6.1. Mérnöki alkalmazások	15
2.6.2. Pénzügyi alkalmazások	18
3. ICA alkalmazása kötvényhozamok dekompozíciójára (Madarász László)	21
3.1. Az empirikus tesztek menete	22
3.2. Az empirikus tesztek kivitelezése	23
3.2.1. Adatpótlási módszerek hatása	23
3.2.2. Komponensek számának vizsgálata	28
3.2.3. Adatrobosztusság vizsgálata	30
3.2.4. Elemszám-robosztusság vizsgálata	34
3.2.5. Pénzügyi sokkok hatásának vizsgálata	37
3.3. Eredmények összefoglalása	39
4. Összefoglalás (Madarász László)	41
Függelék	46
A. Transzformáció sűrűsége	46
B. Kumulánsok	48
C. Ábrák és táblázatok	49
C.1. Adatpótlási módszerek és leíró statisztikáik	49
C.2. Adatpótlási módszerek hatása az ICA-ra	52
C.3. Különböző adatpótlási módszerekkel kapott komponensek összehason- lítása	54

C.4.	Mesterséges sokkok vizsgálata	55
C.5.	Pénzügyi sokkok vizsgálata	60

Ábrák jegyzéke

2.1. A főkomponens analízis szemléltetése	10
2.2. A nemlineáris főkomponens analízis szemléltetése	11
2.3. Az ICA-t megvalósító neurális hálózat	13
2.4. A megvalósított ICA algoritmus eredményei két jel esetén	16
2.5. A megvalósított ICA algoritmus eredményei négy jel esetén	17
3.1. A hozamokat meghatározó komponensek	21
3.2. FastICA eredményei a teljes idősorra	24
3.3. Átlagos keresztkorreláció alakulása extrém sokkolás esetén	35
3.4. A neurális ICA elemszám-robosztusságának vizsgálata	36
3.5. A FastICA elemszám-robosztusságának vizsgálata	37
C.1. Konstans módszerrel pótoltt kötvényhozamok idősorai	49
C.2. Lineárisan interpolált kötvényhozamok idősorai	50
C.3. GARCH módszerrel pótoltt kötvényhozamok idősorai	51
C.4. Sokkolások szisztematikus elemkiválasztással	57
C.5. Sokkolások véletlen elemkiválasztással	58
C.6. Átlagos abszolút eltérés alakulása konstans sokkolás esetén	59
C.7. Orosz válság idősorai	60
C.8. Orosz válság komponensei FastICA-val	61
C.9. Orosz válság komponensei neurális ICA-val	61
C.10. Dot-com válság idősorai	62
C.11. Dot-com válság komponensei FastICA-val	62
C.12. Dot-com válság komponensei neurális ICA-val	63
C.13. Subprime válság idősorai	63
C.14. Subprime válság komponensei FastICA-val	64
C.15. Subprime válság komponensei neurális ICA-val	64

Táblázatok jegyzéke

3.1. Leíró statisztikák a hiányzó adatok nélküli idősorokra	23
3.2. Komponensek átlagos abszolút keresztkorrelációi öt komponens esetén	25
3.3. Különböző imputálási eljárások komponenseinek átlagos eltérései	27
3.4. Neurális ICA komponensek átlagos keresztkorrelációi a komponensszám függvényében	28
3.5. MAPE értékek különböző komponensszámok esetén	29
3.6. Komponensek korrelációi az idősorok számának függvényében	30
3.7. Idősorok komponenseinek robusztussága elemek véletlen és szisztematikus kiválasztásával	32
3.8. Sokkolás utáni komponensek korrelációs mátrixának átlagos legnagyobb elemei	34
C.1. Leíró statisztikák a konstans módszerrel pótoltt idősorokra	49
C.2. Leíró statisztikák a lineárisan interpolált idősorokra	50
C.3. Leíró statisztikák a GARCH-modellel pótoltt idősorokra	51
C.4. Konstans módszerrel pótoltt idősorok korrelációs mátrixa	52
C.5. Konstans módszerrel pótoltt idősorok komponenseinek korrelációs mátrixa . . .	52
C.6. Lineárisan interpolált idősorok komponenseinek korrelációs mátrixa	53
C.7. GARCH-modellel pótoltt idősorok komponenseinek korrelációs mátrixa	53
C.8. GARCH és Lineáris idősorok komponenseinek korrelációs mátrixa	54
C.9. GARCH és Konstans idősorok komponenseinek korrelációs mátrixa	54
C.10. Lineáris és Konstans idősorok komponenseinek korrelációs mátrixa	54
C.11. Idősorok komponenseinek robusztussága elemek véletlen és szisztematikus kiválasztásával	55
C.12. Idősorok komponenseinek robusztussága elemek véletlen és szisztematikus kiválasztásával	55
C.13. Idősorok komponenseinek robusztussága elemek véletlen és szisztematikus kiválasztásával	55
C.14. Elemenkénti átlagos abszolút eltérések konstans sokkolás esetén	56
C.15. Elemenkénti átlagos abszolút eltérések azonos véletlen sokkolás esetén	56
C.16. Elemenkénti átlagos abszolút eltérések elemenkénti véletlen sokkolás esetén . .	56

1. Bevezetés

A kötvényhozamok elemzése mind pénzügyi, mind gazdasági szempontból kiemelt jelentőségű. Mind pénzügyi árazáshoz, mind monetáris policy-k meghatározásához szükséges a hozamok változására ható tényezők azonosítása. Monostori (2010) dolgozatában számolt be arról, hogy melyek azok a főbb változók, amelyek befolyásolják a kötvényhozamok alakulását, azonban empirikus eredmények hiányában hipotézise nem támasztható alá teljes bizonyossággal. A szerző a magyar kötvényhozamokat úgy tekintette, mint különböző kockázati tényezők miatti felárak összegét. Kiinduló hipotézisünk, hogy a különböző országok kötvényhozamaira ugyanazok a látens tényezők hatnak, azonban eltérő mértékben.

Felmerül tehát a kérdés, hogy van-e olyan módszer, amivel lehetséges a hozamok látens komponensekre való dekompozíciója, illetve ha van, milyen jellemzői, határai vannak ennek a módszernek.

Ehhez hasonló problémák megoldására a mérnöki gyakorlatban már évek óta elterjedt módszer az ún. 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 felbontása. Mivel a fentiekben felvázolt probléma esetén joggal feltételezhetjük, hogy a kötvényhozamokban jelen lévő különböző rejtett komponensek bár nem függetlenek, de függetleníthetők egymástól, az ICA ideális megoldásnak tűnik a fenti probléma megoldására.

A szakirodalomban számos ilyen megoldás fellelhető, magyar vonatkozásban azonban nem találtunk releváns eredményt. Ennek megfelelően a dolgozatban az ICA általános bemutatására, valamint pénzügyi alkalmazhatóságának vizsgálatára helyezzük a hangsúlyt.

A dolgozatban először az ICA általános módszertanát mutatjuk be, megvalósítási lehetőségeit, illetve hasonlóságát a pénzügyi területen inkább elterjedt főkomponens analízishez. Ezután ismertetjük egy általunk megvalósított neurális hálózat alapú ICA architektúra szerkezetét és működését, illetve az ICA általános felhasználási lehetőségeit mérnöki és pénzügyi területen egyaránt.

Ezután a módszer tulajdonságainak részletes vizsgálatát végezzük el kötvényhozamok dekompozíciójára alkalmazva azt. Az empirikus vizsgálat során 5 ország kötvényhozamait elemezzük, ismertetjük az idősorok és független komponenseik általános tulajdonságait, illetve az idősorok hiányzó adatainak pótlásának hatását az ICA algoritmus eredményére. Megvizsgáljuk továbbá, hogy a komponensek számának megválasztása milyen hatással jár.

A további tesztek a módszer robusztusságát vizsgálják, mind adatrobusztusság, mind elemszám-

robusztusság tekintetében. A tesztek kivitelezéséhez a teljesítmény megítélésére alkalmas mérőszámokat definiálunk, majd nagyszámú tesztet végzünk az általános viselkedés feltárása céljából.

Végül az idősorok ablakolásának hatását vizsgáljuk, különböző válságok környezetében ablakolva azokat. A komponensek pénzügyi jelentéstartalommal való felruházása jelen dolgozatnak nem célja, de pozitív eredmények esetén ennek megvalósítása egy további kutatási lehetőség.

2. A független komponens analízis

2.1. A koktélparti-probléma

Az ICA alkalmazásának leggyakoribb példája a vak forrásszétválasztás (blind source separation, BSS). Ebben az esetben a feladat több, rendelkezésre álló időfüggvény független komponenseinek meghatározása, pusztán az időfüggvényekből nyerhető adatok alapján.

Egy szemléletes példa erre egy 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 csak valamilyen zajt.

Formálisan megfogalmazva ugyanezt a problémát: keresett N darab statisztikailag teljesen független valószínűségi változó, $S_1, S_2, \dots, S_N$, melyek a koktélparti-probléma beszélőit reprezentálják. Legyen adott – az egyszerűség kedvéért – szintén N darab megfigyelt valószínűségi változó: $X_1, X_2, \dots, X_N$, amelyek a mikrofonok által felvett jeleket jelentik. Feltételezésünk a következő:

$$X_i = \sum_{j=1}^N a_{ij} S_j, \quad i = 1, 2, \dots, N, \quad (2.1)$$

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

$$\underline{X} = \underline{A} \underline{S}, \quad (2.2)$$

ahol $\underline{X} = [X_1, X_2, \dots, X_N]^T$, $\underline{S} = [S_1, S_2, \dots, S_N]^T$, $\underline{A}$ pedig az ún. keverőmátrix, elemei az a_{ij} együtthatók, 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 nonsinguláris $\underline{A}$ esetén) a

$$\underline{S} = \underline{A}^{-1} \underline{X} \quad (2.3)$$

probléma megoldását jelenti. További megfontolásokat igényel, ha a korábbi feltételezéseink nem teljesülnek, azaz ha például $\mathbf{A}$ szinguláris, különböző számú megfigyelt jelünk és komponensünk van, kapcsolatuk nemlineáris, a méréseket zaj terheli, stb. Az esetek ismertetése meghaladná a dolgozat kereteit, következményeiket a szakirodalom bőségesen tárgyalja (Hyvärinen, Karhunen & Oja 2001).

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

2.2. A függetlenség eldöntésének lehetőségei

A koktélparti probléma megoldására törekedve felmerülhet a kérdés, hogy megoldja-e a problémát az $\underline{X}$ megfigyelések fehéritése, azaz ha úgy transzformáljuk őket, hogy várható értékük egy, korrelációs mátrixuk pedig egységnyi legyen.

Az így előállt $\underline{V}$ fehéritett változók korrelálatlanok lesznek ugyan, de nem feltétlenül lesznek függetlenek. Mivel $\underline{V}$ bármely ortogonális transzformációja szintén fehér¹, ezért csupán ez a feltétel nem elegendő annak eldöntésére, hogy $\underline{V}$ a valódi független komponenseket tartalmazza-e.

Másrészt viszont látható, hogy érdemes fehéritett adatokat használni az ICA számításakor, hiszen az $\underline{S}$ független komponensek korrelációs mátrixa biztosan az egységmátrix, ami azt jelenti, hogy a fehéritett adatokra felírt keverőmátrix ortogonális lesz². Ha tehát fehéritett adatokon dolgozunk, a (2.2) egyenlet a következő formát ölti:

$$\underline{V} = \mathbf{V}\underline{X} = \mathbf{V}\mathbf{A}\underline{S} = \mathbf{W}^T \underline{S}, \quad (2.4)$$

ahol a fehéritett adatokat a $\mathbf{V}$ mátrix által reprezentált lineáris transzformációval állítjuk elő, a $\mathbf{W}^T$ ortogonális mátrix pedig az adatokat fehéritő keverőmátrix, mely egyben az ún. szétválasztómátrix inverze – azaz ortogonális mátrix esetén a transzponáltja. Amint azt a 2.4. pontban is látni fogjuk, a szétválasztómátrix a független komponensek elkülönítésére alkalmas, ahogy az a (2.4) egyenletből is látható.

A probléma megoldásához tehát más úton kell elindulnunk. Alakítsuk át a (2.2) egyenletet a következőképpen:

$$\underline{Y} = \mathbf{b}^T \underline{X} = \mathbf{b}^T \mathbf{A} \underline{S} = \mathbf{q}^T \underline{S}, \quad (2.5)$$

¹ $\mathbf{z} = \mathbf{U}\mathbf{v}$ esetén $\mathbb{E}(\mathbf{z}\mathbf{z}^T) = \mathbb{E}(\mathbf{U}\mathbf{v}\mathbf{v}^T\mathbf{U}^T) = \mathbf{U}\mathbf{I}\mathbf{U}^T = \mathbf{I}$, ha $\mathbf{U}$ ortogonális, és $\mathbf{v}$ fehér.

² $\mathbf{v} = \mathbf{W}^T \mathbf{s}$ esetén $\mathbf{I} = \mathbb{E}(\mathbf{v}\mathbf{v}^T) = \mathbb{E}(\mathbf{W}^T \mathbf{s}\mathbf{s}^T \mathbf{W}) = \mathbf{W}^T \mathbf{W}$.

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

A centrális határeloszlás tétel szerint azonos eloszlású, egymástól teljesen független valószínűségi változók összege normális eloszláshoz tart. Ez tulajdonképpen azt jelenti, hogy ha független valószínűségi változók egy összegéhez egy olyan valószínűségi változót adunk, mely az előbbiektől független, akkor az így nyert összeg inkább fog hasonlítani egy normális eloszlásra, mint amennyire az összegzés előtt hasonlított. Eszerint ha a $\mathbf{q}^T \underline{S}$ lineáris kombinációban $\mathbf{q}$ elemeit (legyenek ezek a súlyok) megváltoztatjuk, és az S -ek függetlenek, akkor az összeg annál kevésbé fog hasonlítani a normális eloszlásra, minél inkább csak egyetlen S_i határozza meg az összeg értékét. Ha ugyanis az összeg kevésbé hasonlít a normális eloszlásra, az csak azért lehet, mert az Y összegben kevesebb S_i konvolválódik. Célunk pedig éppen az, hogy Y minél inkább hasonlítson az egyik S_i komponensre, azaz hogy $\mathbf{q}$ változásának hatására – melyet nyilván $\mathbf{b}$ változtatásával érhetünk el – minél kevésbé legyen normális eloszlású.

A fenti gondolatmenet fontos következménye, hogy az S_i -k között legfeljebb egy lehet normális eloszlású, hiszen két normális eloszlású komponenszt ezzel a módszerrel nem tudunk megkülönböztetni. Hogy maximalizáljuk $\mathbf{b}^T \underline{X}$ nemnormalitását, egy olyan mérőszámra van szükség, mely információt ad arról, hogy egy valószínűségi változó eloszlása mennyire hasonlít a normális eloszlásra.

2.2.1. Csúcsosság

Egy ilyen mérőszám a csúcsosság (kurtosis), melynek előnyös tulajdonsága, hogy normális eloszlás esetén három, annál csúcsosabb eloszlások esetén nagyobb, míg ellenkező esetben kisebb³. Egy μ várhatóértékű, σ szórású Y valószínűségi változó csúcsossága:

$$\text{kurt}(Y) = \frac{\mathbb{E}(Y - \mu)^4}{\sigma^4}. \quad (2.6)$$

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).

³Annak érdekében, hogy az összefüggés szemléletesebb legyen, szokásos az ún. excess kurtosis használata, melynek értéke a fenti definícióval $\text{kurt}(Y) - 3$. Így a normális eloszlásnál csúcsosabb eloszlások csúcsossága pozitív, míg a kevésbé csúcsosaké negatív lesz.

2.2.2. Negentrópia

A lapultsággal, mint a nemnormalitás mérőszámával kapcsolatban azt a gyakorlati megfigyelést kell azonban tennünk, hogy nagyon érzékeny az outlierekre⁴, nem robusztus mérőszáma a nemnormalitásnak. Egy hasonló, ám kedvezőbb statisztikai tulajdonságokkal rendelkező mérőszám a differenciális entrópia:

$$H(Y) = - \int p_Y(v) \log p_Y(v) dv, \quad (2.7)$$

ahol $p_Y(\cdot)$ az Y valószínűségi változó sűrűségfüggvénye. Érdemesebb azonban egy olyan entrópia alapú mérőszámot használni, melyen keresztül egy valószínűségi változó nemnormalitása közvetlenül jelenik meg. Ilyen mérőszám a negatív normalizált differenciális entrópia, azaz a negentrópia:

$$J(Y) = H(Y_{norm}) - H(Y), \quad (2.8)$$

ahol $H(Y_{norm})$ az Y -al azonos várható értékű és szórású normális eloszlás entrópiája. Ez a jellemző mindig pozitív⁵, azaz az összes eloszlás közül a normális eloszlás negentrópiája a legkisebb, a nemnormalitás mérőszámának tehát ez is kiválóan megfelel. Számítása azonban nehézkes, mert a valószínűségi sűrűségfüggvény pontos ismerete kellene hozzá. Optimalizációs algoritmus megvalósításakor emiatt Y sűrűségfüggvényének valamilyen közelítésére van szükség $J(Y)$ becsléséhez (Prasad, Saruwatari & Shikano 2005).

2.2.3. Maximum likelihood becslés

Az ICA becsléséhez használható további módszer a maximum likelihood (ML) becslés. A maximum likelihood becslés a modern statisztikai becsléselemlet egyik legfontosabb módszere. Elnagyolt megfogalmazásban annak a valószínűségét számítja ki, hogy adott sokasági eloszlást feltételezve eredményül épp az a minta adódik, ami ténylegesen adott⁶. Itt a likelihoodok kiszámítása zajmentes ICA-modellre a lineáris transzformáció sűrűségének⁷ meghatározásán alapul. Továbbra is adott a (2.2) egyenletnek megfelelő összefüggés, ahol $\mathbf{A}$ a keverőmátrixot jelenti. Ekkor a transzformált valószínűségi változó sűrűsége a következő formulával reprezentálható (Hyvärinen et al. 2001):

⁴Outlier-eknek az adatsorból szélsőségesen kiugró értékeket nevezzük.

⁵Adott szórású és várható értékű eloszlások közül mindig a normális eloszlás a legnagyobb entrópiájú.

⁶Részletesebben lásd (Hunyadi 2009), vagy (Hajdú 2003).

⁷Részletesebben lásd A. függelék.

$$p_{\underline{\mathbf{x}}}(\mathbf{x}) = \left| \frac{1}{\det \mathbf{A}} \right| p_{\underline{\mathbf{s}}}(\mathbf{A}^{-1} \mathbf{x}) = |\det \mathbf{B}| p_{\underline{\mathbf{s}}}(\mathbf{s}) = |\det \mathbf{B}| \prod_i p_i(s_i), \quad (2.9)$$

ahol $\mathbf{B} = \mathbf{A}^{-1}$ és p_i az i -edik független komponens sűrűsége. A fenti egyenlőség kifejezhető $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)^T$ és $\mathbf{x}$ függvényeként a következő egyenlőséggel:

$$p_{\underline{\mathbf{x}}}(\mathbf{x}) = |\det \mathbf{B}| \prod_i p_i(\mathbf{b}_i^T \mathbf{x}). \quad (2.10)$$

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

$$L(\mathbf{B}) = \prod_{t=1}^T \prod_{i=1}^n p_i(\mathbf{b}_i^T \mathbf{x}(t)) |\det \mathbf{B}|. \quad (2.11)$$

Algebrailag egyszerűbb azonban a log-likelihooddal számolni, ami a következő formában adott:

$$\log L(\mathbf{B}) = \sum_{t=1}^T \sum_{i=1}^n \log p_i(\mathbf{b}_i^T \mathbf{x}(t)) + T \log |\det \mathbf{B}|. \quad (2.12)$$

Az egyenlőség jobb oldalán szereplő T -szerinti summázásnál kihasználva, hogy az átlag a várható érték közelítése, a bal oldalt T -vel osztva és a jobboldalon az átlagot a várható érték operátorral helyettesítve, a következő összefüggést kapjuk.

$$\frac{1}{T} \log L(\mathbf{B}) = \mathbb{E} \left(\sum_{i=1}^n \log p_i(\mathbf{b}_i^T \mathbf{x}) \right) + \log |\det \mathbf{B}| \quad (2.13)$$

Itt a várható érték nem az elméleti várható értéket jelöli, hanem a mintából számított átlagot. A kérdés természetesen az, hogy hogyan kaphatjuk meg $\mathbf{B}$ értékeit. Ha adottak a p_i sűrűségek, akkor a fenti egyenlőségben a jobb oldalon szereplő első tag éppen akkor maximális, ha $\mathbf{y} = \mathbf{B}\mathbf{x}$ egyenlőség teljesül, ahol $\mathbf{y}$ éppen a független komponensek megfigyelt értékeit adja.

A probléma ebben az esetben azonban kettős. Egyrészt $\mathbf{B}$, másrészt a p_i sűrűségek meghatározására van szükség. Amennyiben nincs sejtésünk a komponensek sűrűségéről, akkor nemparaméteres becslésre van szükség ezek meghatározásához, melyre a korábban megjelölt források több megoldási lehetőséget is ismertetnek.

2.2.4. Kölcsönös információ

Valószínűségi változók függetlenségének egy másik kiváló mérőszáma lehet a kölcsönös információ. A (2.7) egyenlet jelöléseit használva n darab valószínűségi változó kölcsönös információja:

$$I(Y_1, Y_2, \dots, Y_n) = \sum_{i=1}^n H(Y_i) - H(\underline{Y}) \quad (2.14)$$

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

$$H(\underline{Y}) = - \int \int \cdots \int p_{\underline{Y}}(v_1, v_2, \dots, v_n) \log p_{\underline{Y}}(v_1, v_2, \dots, v_n) dv_1 dv_2 \cdots dv_n. \quad (2.15)$$

Látható, hogy ez tulajdonképpen az egyes vetületi entrópiáknak és az együttes entrópiának a különbsége, függetlenség esetén ez a mérőszám nulla.

Az A. függelékben leírtaknak megfelelően meghatározhatjuk a (2.3) egyenlettel adott transzformációval transzformált Y_i valószínűségi változók kölcsönös információját (ismét az $\mathbf{B} = \mathbf{A}^{-1}$ jelöléssel):

$$I(Y_1, Y_2, \dots, Y_n) = \sum_{i=1}^n H(\mathbf{b}_i^T \underline{X}) - H(\underline{X}) - \log |\det \mathbf{B}|. \quad (2.16)$$

Ez az egyenlet szintén használható egy optimalizációs eljárás költségfüggvényeként, amennyiben $H(Y_i) = \mathbf{b}_i^T \underline{X}$ és $H(\underline{X})$ 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 kimutatható. A részletes levezetés Hyvärinen et al. (2001) könyvében található meg.

2.2.5. Kumuláns tenzor

Ahogy azt a csúcosság esetében is láthattuk, valószínűségi változók függetlenségének vizsgálatkor a negyedrendű statisztikák nagy segítséget nyújthatnak. Nem meglepő tehát, hogy negyedrendű kumulánsok⁹ vizsgálatával a független komponensekre való felbontás szintén elvégezhető.

⁹A kumulánsok leírása a B. függelékben található.

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

$$\mathbf{F}(\mathbf{M})_{ij} = \sum_{kl}^n m_{kl} \cdot \text{cum}(S_i, S_j, S_k, S_l), \quad (2.17)$$

ahol $\mathbf{F}(\mathbf{M})_{ij}$ a tenzor általi transzformáció eredményének ij -edik eleme, m_{kl} pedig a transzformált $\mathbf{M}$ mátrix kl -edik eleme. Ez a négydimenziós tenzor nyilván szimmetrikus¹⁰, tehát diagonalizálható, azaz létezik olyan $\mathbf{M}$ sajátmátrix és λ sajátérték, hogy:

$$F(\mathbf{M}) = \lambda \mathbf{M}. \quad (2.18)$$

Megmutatható, hogy a tenzornak n nemnulla sajátértéke van, melyek éppen az S_i valószínűségi változók csúcosságaival egyenlők. Állítsunk elő az $\underline{X}$ bemeneti adatokból a $\underline{V}$ fehérített adatokat. Megmutatható az is (Hyvärinen et al. 2001), hogy ekkor az $\mathbf{M}$ sajátmátrixok mindegyike $\mathbf{M}_i = \mathbf{w}_i^T \mathbf{w}_i$ alakú, azaz a szétválasztómátrix egy $\mathbf{w}_i$ oszlopának önmagával 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 nagy-méretű tenzorok miatt sok memóriát is igényel, így jellemzően csak kisdimenziós esetekben használják.

2.3. ICA és PCA

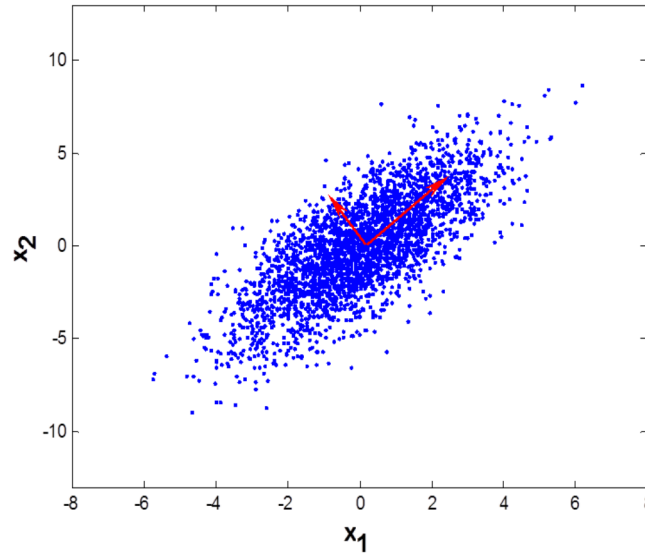
Az előzőekben leírt módszer lényegének a megértéséhez érdemes összehasonlítani azt egy másik hasonló célú, pénzügyi adatok elemzéséhez gyakrabban használt módszerrel, a főkomponens analízissel (principal component analysis, PCA).

A PCA célja valamilyen adathalmaz ún. főkomponenseinek a megtalálása, melyek a lehető legtöbb információt tartalmazzák az adatpontok összességéről. A probléma érzékeltetése a 2.1. ábrán látható, ahol egy kétdimenziós adathalmazt ábrázolunk. Amint látható, az adatpontok egy elnyújtott ellipszis mentén helyezkednek el, és ha azokat az ellipszis főtengelyeivel párhuzamos bázisban írjuk fel, több szempontból is előnyösebb reprezentációhoz jutunk. Először is ha az ellipszis kistengelyének megfelelő irányban a koordinátát elhagyjuk – azaz a pontokat a nagytengelyre vetítjük –, a lehető legkisebb mértékű hibát¹⁰ okozzuk. Másrészt az is látható, hogy a nagytengely irányában maximális az adatok varianciája, azaz a főkomponensek az egyes

¹⁰ $\text{cum}(S_i, S_j, S_k, S_l)$ értéke nem függ i, j, k és l sorrendjétől.

¹⁰Valamilyen megállapodás szerinti hibát kell ez alatt értenünk, PCA esetén a konvenció szerint legkisebb négyzetes értelemben értendő.

adatpontok halmazban elfoglalt helyét is jobban érzékeltetik. A PCA tehát főkomponenseket állít elő, melyeket fontossági sorrendbe rendez úgy, hogy az utolsó elhagyása jelentse a legkisebb információvesztést. Ezt úgy is megfogalmazhatjuk, hogy a legfontosabb főkomponensnek olyannak kell lennie, hogy a lehető legkisebb hibával reprezentálja az adathalmaz egészét.



2.1. ábra. A PCA szemléltetése. Amint látható, a nyilak irányába mutató bázisvektorokkal felírt koordinátákkal jobb képet kaphatnánk arról, hogy az egyes adatok hol helyezkednek el az ellipszisen belül.

Az ICA ezzel szemben nem csak a korrelálatlanságot, de a függetlenséget is előírja az egyes komponensek számára. Amint az tehát sejthető, az ICA a PCA egyfajta kiterjesztése, hiszen a PCA-hoz hasonlóan az adatok - illetve az ICA esetén jellemzően adatsorok - egy speciális reprezentációját keresi. A két módszer rokonságának megértése érdekében érdemes a PCA formális matematikai leírásán keresztülhaladnunk.

Ehhez használjuk fel a következő definíciókat: legyen $\mathbf{x}=[x_1, x_2, \dots, x_n]^T$ ¹¹ a bemeneti adatok egy n dimenziós vektora, $\mathbf{y}=[y_1, y_2, \dots, y_n]^T$ a transzformált adatoké, a $\mathbf{W}$ ortogonális transzformációs mátrix oszlopai pedig a transzformált tér $\mathbf{w}_i=[w_{i1}, w_{i2}, \dots, w_{in}]^T$ bázisvektorai. A PCA probléma alapjául szolgáló egyenlet tehát:

$$\mathbf{y} = \mathbf{W}\mathbf{x}. \quad (2.19)$$

Keressük azt a $\mathbf{W}$ transzformációs mátrixot, melyre igaz, hogy $\mathbf{x}$ -et oszlopterének¹¹ bármely m dimenziós alterére vetítve $\mathbf{x}$ legkisebb négyzetes hibájú becslését kapjuk, azaz minimalizáljuk

¹¹Természetesen ez lehet valószínűségi vektrováltozó egy realizációja is, a módszer általános felírásakor azonban általában skalárokkal dolgoznak, így a továbbiakban mi is ezt a jelölésrendszert követjük.

¹¹A mátrix oszloptere az oszlopai által kifeszített altér, mely a mátrix által reprezentált transzformáció képtere.

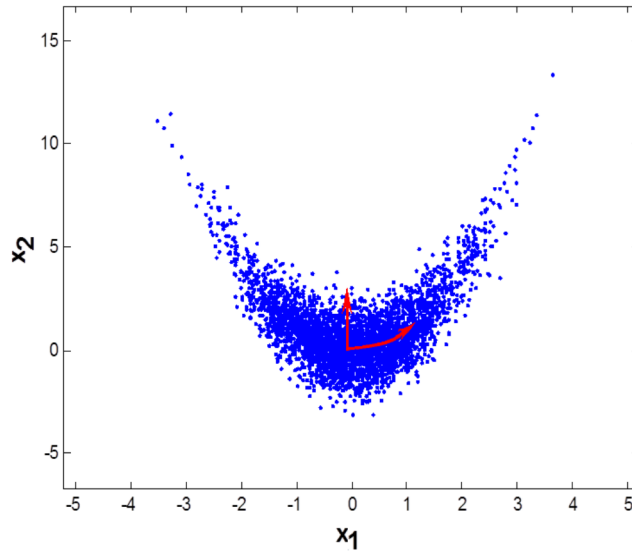
a következő költségfüggvényt:

$$C_{PCA}(\mathbf{W}) = \mathbb{E} \left\| \mathbf{x} - \sum_{i=1}^m (\mathbf{w}_i^T \mathbf{x}) \mathbf{w}_i \right\|^2 \quad (2.20)$$

Felmerül a kérdés, hogy kiterjeszthető-e a PCA olyan esetekre, mikor a keresett főkomponensek a bemeneti adatok valamilyen nemlineáris transzformációja által adóttak, ahogy a 2.2. ábrán ábrázolt esetben. a (2.20) lineáris kritériumot ilyen esetekben a főkomponensekre alkalmazott $g_i(\cdot)$ nemlineáris függvényekkel kell módosítanunk, a következőképpen:

$$C_{NLPCA}(\mathbf{W}) = \mathbb{E} \left\| \mathbf{x} - \sum_{i=1}^m g_i(\mathbf{w}_i^T \mathbf{x}) \mathbf{w}_i \right\|^2 = \mathbb{E} \left\| \mathbf{x} - \mathbf{W}^T \mathbf{g}(\mathbf{W}\mathbf{x}) \right\|^2, \quad (2.21)$$

ahol $\mathbf{W}$ a $\mathbf{w}_i$ oszlopvektorokból létrehozott mátrix, $\mathbf{g}(\mathbf{W}\mathbf{x})$ pedig egy oszlopvektor, i -edik eleme $g_i(\mathbf{w}_i^T \mathbf{x})$.



2.2. ábra. Az NLPCA szemléltetése. Amint látható, a főkomponensek ezúttal nemlineáris kapcsolatban állnak az $\{x_1, x_2\}$ bázis koordinátaival.

A 2.2. pontban említett okokból állítsuk elő az $\mathbf{x}$ bemeneti adatokból a $\mathbf{v}$ fehérített adatokat, és tegyük fel, hogy a fehérítés utáni adatok és $\mathbf{y}$ dimenziója megegyezik. Megmutatható (Hyvärinen et al. 2001), hogy ekkor a (2.19) egyenlet jelöléseit használva igaz a következő összefüggés:

$$C_{NLPCA}(\mathbf{W}) = \mathbb{E} \|\mathbf{v} - \mathbf{W}^T \mathbf{g}(\mathbf{W}\mathbf{v})\|^2 = \mathbb{E} \|\mathbf{y} - \mathbf{g}(\mathbf{y})\|^2 = \sum_{i=1}^m \mathbb{E} [y_i - g_i(y_i)]^2. \quad (2.22)$$

Legyen minden i -re igaz, hogy

$$g_i(y) = y^2 + y. \quad (2.23)$$

Látható, hogy ekkor a (2.22) egyenlet éppen a fehérített adatok esetén érvényes a (2.6) egyenlettel lesz ekvivalens:

$$C_{NLPCA}(\mathbf{W}) = \sum_{i=1}^m \mathbb{E} y_i^4. \quad (2.24)$$

2.4. ICA megvalósítása nemlineáris PCA-hálóval

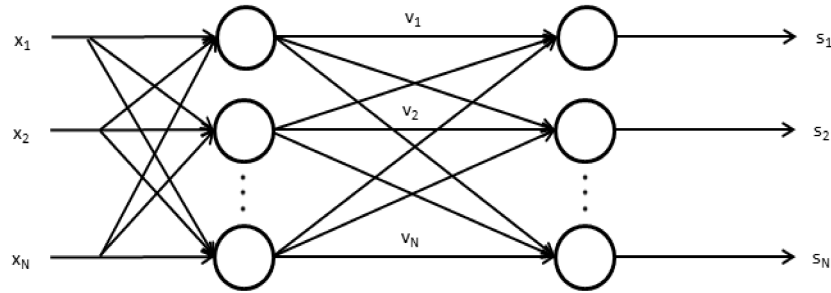
A (2.24) egyenletnek megfelelően tehát az NLPCA elvégzése tulajdonképpen egyenértékű az ICA probléma megoldásával, ha a megfelelő műveleteket fehérített adatokon végezzük.

Mind az adatok fehérítésére, mind az NLPCA megoldására kézenfekvő módszer a neurális hálózatos architektúra használata (Horváth, Altrichter, Pataki, Strausz, Takács & Vaylon 2006). A neurális hálózat a 2.3. ábrán látható módon épül fel. A hálózat j -edik adatfeldolgozó eleme – melyeket neuronnak nevezünk, az ábrán körök jelölik – egy-egy nemlineáris műveletet végez: ha az egység n -dimenziós bemenete $\mathbf{v}$, kimenete pedig s , akkor az egység az

$$s_j = g_j \left(\sum_{i=1}^n w_{ij} \mathbf{v}_i \right) \quad (2.25)$$

nemlineáris műveletet végzi, ahol a w_{ij} oefficiensek az egység ún. súlyai. A súlyokat minden probléma esetén az adott problémának megfelelő költségfüggvény minimalizálásával tanítjuk meg a hálózatnak. Ezeket az algoritmusokat tanítási szabályoknak nevezzük. A költségfüggvény megfogalmazásakor kell figyelembe vennünk a $g_j(\cdot)$ leképezés megfelelő megválasztását is – több neuronból álló hálózatok esetén jellemzően $g_j(\cdot) = g(\cdot)$ minden j -re.

Feladatunk tehát – visszatérve a (2.2) egyenlet jelöléseihez, felhasználva a (2.22) egyenlet tanulságait – a megfigyelt $\mathbf{x}$ adatokból az $\mathbf{a}$ független komponensek értékeinek előállítását. Ezt a $\mathbf{V}$ mátrix segítségével előállított $\mathbf{v}$ fehérített adatok neurális hálózattal való feldolgozásával érhetjük el. A klasszikus, 2.25 egyenlet szerinti megközelítés helyett a 2.3. ábrának megfelelő felépítésű, lineáris hálózatot alkalmazunk, a nemlinearitást csak a második réteg tanításához



2.3. ábra. Az ICA-t megvalósító neurális hálózat

felhasználva, hiszen a (2.1) egyenlet szerint a keresett értékek és a bemenet közötti kapcsolat lineáris. A hálózat súlyait megfelelően a $\mathbf{W}$ szétválasztómátrixba gyűjtve a következő egyenlet tagjait kell tehát meghatároznunk:

$$\mathbf{s} = \mathbf{W}\mathbf{v} = \mathbf{W}\mathbf{V}\mathbf{x}. \quad (2.26)$$

Az adatfehérítéshez a $\mathbf{v}=\mathbf{V}\mathbf{x}$ egyenlet $\mathbf{V}$ együtthatóira van szükség, melyeket szintén meghatározhatjuk neurális hálózat segítségével, a k -adik tanítási lépésben a következő szabályt alkalmazva (Plumbey 1993):

$$\mathbf{V}(k+1) = \mathbf{V}(k) + \mu(k) \left[\mathbf{V}(k)\mathbf{x}(k)\mathbf{x}(k)^T \mathbf{V}(k)^T - \mathbf{I} \right] \mathbf{V}(k), \quad (2.27)$$

ahol μ az ún. tanítási vagy bátorsági tényező.

A második, a független komponensek szétválasztását végző, $\mathbf{v}$ bemenetű neuronréteg tanítását a következő egyenlet szerint végezhetjük (Wang & Karhunen 1996):

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k) \mathbf{g}(\mathbf{W}(k)\mathbf{v}(k)) \mathbf{v}(k)^T + \alpha (\mathbf{I} - \mathbf{W}(k)\mathbf{W}(k)^T) \mathbf{W}(k), \quad (2.28)$$

ahol α egy konstans a $[0,5,1]$ intervallumból – a legjobb eredményeket 0,5 esetén kaptuk.

Megmutatható, hogy a (2.28) egyenletnek megfelelő algoritmussal tanított háló az $\sum_{i=1}^n \mathbb{E}f(s_i)$ költségfüggvényt minimalizálja, ahol az $f(\cdot)$ függvény $g(\cdot)$ integrálja. Ha tehát a (2.24) egyenletnek megfelelő költségfüggvény minimalizálására törekszünk, $g(t)=4t^3$ választással kell élnünk. Ezzel a választással azonban az algoritmus nem stabil (Horváth et al. 2006), így szerencsésebb a neurális hálózatok körében elterjedt $g(t) = \tanh(t)$ függvény használata, melynek Taylor-soros közelítésében a t^3 tag dominál.

A hálózatok tanítását az idősor minden k elemére elvégezzük, az idősoron akár többször is végighaladva egészen addig, míg a $\mathbf{W}$ és $\mathbf{V}$ mátrixok együtthatói nem konvergálnak. Ehhez elengedhetetlen a μ és η bátorsági tényezők megfelelő megválasztása. Az általunk megvalósított algoritmus adaptív bátorsági tényezőket használ, azokat egy kezdeti – megfelelően nagy – értékről addig csökkenti, míg azok megfelelő mértékű konvergenciát nem biztosítanak.

Fontos megjegyeznünk az ICA analízissel kapcsolatban, hogy a keletkezett komponensek nem egyértelműek. Egyrészt a komponensek sorrendje tetszőleges lehet, tehát többszöri futtatás esetén a komponensek sorrendje cserélődhet, másrészt az eredményül kapott komponensek az eredeti $\mathbf{s}$ vektor elemeinek bármilyen skalárszorosai lehetnek – ez ugyanis a függetlenséget nem befolyásolja, és csak a keverőmátrix elemeinek értékét kellene megváltoztatni, melyek szintén az ICA által meghatározottak. Éppen ezért a továbbiakban az ICA minden eredményét a $[-1,1]$ tartományra normálva ábrázoljuk, illetve a különböző idősorok esetén kapott komponensek összehasonlításakor a kapott komponenseket sztenderdizálni fogjuk, tehát 0 várható értékű, 1 szórású idősorokat állítunk elő.

2.5. FastICA algoritmus

Az ICA egyik leggyakoribb implementálási módja a fixpont-megközelítésre alapuló FastICA algoritmus (Hyvärinen & Oja 2000), melyet a dolgozatban a neurális architektúrával való összehasonlításhoz használunk.

Az algoritmus feltételes optimalizációt végez, a $C_{FastICA} = \mathbb{E}(f(\mathbf{w}^T \mathbf{x}))$ költségfüggvény minimumát keresi $\|\mathbf{w}\|^2 = 1$ feltétel mellett, ahol a $f(\cdot)$ tetszőleges függvény, deriváltját $g(\cdot)$ -vel jelöljük. Az algoritmus alapváltozata külön-külön határozza meg a független komponenseket, lépései egy független komponens meghatározásához (a korábbi jelöléseket használva):

1. Kezdeti – véletlen – $\mathbf{w}_0$ súlyvektor kiválasztása.
2. $\mathbf{w}_{k+1} = \mathbb{E}(\mathbf{x}g(\mathbf{w}_k^T \mathbf{x})) - \mathbb{E}(g'(\mathbf{w}_k^T \mathbf{x})) \mathbf{w}_k$.
3. $\mathbf{w}_{k+1} = \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_{k+1}\|}$.
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 tulajdonságai a módszernek, hogy a $g(\cdot)$ függvény tetszőleges lehet – természetesen az algoritmus eredménye szempontjából a megválasztás nem közömbös, bizonyos függvények esetén lapultság, illetve negentrópia alapú megközelítéseket kaphatunk. Az algoritmus láthatóan nem igényli bátorsági tényezők hangolását, konvergencia-tulajdonságai pedig kiválóak.

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 kereteit, ehhez vö.: Horváth et al. (2006), Hyvärinen & Oja (2000), vagy Hyvärinen et al. (2001).

2.6. Alkalmazási példák

Az ICA alkalmazási lehetőségeinek szemléltetésére a 2.4. pontban ismertetett struktúra használata a legkézenfekvőbb megoldás, ezzel egyszersmind a létrehozott architektúra funkcionalitása is ellenőrizhető.

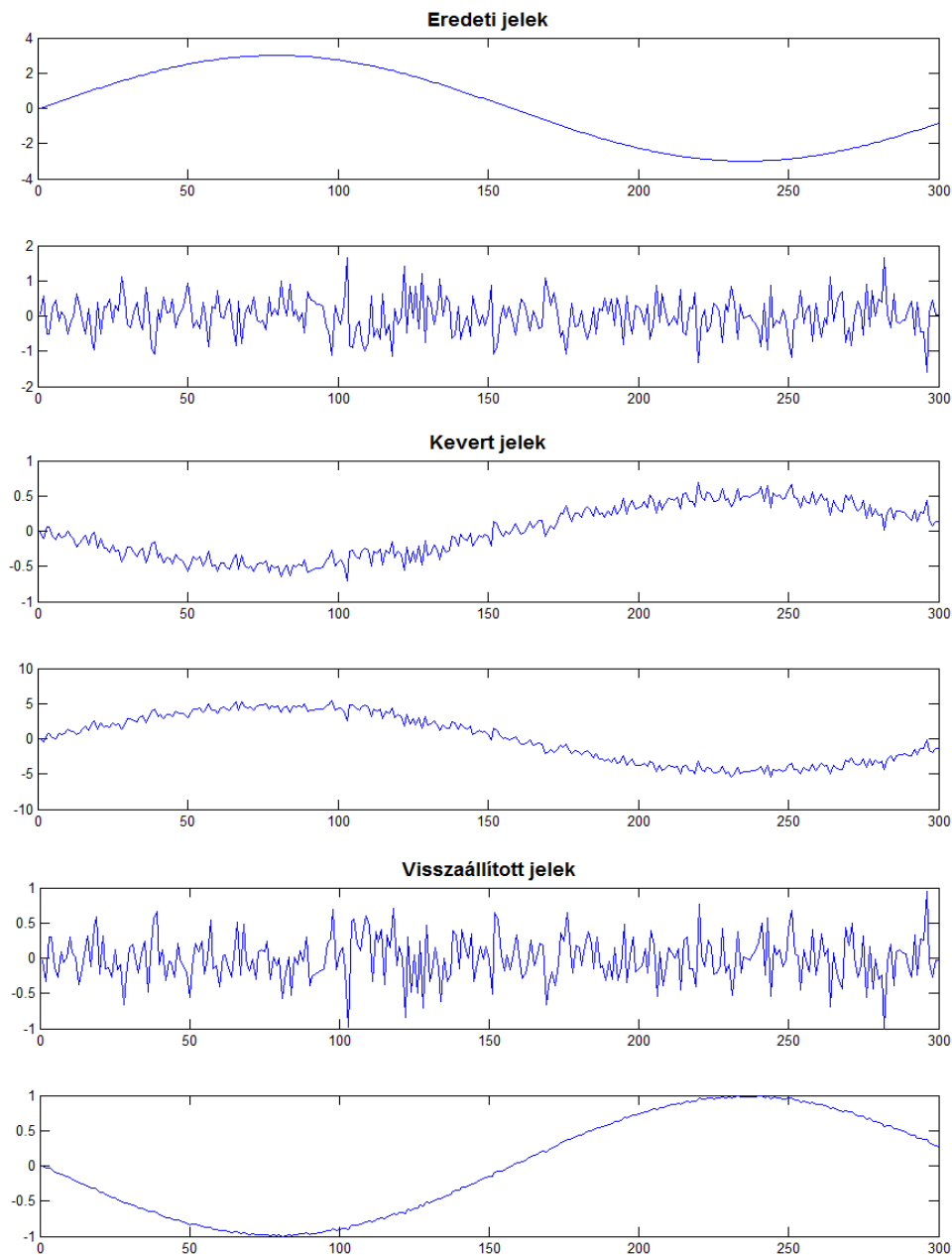
Alkossunk két tetszőleges, de ismert idősorból két kevert idősort egy keverőmátrix segítségével, majd vizsgáljuk meg, hogy a módszer képes-e visszaállítani az eredeti idősorokat. Mivel az ICA elsősorban mérnöki problémák megoldása esetén elterjedt, legyen a két idősor – melyet ezen a területen gyakran egyszerűen jeleknek neveznek – ilyen témakörben használatos: az egyik az $x \in [0,3]$ intervallumon értelmezett $f(x) = 3 \sin(x)$ függvény, a másik pedig egy Y egyenletes eloszlású valószínűségi változó által generált mintasorozat.

A keverőmátrix elemeinek pontos értéke az eredmény szempontjából nem releváns, mivel ahogy azt korábban említettük, az ICA által előállított független komponensek az eredetiek tetszőleges skalárszorosai lehetnek. Így ha a visszaállított jelek valamilyen skalárszorosai megfelelnek az eredetieknek, a keverőmátrix oszlopai esetén ugyanez a helyzet fog fennállni. A két eredeti jel, a kevert jelek, valamint a visszaállított jelek a 2.4. ábrán láthatóak.

Amint a 2.4. ábrán is látható, a módszer működőképes, a visszaállított jelek szinte tökéletesen megfelelnek az eredetiek skalárszorosainak. Érdeemes megjegyezni, hogy a mintaszám csökkenése, a jelek számának növekedése, illetve a keverőmátrix kondicionáltsága jelentős hatást gyakorolhatnak az algoritmusra, így szerencsétlen esetben a visszaállítás a képen láthatónál rosszabb minőségben is történhet, különösen több jel esetén. A 2.5. ábrán látható, hogy négy jel dekompozícióját az algoritmus szintén képes elvégezni, az egyes visszaállított jeleket terhelő zaj azonban megnő, illetve a dekompozíció sem olyan tökéletes, mint az előző, egyszerűbb esetben. Mindezek ellenére az eredeti jelek jól felismerhetők ebben az esetben is.

2.6.1. Mérnöki alkalmazások

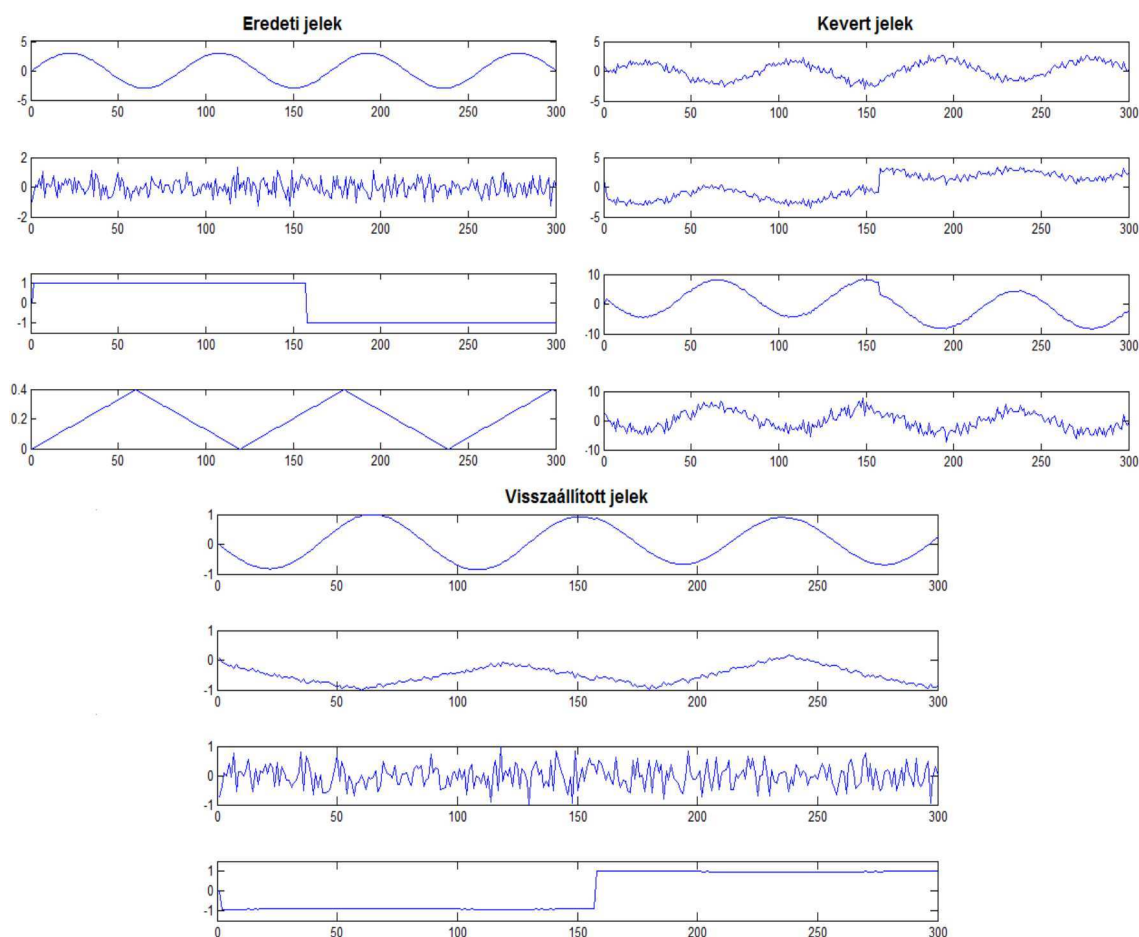
Számos olyan probléma merül fel a mérnöki gyakorlatban, melynek során adatok függetlenítésére, illetve különböző komponensek elkülönítésére van szükség. Ennek megfelelően az ICA az évek során számos esetben bizonyította alkalmazhatóságát. A téma bőséges irodalma is a módszer elterjedtségére utal.



2.4. ábra. A megvalósított ICA algoritmus eredményei két jel esetén, az algoritmust 300-300 mintaelem-re lefuttatva

Legjellemzőbb felhasználási módjai a dekompozíciós feladatok. Ezek közül is a legegyszerűbb eset az, ha a (2.1) egyenlet változóinak száma egy. 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 & Ulrych (2003) által leírt módszerrel lehetséges.

A többdimenziós dekompozíciós feladatok nagy része a korábban említett BSS probléma megoldását igényli. Ez a klasszikus, a (2.3) egyenlet által megfogalmazott feladat, megoldása számos gyakorlati problémára nyújthat megoldást. A BSS probléma klasszikus példája – a koktélparti probléma mintájára – a hang- illetve beszédfelismerés, több mikrofon által rögzített



2.5. ábra. A megvalósított ICA algoritmus eredményei négy jel esetén, az algoritmust 300-300 mintaelemre lefuttatva

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, Lewicki, Girolami & Sejnowski 1999).

Érdemes megjegyezni, hogy a BSS problémát bármilyen adatstruktúra esetén megfogalmazhatjuk, akkor is, ha az egyes független komponensek maguk is több dimenziósak. Ilyen például az arcfelismerés esete. Ekkor a kevert kép, vagy képek több, független képpé való dekompozícióját keressük. A későbbi felismerés szempontjából különösen fontos a lokális torzulások kezelése, mely körültekintő komponensekre bontást igényel (Kim, Choi, Yi & Turk 2005).

További alkalmazási lehetőség bármilyen többdimenziós mért adat dekompozíciója, Liu, Pearlson, Windemuth, Ruano, Perrone-Bizzozero & Calhoun (2007) például agyi fMRI¹² felvételek feldolgozására használták a módszert. A szerzők a felvételeken tapasztalható elváltozások, és az egyponos nukleotid-polimorfizmus (single nucleotide polymorphism, SNP) kap-

¹²A 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.

csolatát vizsgálták. Az SNP a DNS egy bázispárjának mutációja, mely a populáció legalább 1%-ában azonos helyen megjelenik. Az SNP-k az emberi genetikai variációk nagyjából 90%-át teszik ki, hatásuk feltárása a genetika egyik intenzíven kutatott területe (Genomic Science Program 2012). A vizsgálathoz a mért adatok dekompozícióját egy úgynevezett 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, Theis, Stadthanner, Lang, Tome & Teixeira 2004), mivel a módszer a különböző mért jelek és a zaj elkülönítésében igen robusztus eszköznek bizonyul – ezt a tulajdonságát már a 2.4. ábra is mutatta. További, nem triviális alkalmazása lehet a módszernek különböző digitális képek vízjelezése is. Digitális csatornák védelmére gyakran alkalmaznak ún. vízjelezést, például dokumentumok hitelesítése esetén. Az adatokhoz valamilyen érzékeny információt adnak hozzá, mely az adatok sérülése esetén – például jogtalan módosítás miatt – jelzi a sérült adatok jelenlétét a fogadó fél számára. Lehetséges a vízjel hozzáadása az adatok független komponenseihez is (Bounkong, Toch, Saad & Lowe 2003), mely további védelmet nyújt az adatok sérülése ellen.

2.6.2. Pénzügyi alkalmazások

Az előbbieken felsorolt alkalmazások mellett a módszer népszerűsége a gazdasági alkalmazások területén is egyre nő, hiszen a gazdasági adatok elemzésekor sok, a felsoroltakhoz hasonló jellegű probléma merül fel. Ezen a területen a módszert elsősorban pénzügyi adatok és modellek vizsgálatára alkalmazták. Ennek oka, hogy módszer nagy adatmennyiség esetén hatékonyabb, így nagyfrekvenciás idősorok esetén rendkívül jól alkalmazható, míg más, például makro idősorok esetén, ahol jellemzően havi, negyedéves vagy éves adatok állnak rendelkezésre, kevésbé hatékony. A magyar gazdasági szakirodalom eddig nem foglalkozott a módszer használatával, a kutatás során nem is találtunk releváns magyar nyelvű irodalmat a témában. A külföldi folyóiratokban 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 nyújtunk az általunk feldolgozott pénzügyi alkalmazásokkal kapcsolatos szakirodalomról.

A portfóliókezelés 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. Egy portfóliót immuni-

záltnak tekintünk, ha az átlagideje nulla, vagyis a kamatláb kicsiny megváltozása nem befolyásolja a portfólió értékét. Ahhoz azonban, hogy ezt megvalósítsák, a gyakorlatban nagyszámú kötvény átlagidejét kellene kiszámítani, tehát az immunizáció hatékonysága és a számításigény között erős trade-off van. A számításigény csökkentésére általában főkomponenseket használnak, amelyek egy adott elemű kötvényportfólió varianciájának a lehető legnagyobb hányadát magyarázzák, és ezek segítségével „állítják be” az átlagidőt nullára. Gonzalez & Nave (2010) cikkükben PCA és ICA megközelítést használva vizsgálták, 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 & Salinelli (2003) is.

Egy másik gyakran használt pénzügyi modell az APT¹³, amely a faktormodellek körébe sorolható. A modell segítségével lehetőség van egy adott értékpapír hozamának faktorokra való felbontására. Cha & Chan (2000) cikkükben részvényhozamokra alkalmazták a független komponens analízist, hogy felállítsanak egy többfaktoros modellt. Szintén részvényhozamok dekompozíciójára használták az ICA-t Back & 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-modellek¹⁴ a volatilitás folyamatát modellezik. Wu & Yu (2005) a GARCH-modellekkel kötötte össze az ICA használatát. A szerzők azt vizsgálták, hogy a hozamokra illesztett VAR-modellek reziduumaiknak PCA és ICA által dekomponált változókra illesztett volatilitás-modellek 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 le az idősort mint a főkomponenseken alapuló modellek.

Szintén ezen a területen, de az előrejelzések megbízhatóságát vizsgálta Chiu, Lee & Lu (2009) Nikkei225 és TAIEX indexek részvényeinek hozamaira. A vizsgálat során a független komponens elemzést a zaj idősorából való kiszűrésére használták. A korábbiakhoz hasonlóan a PCA alapú megközelítéssel vetették össze az előbbi módszert. A zaj kiszűrése után a komponensekre szupport-vektor regressziót (support vector regression, SVR) használtak az előrejelző modell becsléséhez. A tesztek során több mutatót is vizsgáltak az előrejelzések megbízhatóságának megállapítására. Az előrejelzési hiba mérését MAD (mean absolute deviation), RMSE (root mean squared error) és NMSE (normalized mean squared error) mutatókkal, míg az előrejelzés adekvátságának mérését DS (directional symmetry), CP (correct up trend) és CD (correct

¹³APT - Arbitrage Pricing Theory, lásd (Ross 1976).

¹⁴GARCH - Generalized Autoregressive Conditional Heteroskedasticity, lásd Bollerslev (1986).

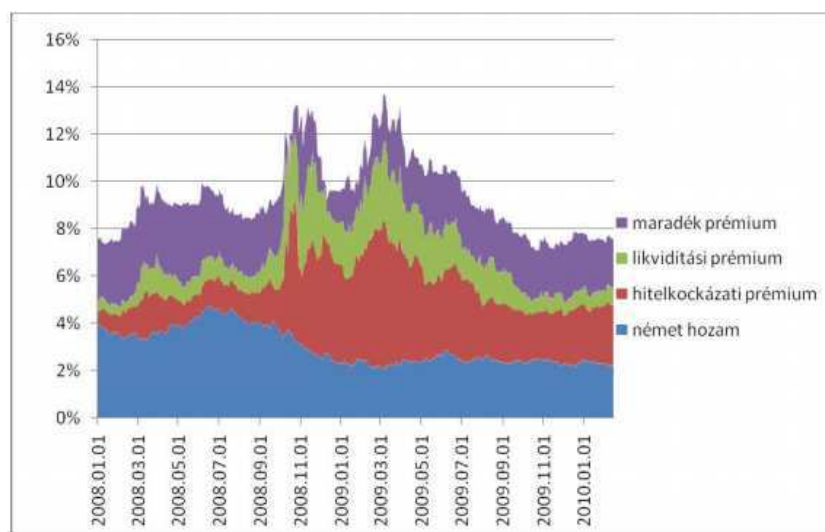
down trend) mutatókkal végezték. Eredményeik alapján minden mutató esetén jobb eredményt szolgáltatott az ICA-SVR módszer.

A korábban leírtaknak megfelelően az adatelemzési módszerek egyre elterjedtebb eszközét képezik a neurális hálók. A módszert pénzügyi területen elsősorban előrejelzésre használják. Liu & Wang (2011) ICA és 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 (root mean squared error) mutatókat számítottak, mindkét mutató esetén az ICA-BP modell teljesített jobban.

3. ICA alkalmazása kötvényhozamok dekompozíciójára

A továbbiakban az ICA sajátosságait teszteljük kötvényhozam-adatok felhasználásával. Az első részben röviden bemutatjuk a kötvényhozamok dekompozíciójának pénzügyi hátterét, és egyben rávilágítunk arra, hogy az ICA miért alkalmazható ennek a problémának a megoldására. Az empirikus tesztek során nem célunk a komponensek pénzügyi tartalmának meghatározása, sokkal inkább a módszer sajátosságainak vizsgálata lesz a középpontban.

Monostori (2010) dolgozatában bemutatta, hogy a kötvényhozamok mögött elméleti szempontból különböző komponensek állnak, és ezen komponensek hatására változik a kötvényhozamok szintje. A szerző ötéves fix kamatozású államkötvényekre végezte el vizsgálatát, azonban adathiány miatt a komponensek meghatározását nem tudta megvalósítani empirikus úton. Dolgozatának arra az eredményére alapozzuk vizsgálatunkat, miszerint a kötvényhozamok változásában olyan változók játszanak szerepet, mint a hitelkockázati felár, likviditási felár, kamatkockázati felár stb.. A szerző négy komponenst azonosított a kötvényhozamok esetén. A probléma azonban, hogy ezek a felárak korreláltak, és így nem választhatók el egymástól feláranként, vagyis a teljes kötvényhozamok nem egyszerűen a prémiumok összegeként adódik. A kötvényhozamok változásában közrejátszó látens változók kiszámítása és az ezekhez köthető pénzügyi jelentéssel bíró változók meghatározása kulcsszerepet játszhat számos esetben. Egyfelől szeparálhatóvá válnak az egyes kockázati tényezők hatásai, így lehetőség van célzott beavatkozásra az adott területen. Monetáris szempontból például egyáltalán nem elhanyagolható, hogy az adott sokkhatásra a jegybank kamatemeléssel vagy devizapiaci tranzakciókkal reagáljon. Szintén fontos lehet például pénzügyi termékek árazásában, ha az egyes hatásoknak megfelelő kockázati prémiumot mérni lehet.



3.1. ábra. A hozamokat meghatározó komponensek (Monostori 2010)

Dolgozatunkban a kötvényhozamok dekompozíciójának pénzügyi háttérét a fent említett dolgozat szolgáltatja. Módszertanilag pedig azért van lehetőség az ICA használatára, mivel a pénzügyi idősorok eloszlása jellemzően leptokurtikus (Cont 2001), ami az ICA alkalmazhatóságának egyik feltétele.

3.1. Az empirikus tesztek menete

A továbbiakban 5 ország (Egyesült Államok, Nagy Britannia, Németország, Svédország, Magyarország) 1997-2012 közötti ötéves államkötvények hozamait vonjuk be az elemzésbe. A megfigyelések napi gyakoriságúak, ami összesen 3927 napot jelent országonként.

Az első problémát az okozza, hogy a kereskedési napok az ünnepnapok eltérése miatt nem egyeznek minden esetben. A kérdés ebben az esetben az, hogy jelentősen befolyásolja-e az eredményeinket, ha azt a napot, amikor valamelyik országban nem volt kereskedés, elhagyjuk a mintából. Az összevetéshez három adatszámítási eljárást alkalmazunk. Az első eljárás során a hiányzó adatokat az azt megelőző napi hozammal pótoljuk. A második esetben lineáris interpolációt használunk, míg a harmadik esetben az adott hiányzó adatig meglévő megfigyelésekre GARCH modellt illesztünk, majd a becsült paraméterek segítségével egynapos előrejelzést készítünk a hiányzó adatra, így járva el az összes hiányzó adat esetén. Várakozásaink szerint nem lesz jelentős hatása az adatszámítási eljárás konkrét módjának, tekintve, hogy országonként a megfigyelt időszak csupán hozzávetőleg öt százalékában nem állt rendelkezésre adat.

Az ICA tesztelését az adatszámítási módszerek komponensekre való hatásának elemzésével kezdjük. Megvizsgáljuk a különböző imputálási módszerekkel kapott idősorok független komponenseit, ezek egyidejű korrelációját az azonos adatszámítási eljárás többi komponensével, illetve összehasonlítjuk a különböző módszerekkel imputált adatokból számított komponenseket is. Elvégezzük továbbá a két ismertetett ICA algoritmus összehasonlítását is.

Ezután a kötvényhozamok mozgását meghatározó komponensek számát határozzuk meg. Ehhez a komponensekre számított korrelációs mátrixot használjuk. A komponensek számának helyes megválasztása esetén nyilvánvalóan egy az egységmátrixhoz nagyon hasonló mátrixot kell kapnunk a független komponens analízis után, vagyis a komponensek önmagukkal való korrelációja 1, míg a többi független komponenssel való korrelációja majdnem 0 lesz. A komponensek számának változtatásával becsülhetjük tehát meg a kötvényhozamok mögött rejlő látens változók számát. Ezután megvizsgáljuk a használt idősorok száma és a független komponensek száma közötti összefüggést is.

A megfelelő komponensszám meghatározása után az ICA robusztusságát vizsgáljuk. A kérdés ebben az esetben, hogy a komponensek mennyire robusztusak, vagyis a korrelációk hány mintaelem megváltoztatása, sokkolása esetén változnak meg. A teszteket elvégezzük mindkét ismertetett algoritmus esetén, különböző sokkolási módszereket eltérő mértékben alkalmazva.

A módszer vizsgálatát az adatok különböző ablakozásával is elvégezzük. Ebben a részben

arra vagyunk kíváncsiak, hogy adatelemek az idősorból való elhagyása mennyire torzítja az ICA által kiszámított komponenseket. Ezt egyrészt az adatok szisztematikus elhagyásával, másrészt az idősorok ablakozásával vizsgáljuk. Az ablakozás egyben lehetőséget nyújt pénzügyi szempontból érdekesebb időszakok vizsgálatára is, így az ablakokat jelentősebb pénzügyi sokkok, válságok környezetébe helyezzük.

3.2. Az empirikus tesztek kivitelezése

3.2.1. Adatpótlási módszerek hatása

Azon napok elhagyása, amikor valamelyik országban nincs adatunk, 405 mintaelemmel csökkenti az idősorokat. A három adatpótlási eljárás után kapott adatok leíró statisztikáit, illetve a hozamok idősorait a C.1.-C.3. táblázatok, és a megfelelő ábrák ismertetik. A statisztikák esetén az egyes megközelítések között alig figyelhetünk meg különbséget, az eredmények legtöbbször még ezredekben sem térnek el egymástól.

3.1. táblázat. Leíró statisztikák a hiányzó adatokat tartalmazó napok elhagyásával előállt idősorokra

Statisztika	US	GB	D	SE	HU
Átlag	3,901	4,473	3,507	3,896	9,257
Szórás	1,526	1,437	1,062	1,171	2,984
Módusz	4,550	7,040	4,730	3,810	6,850
Medián	4,010	4,753	3,624	3,947	8,270
Ferdeség	-0,111	-0,628	-0,532	-0,226	1,434
Csúcsosság	-0,871	-0,141	-0,395	-0,604	1,068
IQR	2,268	1,330	1,643	1,695	2,698
MAD	1,288	1,104	0,867	0,956	2,253
JB-teszt	118,637	234,131	188,707	83,489	1375,039

Az első adatpótlási módszer a nulladrendű interpoláció, azaz a hiányzó adatokat mindig az előző napi értékkel helyettesítjük, illetve ha az első adat hiányzik, a következő napival. Az idősorokat a C.1. ábra, a leíró statisztikákat a C.1. táblázat szemlélteti. A második esetben az adatokat elsőrendűen interpoláljuk, azaz a t pontban az $f(t)$ adatot a szomszédos t_0 és t_1 pontokból határozzuk meg:

$$f(t) = f(t_0) + \frac{f(t_1) - f(t_0)}{t_1 - t_0}(t - t_0) \quad (3.1)$$

Az idősorokat a C.2. ábra, a leíró statisztikákat a C.2. táblázat szemlélteti. Amint látható, lényeges eltérés nincs az eddigi két módszer között. Végül az ARIMA-GARCH-modellel¹⁵

¹⁵Az adatpótlás során először egységgyök jelenlétére tesztelünk, majd a megfelelő ARMA(p,q) modell meghatározása után GARCH(p*,q*)-modellt becslünk. A modellek meghatározásánál mindig a hiányzó adatig meglévő megfigyeléseket használjuk, majd a következő hiányzó adatnál ismét újrabecsüljük az együtthatókat. Modellszelekciós kritériumként az SIC, LM és ARCH-LM statisztikákat használjuk.

(lásd pl. Hamilton (1994) vagy Bollerslev (1986)) történő pótlás hatását vizsgáljuk, a kapott idősorok a C.3. ábrán, a leíró statisztikák a C.1. táblázatban láthatóak.

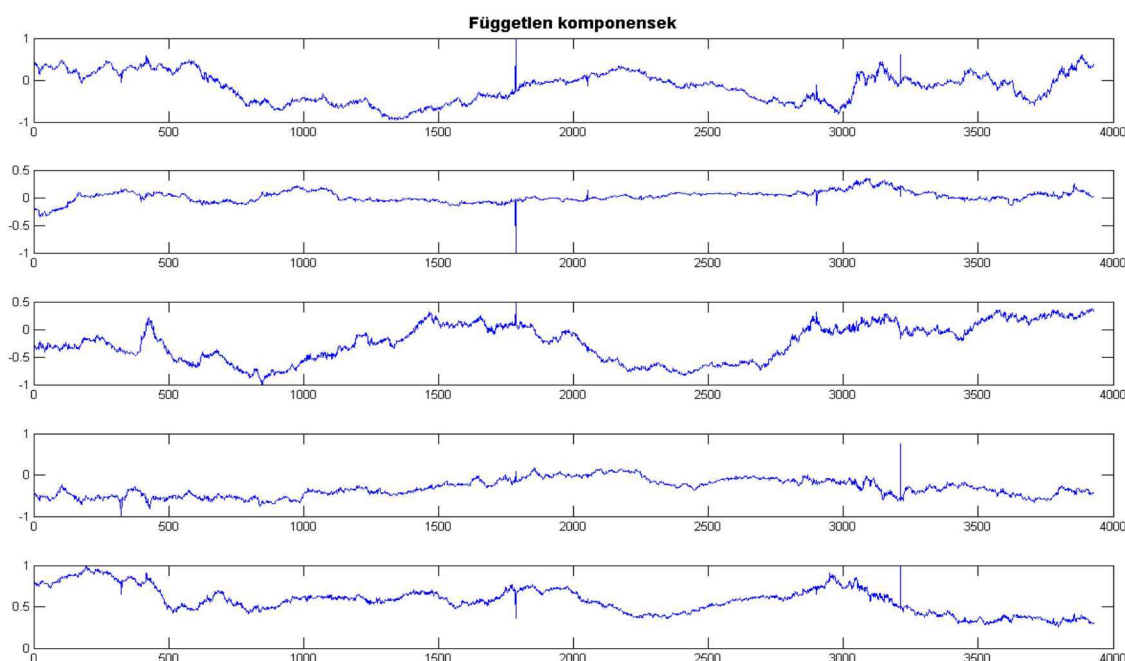
$$\Phi(L)Y_t = c + \Theta(L)\varepsilon_t \quad (3.2)$$

$$\varepsilon_t = \sigma_t \eta_t$$

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^q \beta_i \sigma_{t-i}^2 \quad (3.3)$$

Adatpótlás hatása a komponensekre

Amint az a korábban felsorolt táblázatokban látható, az adatpótlási módszerek különbsége nincs hatással a leíró statisztikákra. Fontosabb kérdés azonban, hogy a független komponensekre, illetve azok egyidejű korrelációjára hatással van-e. A kérdés megválaszolásához elő kell állítanunk a komponenseket minden adatpótló eljárásra kapott idősor esetén, majd minden esetben ki kell számítanunk a komponensek korrelációs mátrixát. Fontos megjegyzés, hogy a kereszt-korrelációk számításakor a negatív korrelációs koefficienseket épp úgy korrelációknak kell tekintetnünk, mint a pozitívokat, mivel az ICA kimenetén a komponensek tetszőleges konstans szorzója -1 is lehet. Ha tehát két komponens között a korreláció negatív, de az együtttható értéke nagy, akkor a két komponens korrelál, éppen csak az egyik -1-szeres szorzóval jelent meg az algoritmus kimenetén. Az elkülönített független komponensek a 3.2. ábrán láthatóak.



3.2. ábra. A FastICA algoritmus által meghatározott független komponensek a teljes idősorra.

Várakozásaink szerint két eredmény lehetséges a korrelációs mátrixokkal kapcsolatban. Ha az adatpótlás nincs jelentős hatással, akkor a módszer közel azonos komponenseket állít elő, így a korrelációs mátrixban öt darab 1 körüli érték fog szerepelni, míg a többi elem nulla. Másfelől még nem vagyunk bizonyosak benne, hogy hány „strukturális” komponenszt kapunk, így elképzelhető, hogy egy vagy több komponens csak a zajt szűri, amelyek azonban változhatnak az adatpótlás hatására, és lehet, hogy függetlenek is lesznek, így a közöttük fellépő korreláció nulla lesz. A korrelációs mátrixokat a C.5.-C.7. táblázatok tartalmazzák.

A módszerrel kapcsolatban fontos megjegyzés, hogy egyrészt a keresztkorrelációs értékek láthatóan alacsonyok, egy bizonyos határ alá azonban nem tudjuk csökkenteni őket. Ennek okát abban látjuk, hogy a (2.27) és a (2.27) egyenletekkel leírt optimalizációs eljárások költségfüggvényeinek valószínűleg ebben a korrelációs nagyságrendben vannak az első lokális minimumhelyei. Mivel az optimumkereső algoritmusok alapváltozatai egy lokális minimum megtalálása után nem képesek azt elhagyni, felmerülhet a kérdés, hogy a megtalált minimumok globális minimumok-e.

Ennek kiderítésére a FastICA algoritmus egy szabadon hozzáférhető megvalósítását használjuk, mely fixpont algoritmuson alapul, és mindig a mintapontok által biztosított maximális pontosságú becslést adja (Horváth et al. 2006). Ahhoz, hogy egyetlen értékbe tömörítsük a különbséget, a keresztkorrelációk abszolút értékének átlagát számítjuk ki minden korrelációs mátrixra. Az eredményeket a 3.2 táblázat tartalmazza.

3.2. táblázat. Komponensek átlagos abszolút keresztkorrelációi 5 komponens esetén, egy-egy imputálási módszer komponenseit összehasonlítva

Adatpótló eljárás	Konstans	Lineáris	GARCH
Neurális ICA	0,0276	0,0248	0,0276
FastICA	$1,67 \cdot 10^{-15}$	$1,9 \cdot 10^{-15}$	$2 \cdot 10^{-15}$

Látható, hogy egyrészt az átlagos korrelációk nem különböznek lényegesen, azaz az adatpótlás módja – mint ahogy az ilyen kevés hiányzó adat esetén várható volt – nincs jelentős hatással a független komponensek önmagukkal vett átlagos korrelációira. Másrészt látható, hogy a FastICA valóban optimális megoldást biztosít, a keresztkorrelációk alig mérhetőek. Az itt leírt tapasztalatok miatt, valamint, mivel a FastICA konvergenciasebessége az általunk használt rendszeren három nagyságrenddel nagyobb a neurális architektúráénál, a továbbiakban a nagy műveletigényű tesztek a FastICA használatával fogjuk végezni.

Különböző imputálások után kapott komponensek összehasonlítása

Az egyes adatpótlási módszerek komponenseinek az azonos módszerrel kapott többi komponenssel való korrelációja mellett érdekes lehet az is, hogy az eltérő módszerekkel imputált idősorokból számított komponensek mennyire korrelálnak egymással. Hipotézisünk szerint az eltérő adatpótlási módszerek után kapott idősorokból számított komponensek korrelációs mátrixa

egy permutációs mátrix, tehát nem lesz lényeges eltérés az egyes módszerek által szolgáltatott komponensek között, csupán a komponensek sorrendje véletlenszerű, tehát az egyhez közeli értékek nem feltétlenül a mátrix főátlójába kerülnek.

Amennyiben nagy korrelációs értékeket kapunk, további kérdés, hogy az adatpótlás elemenként és idősoronként mekkora eltérést generál a komponensek értékeiben.¹⁵ Ennek mérésére a következő mutatószámot használjuk:

$$0 \leq \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \frac{||s_{it}^{(a)}| - |s_{it}^{(b)}||}{|s_{it}^{(a)}| + |s_{it}^{(b)}|} \leq 1. \quad (3.4)$$

A korrelációs mátrixban szereplő értékek alapján azonban több esetben nem dönthető el egyértelműen, hogy melyik komponens felel meg a másik adatpótló eljárás által előállított idősorokkal számolt komponensnek, így azzal a feltételezéssel élünk, hogy a leginkább korreláló tagokat „összetartozónak” tekinthetjük. Ennek megfelelően azon két komponens közötti átlagos és elemenkénti eltérést számítjuk ki, amelyek között a legmagasabb a korreláció.

A képletben¹⁵ szereplő $s_{it}^{(a)}$ és $s_{it}^{(b)}$ az a és b adatpótlási módszerrel előállított idősorokra készített i -edik komponensének t időpontbeli sztenderdizált értékét jelölik. A sztenderdizálásra azért van szükség, mivel az ICA kimenetén az egyes komponensek bármilyen skalárszorosa megjelenhetnek, így szükség van a komponensek eltérő skálájának eliminálására. A számlálóban azért kell a komponenselemek abszolút értékét venni, mert ahogy azt korábban említettük, elképzelhető, hogy két azonosnak tekinthető komponens egymással negatívan korrelál. Mivel azonban a teljes eltéréssel normálunk, így előfordulhat, hogy akkor is 1 értéket kapunk, amikor a komponensek közötti korreláció nem tökéletesen negatív.

A mutatószám 0 és 1 között méri az egy elemre jutó különbség arányát, tehát azt mutatja meg, hogy a két módszer által előállított idősorra készített komponensek elemei között átlagosan mekkora arányú az eltérés. A mutatót komponensenként is kiszámítjuk¹⁶, ebben az esetben komponensenként fog 0 és 1 között mérni, vagyis a komponensenkénti átlagos hatást mutatja. Egy további mérőszámként fogjuk használni a komponenselemek eltéréseinek számát. A mutatónak akkor lesz haszna, ha az előbbi esetben nem kapunk nagy értékeket, mert például a hosszú idősor tompítja az eltérések hatását, és az egy elemre jutó átlagos különbség kicsi lesz. Azt azonban az eddig felsorolt mutatók alapján nem tudjuk meghatározni, hogy az adatpótlás hány komponenselemre volt hatással. Legyen emiatt egy további mérőszám a következő:

¹⁵Túl kicsi vagy egymáshoz közeli korrelációs értékek esetén a legfőbb problémát a komponensek azonosítása jelentené.

¹⁵A mutatónak az AAD: Average Absolute Difference nevet adtuk.

¹⁶AADC: Average Absolute Difference for a Component.

$$0 \leq \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \mathbf{1}_{\{d>0\}} \leq 1. \quad (3.5)$$

Ahol az $\mathbf{1}$ indexében szereplő d értéke a fenti mutatószám számlálójában szereplő különbség értéke¹⁷. A mutató ebben az esetben arányszámot jelöl, viszont érdekesebb komponensenként vizsgálni az értékét, ekkor 0 és T között mér. Hipotézisként kimondható, hogy komponensenként ez a szám a pótoló elemek számával kell, hogy megegyezzen (az összelemszámra pedig az összes pótoló adat számával, relatív esetben pedig a hiányzó adatok száma osztva T -vel).

A C.8.-C.10 táblázatokban szereplő eredmények alapján látható, hogy az adatpótlás komponensekre gyakorolt hatása nem jelentős. A várt eredményeket kapjuk, hogy ti. öt egyhez közeli érték szerepel a mátrixokban. A keresztkorrelációk nagysága több esetben is jelentős, de ez a C.5.-C.7. táblázatokban is ezt tapasztaltuk. További megemlíthető eredmény, hogy nincs kifejezett különbség a három korrelációs mátrix értékeit tekintve. Előzetes várakozásainknak ez éppen ellentmond, mivel a pótoló hozamok szintbeli különbsége a Konstans-Lineáris esetben kisebb, mint a másik két esetben.

A korrelációs értékek mellett kiszámítottuk az AAD és AADC értékeket az egyes imputálási eljárások komponenseire. A 3.3. táblázatban közölt eredmények alapján látható, hogy komponensenként az egyes elemek abszolút értékben hány százalékban térnek el egymástól. Látható, hogy az elemenkénti abszolút eltérés átlagosan a GARCH-Lineáris esetben a legkisebb, míg a Lineáris-Konstans eljárásokra a legnagyobb. A komponenselemek átlagosan az 22,8 - 31,5 %-ban térnek elnek egymástól, a GARCH-Lineáris esetben két-két olyan komponens is van, amelynek átlagos abszolút eltérése 10%-nál kisebb.

3.3. táblázat. A komponensenkénti és elemenkénti átlagos eltérések, különböző imputálási eljárások után adódó komponenseket összehasonlítva

Metódusok	GARCH - Konst	GARCH - Lin	Lin - Konst
AADC	0,1777	0,0864	0,4058
	0,0542	0,0369	0,3887
	0,3591	0,3604	0,3366
	0,2775	0,2869	0,4004
	0,3665	0,3689	0,0422
AAD	0,2470	0,2279	0,3147

Az eredmények a már korábban leírtakat erősítik meg, miszerint az adatpótlás jelentős hatással van a komponensek értékére, azok korreláltsága viszont nem változik. Kiszámítottuk a fent bemutatott másik mutató értékeit is. Ezek minden esetben azt mutatják, hogy minden komponenselem eltér a „hozzá tartozó” másik metódus idősoraira számított megfelelő komponenselemétől, tehát az indikátorfüggvény szummázott értéke éppen az adatsorok elemeinek számát

¹⁷Ha nem vagy túlságosan is jelentős hatást mutatunk ki ezen mutatóval, akkor érdemes az egy elemre jutó átlagos abszolút különbséggel helyettesíteni a nulla feltételt.

adja – emiatt ezt az eredményt táblázatos formában nem is közöljük.

A fenti eredményeket összegezve elmondhatjuk, hogy bár az adatróplás nincs hatással az egyes kapott komponensek korrelációjára, a különböző adatróplási módszerek után kapott komponensek különböznek. A sztenderdizált komponensek közötti átlagos abszolút eltérés mérhető, átlagosan 25%-os, továbbá az így előállított komponensek minden elemükben különböznek, tehát más-más komponensek állnak elő az adatróplásnak köszönhetően – azzal együtt, hogy azonos módszerrel előállt komponensek korrelációja nagyjából azonos. Az azonban továbbra is kérdés, hogy az eddigi vizsgálatok mennyire megbízhatóak, ugyanis eddig azzal az implicit feltételezéssel élünk, hogy az idősorokat befolyásoló látens változók száma megegyezik az idősorok számával.

3.2.2. Komponensek számának vizsgálata

Az idősorok mögött rejlő komponensek optimális számának meghatározásáról eddig még nem esett szó. (Monostori 2010) négy komponensre bontotta a magyar kötvényhozamokat, azonban azok korrelációja miatt nem biztos, hogy ténylegesen ennyi komponensből épül fel a hozam mint változó. Feltehető, hogy a mi esetünkben azért kaptunk különböző értékeket a komponensekre, mert túl sok látens változót feltételeztünk, így a továbbiakban ezt a kérdéskört járjuk körül.

Az idősorokat befolyásoló komponensszám meghatározásához szintén a korábban használt átlagos abszolút keresztkorrelációkat számítjuk ki, az eredmények a 3.4. táblázatban láthatóak.

3.4. táblázat. ICA komponensek átlagos keresztkorrelációi különböző komponensszámokra, mind az öt idősor felhasználásával, a neurális ICA-t használva

Komponensszám	2	3	4	5
Konstans	0,115	0,095	0,15	0,0276
Lineáris	0,14	0,071	0,035	0,0248
GARCH	0,116	0,114	0,068	0,0276

Megfigyelhető, hogy az imputálás módjától nem függnnek jelentősen az értékek, kiemelkedő különbséget nem láthatunk. Ugyanezeket a teszteket lefutttatva a FastICA algoritmussal azt tapasztaltuk, hogy a keresztkorrelációk minden esetben gyakorlatilag nullák (10^{-15} nagyságrendűek), ezért azokat táblázatosan nem is tüntetjük fel.

Ez felveti a kérdést, hogy különböző komponensszámok esetén mennyi információt őriznek meg a komponensek, azaz az előállított komponensekből milyen pontossággal állíthatóak vissza a kiinduló idősorok. Ezt az összefüggést a 3.5. táblázat szemlélteti, a visszaállítási hiba abszolútértékét átlagos relatív hibaként megadva – ez tulajdonképpen a MAPE (mean absolute percentage error.) mérőszám. A táblázatban szereplő adatok, csakúgy, mint a továbbiakban az összes teszt eredményei, a lineáris pótlási módszerrel előállított adatsorok használatával adódnak. Azért ezt az imputálási módszert használjuk, mert bár a különbségek csekélyek, a neurális

ICA ebben az esetben éri el a legjobb eredményeket.

3.5. táblázat. MAPE értékek különböző komponensszámok esetén Neurális ICA és FastICA algoritmus használata esetén

Komponensszám	2	3	4	5
MAPE Neurális ICA esetén [%]	88,8	80,3	63,7	0,0
MAPE FastICA esetén [%]	76	17,3	13,5	0,0

A táblázat egyrészt rávilágít, hogy alacsony komponensszámok esetén jelentős mennyiségű információvesztés következik be, a független komponensek bár függetlenek – különösen FastICA esetén, ahogy azt korábban említettük –, de az eredeti idősorokat olyan mértékben torzítják, hogy azok felismerhetetlenek lesznek. Ez a hatás szélsőségesen nagygyá válik, ha csupán két komponenset különítünk el. Emellett az is megfigyelhető, hogy mind az öt komponens használva a rekonstrukciós hiba elfogadhatóan kicsi.

A komponensszám csökkenésének a teljesítménnyel való összefüggéséhez meg kell jegyeznünk, hogy a rekonstrukciós hiba a neurális ICA esetén nem feltétlenül a módszer hibájából ered. Ahogy azt korábban láthattuk, az algoritmus az $\mathbf{A}$ keverési mátrix inverzét a $\mathbf{V}$ és a $\mathbf{W}$ mátrixok szorzataként határozza meg. $\mathbf{A}$ kiszámításához ennek az inverzét kellene venni, ha azonban $\mathbf{x}$ és $\mathbf{s}$ dimenziója nem egyezik, ez a mátrix nem invertálható. Ennek megfelelően a $\mathbf{W} \cdot \mathbf{V}$ szorzat pseudoinverze szolgáltatja az eredményt, mely egyrészt a pseudoinverz számításának a pontatlanságával terheli a későbbi komponenseket, másrészt nyilvánvalóan csak közelítő megoldást nyújt.

A komponensszámnak tehát nem szabad lényegesen eltérnie az idősorok számától, ha nem akarunk az idősorok jelentős torzulásával számolni. Kérdés azonban, hogy az idősorok akkor is jól függetleníthetők-e, ha az ICA-t nem az összes komponens felhasználásával végezzük, hanem csak néhányat választunk ki a jelek közül. A válasz igen, a különböző idősorok jól függetleníthetők páronként, hármasával, illetve négyesével is: a FastICA algoritmust használva 10^{-14} nagyságrendű keresztkorrelációkkal számolhatunk. Másfelől sejthető, hogy a komponensek elválasztása nem ugyanolyan könnyű feladat két komponens esetén, mint öt komponensre. A feladat nehézségét a neurális hálózattal megvalósított ICA segítségével érzékeltethetjük, mivel a lokális optimumban ragadás miatt kevés komponens esetén érzékelhetően rosszabb teljesítményt nyújt, ahogy azt a 3.6 táblázat is szemlélteti.

Az átlagos keresztkorrelációkat minden lehetséges kombinációra megvizsgáltuk minden komponensszám esetén, a táblázat ezek átlagait tartalmazza. Megfigyeléseink szerint a neurális architektúrát használva vannak olyan idősor-párok, illetve hármasok, melyek könnyen függetleníthetők, az öt komponens esetén kapott keresztkorrelációs értékek eléréséhez azonban már legalább négy, jól függetleníthető komponens kell. Ez azt mutatja, hogy az idősorok számának növekedésével azok egyre könnyebben lesznek független komponensekre bonthatóak.

3.6. táblázat. Komponensek korrelációi az idősorok számának függvényében, különböző komponensszámokra

Idősorok száma		2	3		4			5			
Komponensek száma		2	2	3	2	3	4	2	3	4	5
Neurális ICA	Átlagos keresztkorreláció	0,23	0,31	0,13	0,25	0,19	0,11	0,56	0,14	0,046	0,025
	Átlagos rekonstrukció	0	89	0	96	73	0	89	80	64	0,0
	MAPE [%]										
FastICA	Átlagos keresztkorreláció	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	Átlagos rekonstrukció	0	51	0	39	65	0	76	18	14	0
	MAPE [%]										

Bár a FastICA esetén ilyen állítást nem fogalmazhatunk meg, a rekonstrukciós hiba tekintetében hasonló tendenciát figyelhetünk meg mindkét módszerrel kapcsolatban. Megfigyelhető, hogy a rekonstrukciós hiba a megfigyelt idősorok és a független komponensek számának azonosága esetén elenyésző. Ez azt mutatja, hogy a rekonstrukciós hiba alapvetően a ki- és bemenetek számának különbségével van összefüggésben, és csak kisebb mértékben függ a konkrét dimenziótól.

Kijelenthető tehát, hogy annak ellenére, hogy a FastICA eredményei alapján a vizsgált országok adatsorai közül bármelyik kettő jól függetleníthető, a korábban ismertetett neurális hálózatot használva több ország együttes vizsgálatához nem elegendő két komponens elkülöníteni, mert ez az idősorok torzulását okozza. Emellett megállapíthatjuk azt is, hogy a neurális ICA bár egyszerűbb jelek esetén jól teljesít, ahogy azt a 2.5. ábra is mutatja, pénzügyi idősorok esetén használata csak jól kidolgozott optimizációs eljárással javasolható, mely képes a lokális minimumokból kiragadni az algoritmust.

Korábbi feltételezésünk tehát, miszerint a kötvényhozamok mögötti rejtett változók számát az ICA által függetlenített komponensek korrelációjával meghatározhatjuk, nem bizonyult helyesnek, hiszen az ICA idősoraink közül tetszőleges két jelet képes jól függetleníteni. Emiatt a továbbiakban az összes vizsgálatot öt komponensre fogjuk elvégezni azzal a feltételezéssel élve, hogy így a lehető legtöbb információt tartalmazó halmazt választjuk.

3.2.3. Adatrobosztuság vizsgálata

A komponensek további vizsgálata során elsődleges feladatnak azt tekintjük, hogy a komponensek érzékenységét vizsgáljuk. Az első fejezetben ismertetett irodalmak nem foglalkoztak azzal a kérdéssel, hogy az ICA milyen körülmények között teljesít jobban a PCA-nál. Nem vizsgálták azt a kérdést sem, hogy a módszer egyáltalán alkalmas-e pénzügyi idősorok függetlenítésére, illetve, ha igen, a kapott komponensek különböző mesterségesen generált sokkokra mennyire rezisztensek. A robusztuság-vizsgálatokat két fő részre bontjuk. Egyrészt a komponenselemek sokkokra való érzékenységét (adatrobosztuság), másrészt a megfigyelt idősor hosszának csökkentéséből fakadó változást (elemszám-robusztuság) vizsgáljuk.

A komponensek mesterségesen generált sokkokra való érzékenységének tesztelését alap-

vetően három dimenzióra oszthatjuk a sokkolás módszere, a sokkolt elemek kiválasztása és a sokkolt elemek száma alapján.

I. Módszer:

1. Konstans sokkok: minden idősorban a kiválasztott elemet ekkora mértékben térítünk el, azaz sokkolunk (pl: 1%, 5%, 10%).
2. Azonos véletlen sokkok: véletlenül választott sokkok valamilyen intervallumon belül, de az azonos sorszámú elemeket ugyanolyan mértékben térítjük el.
3. Elemenkénti véletlen sokkok: véletlenül választott sokkok elemenként, tehát ugyanazokat az elemeket választjuk ki, de mindegyikre eltérő mértékű sokkot alkalmazunk.

II. Elemek kiválasztása:

1. Osztópontos kiválasztás: a teljes idősort a sokkolni kívánt elemek számának függvényében egyenlő hosszúságú intervallumokra osztjuk és az osztópontokat sokkoljuk.
2. Véletlenszerű kiválasztás: 1 és a maximális elemszám, 3927 között véletlenül kiválasztott elemek. Az elemek kiválasztásánál, minden elemszámba újra választjuk az elemeket, de az eltérő mértékű sokkolásnál ugyanazokat az elemeket sokkoljuk, csak különböző mértékben.

III. Elemszám:

1. 1-10 (a lépésszám 1)
2. 10-100 (a lépésszám 10)
3. 100-1000 (a lépésszám 100)

A hatások méréséhez először a nem sokkolt idősorokra állítjuk elő a komponenseket, a későbbiekben ezekhez fogjuk hasonlítani a többi futtatás eredményét. A sokkolt idősorok kiszámítása után a sokkolt és nem sokkolt idősorok alapján előállított komponensek korrelációs mátrixát képezzük a komponensek azonosításához. A hatások mérése itt is két lépcsős. Egyrészt azt vizsgáljuk, hogy a sokkolt és a nem sokkolt idősorok által generált komponensek közötti korreláció hogyan változik, másrészt a fenti mutatószámokhoz hasonló mértékek segítségével számszerűsítjük a komponensek robusztusságát. A teszteket 49200 futás eredményeiből átlagoltuk.

Neurális ICA architektúra

A 3.7. táblázat tartalmazza a neurális ICA-val végzett tesztek eredményeit. A táblázat első sorában szerepel a sokkolási együttható, tehát ekkora volt a sokk mértéke. A táblázatban szereplő értékek az előnn leírtaknak megfelelően a sokkolt idősorokkal kapott, és az eredeti komponensek átlagos keresztkorrelációi. Eredményeink elsősorban a neurális architektúra egy jellegzetességére világítanak rá, nevezetesen arra, hogy érzékeny a lokális minimumokra.

3.7. táblázat. Lineáris módszerrel pótoltt idősorok komponenseinek robusztussága konstans sokkolás esetén különböző elemszámokra, véletlen és szisztematikus elemkiválasztással

Véletlenszerű elemkiválasztás				Szisztematikus elemkiválasztás			
Sokk mértéke	1	5	10	Sokk mértéke	1%	5%	10%
1	0,0298	0,1479	0,0297	1	0,0298	0,0298	0,1481
10	0,0296	0,1477	0,1462	10	0,0298	0,1498	0,0316
50	0,0299	0,0308	0,0345	50	0,1481	0,0298	0,1479
100	0,0296	0,1440	0,1408	100	0,0298	0,1481	0,0300
500	0,1514	0,0329	0,0437	500	0,0298	0,1478	0,0301
1000	0,1481	0,0522	0,0665	1000	0,0298	0,0297	0,1464

Amint az látható, a táblázatban az értékek két érték köré csoportosulnak, 0,15 és 0,03 köré. Ez a két érték független komponensek két halmazának felel meg, azaz két ICA megoldásnak, melyek többé-kevésbé véletlenszerűen állnak elő a sokkolt elemek számának, és a sokkolás mértékének függvényében. A két megoldás a (2.24) egyenletben szereplő költségfüggvény két lokális minimuma, melyek valószínűleg egymáshoz közel helyezkednek el, hiszen már egy adat kis mértékű sokkolása is okozhat eredménybeli eltérést, ahogy azt a táblázat első sora is mutatja.

A sokkolásra való érzékenységgel kapcsolatban a 3.7. táblázat véletlenszerű és szisztematikus sokkokra vonatkozó részét összehasonlítva megfigyelhető, hogy szisztematikus sokkok esetén a független komponensek közötti korrelációk nem változnak érdemben, míg véletlenszerű sokkok esetén azok szórása megnő – különösen nagy sokkolási elemszámoknál. Ez azt jelenti, hogy szisztematikus sokkok esetén a sokkok elkülönítése egy adott komponensbe történik, így a komponensek egymás közötti viszonyai lényegében nem változnak. Véletlen sokkok esetén a sokkolás feltehetően nehezebben különíthető el az idősorokat terhelő zaj jellegű komponensből vagy komponensekből, így ennek elkülönítése az ICA számára is nehezebb feladat. A véletlen sokkok eredményei nem térnek el a konstans sokkokétól, emiatt azokat táblázatos formában nem közöljük.

FastICA algoritmus

A FastICA algoritmus eredményei a C.11.-C.13. táblázatokban találhatóak. A neurális ICA esetén megfigyelt elemkiválasztási módra való érzékenység a FastICA esetén is fennáll. Érdekes eredmény azonban, hogy szisztematikus sokkok esetén, ha a sokk mértéke elegendően nagy, az eredetileg kapott és a sokkolt komponensek közötti átlagos korreláció három nagyságrenddel

nagyobb lesz, elegendően nagyszámú elem sokkolása esetén viszont visszaesik. A jelenség okát, illetve annak magyarázatát, hogy miért leginkább konstans sokkolás esetén fordul elő, nem találtuk meg.

A táblázatokat vizsgálva azt az általános megállapítást tehetjük, hogy 1%-os sokkolás, illetve elegendően kis elemszámú 5%-os sokkolás sem rontja az algoritmus teljesítőképességét számottevően, a komponensek keresztkorrelációja nem jelentős. Ennél nagyobb sokkok, vagy nagyobb elemszámok esetén azonban a véletlenszerű és a szisztematikus sokkok eltérő tendenciát mutatnak. A véletlenszerű sokkok a sokkolás módjától függetlenül torzulást okoznak, 5%-os sokkok esetén már 50-100 elem sokkolása is jelentő keresztkorrelációkat okoz, melyek átlagos értéke az eltérített elemek számának növekedésével nő.

Szisztematikus sokkok esetén viszont az eredeti és a sokkolt idősorokból számolt komponensek keresztkorrelációi nagyságrendekkel nőnek, de csak nagy mértékű sokkok alkalmazásakor. Ez megerősíti a korábbi tapasztalatokat, hogy ti. szisztematikus változtatások esetén a komponensre bontás könnyebb feladat.

A korrelációs mátrixokból kiolvasott eredményeket a függelékben található C.4.-C.5. ábrák még szemléletesebbé teszik. Látható, hogy az elemkiválasztás módja a kiválasztott elemek számának növelésével egyre kisebb mértékben játszik szerepet, főként a nagyobb sokkok esetén. Látható továbbá, hogy a sokk mértékének növelésével a két kiválasztási metódus egymáshoz konvergál. Tehát, ahogy egyre több zajt keverünk az idősorokhoz, annál kevésbé számít, hogy milyen módszerrel tettük.

Az átlagos keresztkorrelációk elemzése mellett fontos, hogy lássuk, mekkora hatást váltanak ki a komponensekben a sokkok. A függelék C.6 ábrája mutatja az 1 és 10%-os konstans sokkolás hatását a komponensek átlagos abszolút eltérésére. Mivel a tendencia hasonló a másik két sokkolási eljárás esetén, így csak ezt az ábrát közöljük. Az ábrán látható, hogy 1%-os sokkolás esetén a szisztematikus elemkiválasztás rosszul reagál bizonyos számú elemenként. Ennek tudható be az átlagos keresztkorrelációs érték „ugrálása” is. Véletlen kiválasztás esetén is tapasztalunk hasonlót, azonban a kiugrások száma kis sokkok esetén alacsonyabb. Meg kell azonban jegyezni, hogy a kiugrások mértéke 40 elemszám után a véletlen kiválasztási eljárással nagyobb, mint a másik esetben. 10%-os sokkoknál a kiugrások mértékében fellépő különbség sokkal hangsúlyosabbá válik. Az elemenkénti átlagos eltérésekkel kapcsolatos eredményeket a C.14.-C.16. táblázatok számszerűsítve is mutatják, azonban a komponensenkénti értékek ezekben nem szerepelnek.

Az előzőekhez hasonlóan kiszámítottuk az egyes sokkolások előtt és után kapott komponensek közötti átlagos maximális korrelációkat extrém mértékű sokkok esetén is, melyeket a 3.8. táblázat tartalmaz. Megfigyelhető, hogy már 10 elem 1%-os eltérítése esetén sincs két azonos idősor a komponensek között, a sokk mértékének növekedésével pedig a korrelációk egyre romlanak.

3.8. táblázat. Komponensek korrelációs mátrixának átlagos legnagyobb elemei extrém sokkok esetén különböző elemszámokra, véletlenszerű elemkiválasztással, FastICA algoritmus használatával

Sokk mértéke	0,00001%	1%	100%
1	1,0000	1,0000	0,8745
10	1,0000	0,9996	0,7814
50	1,0000	0,8144	0,6287
100	1,0000	0,8116	0,6131
500	1,0000	0,7977	0,5210
1000	1,0000	0,7738	0,4483

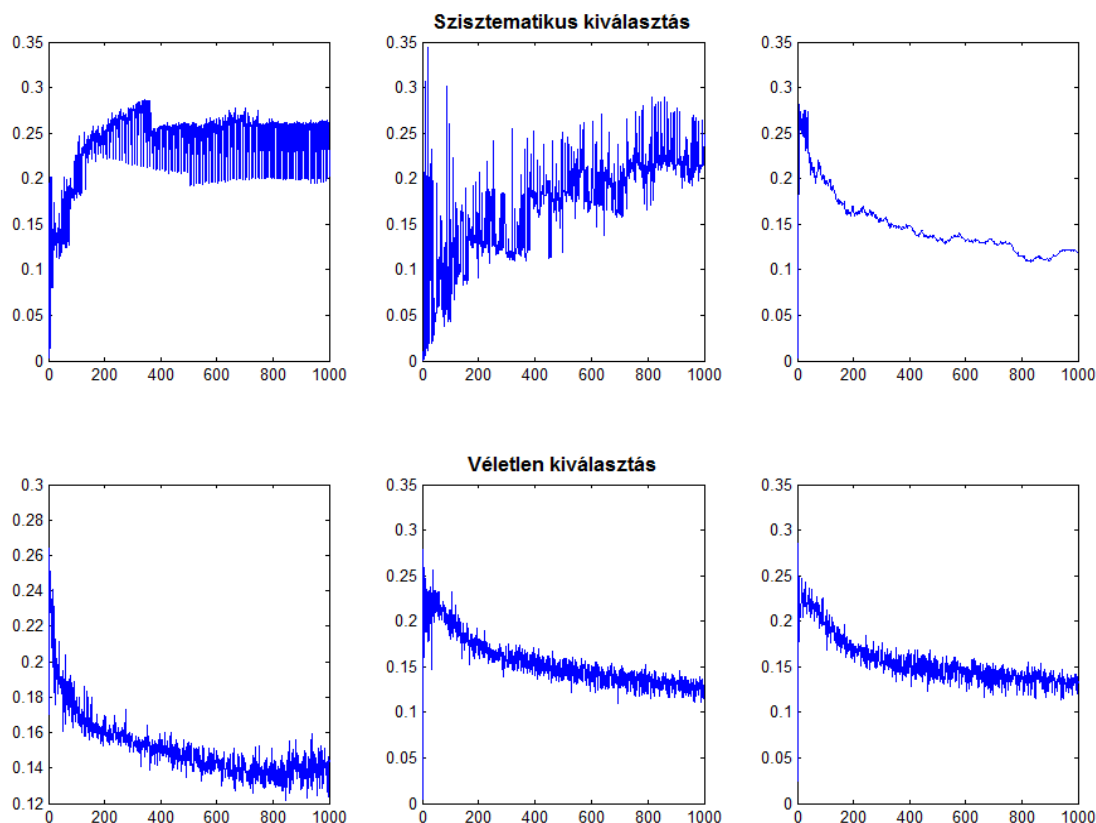
A „zajosítás” hatását a 3.3 ábra mutatja az adatok nagymértékű eltérítése esetén. Minden esetben maximum 100%-kal sokkoltuk az adatokat. Látható, hogy az elemszám növelésével csak szisztematikus elemkiválasztás esetén, konstans és azonos véletlen mértékű sokkolással nő a keresztkorrelációk átlaga. A másik négy esetben látható értékek azonban némiképp meglepőek. Az eredményeket azzal tudjuk magyarázni, hogy a szisztematikus elemkiválasztás, de véletlenszerű nagy sokkok esetén a módszer képes a véletlenszerűséget nagyon jól elkülöníteni, míg az alsó sorban a véletlenszerű kiválasztás kever a folyamatokhoz további zajt, ezért változnak a keresztkorrelációs értékek egy sávban. Ami talán a legmeglepőbb, hogy a keresztkorrelációs értékek csökkennek, ha növeljük a véletlenszerűséget a rendszerben és az utolsó oszlopbeli ábrákon látható sorozatok látszólag egy értékhez konvergálnak.

Kijelenthető tehát, hogy kis zavaró hatásokra mind a neurális ICA, mind a FasICA érzéketlen. Nagyobb sokkok esetén viszont a komponensek elkezdenek torzulni, különösen, ha a sokkolt elemek, vagy a sokkolás módja véletlenszerű. Szisztematikus sokkok esetén ez a hatás kevésbé érvényesül. Ez a jelenség magyarázhatja az adatkótlási módszerekre való érzéketlenséget is, hiszen az adatkótlási módszerek különbözősége kisszámú elem sokkolásaként fogható fel, ráadásul ezek az elemek – az évenkénti ismétlődések miatt – periodikusan fordulnak elő. Jelentősebb sokkolások esetén ez a teljesítmény inkább a sokkolt elemek kiválasztásától és a sokkolás módjától függ, mivel a szisztematikus konstans nagy sokkok nem szeparálható korrelációt visznek a rendszerbe, a véletlenszerűséget növelve azonban a módszer egyre jobban teljesít.

3.2.4. Elemszám-robustusság vizsgálata

A következőkben a módszer elemszám-változásra való érzékenységet vizsgáljuk. Az idősorok elejéről, végéről, illetve mindkét végéről elkezdjük elhagyni az adatokat, majd az így kapott komponenseket az eredetileg számolt komponensekkel hasonlítjuk össze.

Módszertanilag az előzőeknek megfelelően járunk el, tehát a komponensek standardizáltjaival fogunk dolgozni. Minden rövidebb komponenst azzal az eredeti komponenssel azonosítunk, melyre a két jel egyidejű abszolút korrelációja maximális, természetesen az eredeti komponenseknek csak a megfelelően rövidített szakaszait véve figyelembe. Ezt követően kiszámoljuk



3.3. ábra. Komponensek közötti átlagos keresztkorreláció alakulása 100%-os sokkolás esetén. Az első oszlopban a változtatás módja konstans, a másodikban azonos véletlen, a harmadikban pedig elemenkénti véletlen.

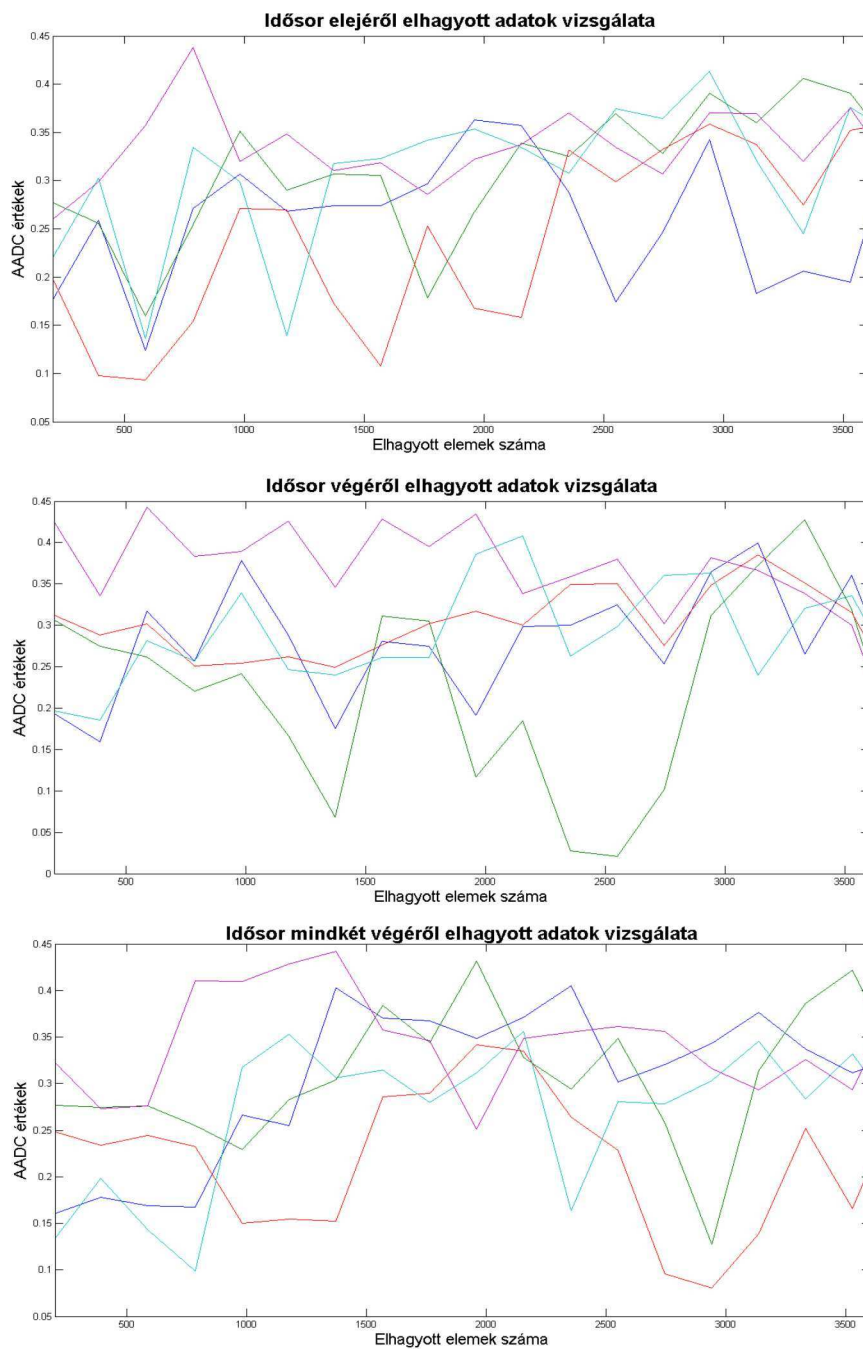
a két komponens standardizáltjának az AADC-mérőszámát. Várakozásaink szerint az AADC mérőszámokban egészen addig nem lesz jelentős különbség, míg az elemszám csökkenése nem teszi nehezebbé az ICA probléma megoldását. Az eredményeket a 3.4. és a 3.5. ábra szemlélteti.

Amint az látható, mindhárom adatelhagyási módszer esetén hasonlóak az eredmények, a komponensek különbségeit nem befolyásolja számottevően az adatok elhagyása, egyik módszer használata esetén sem – a különbségek azonban nem elenyészőek. Ami mindenképpen megjegyzendő, hogy a komponensek között kevés adat elhagyásakor minden esetben felfedezhető egy-két a teljes idősorra számítottakhoz nagyon hasonló komponens. Ezek a komponensek az adatok elhagyásával folyamatosan torzulnak, eltérésük egyre inkább megközelíti a többi komponensét.

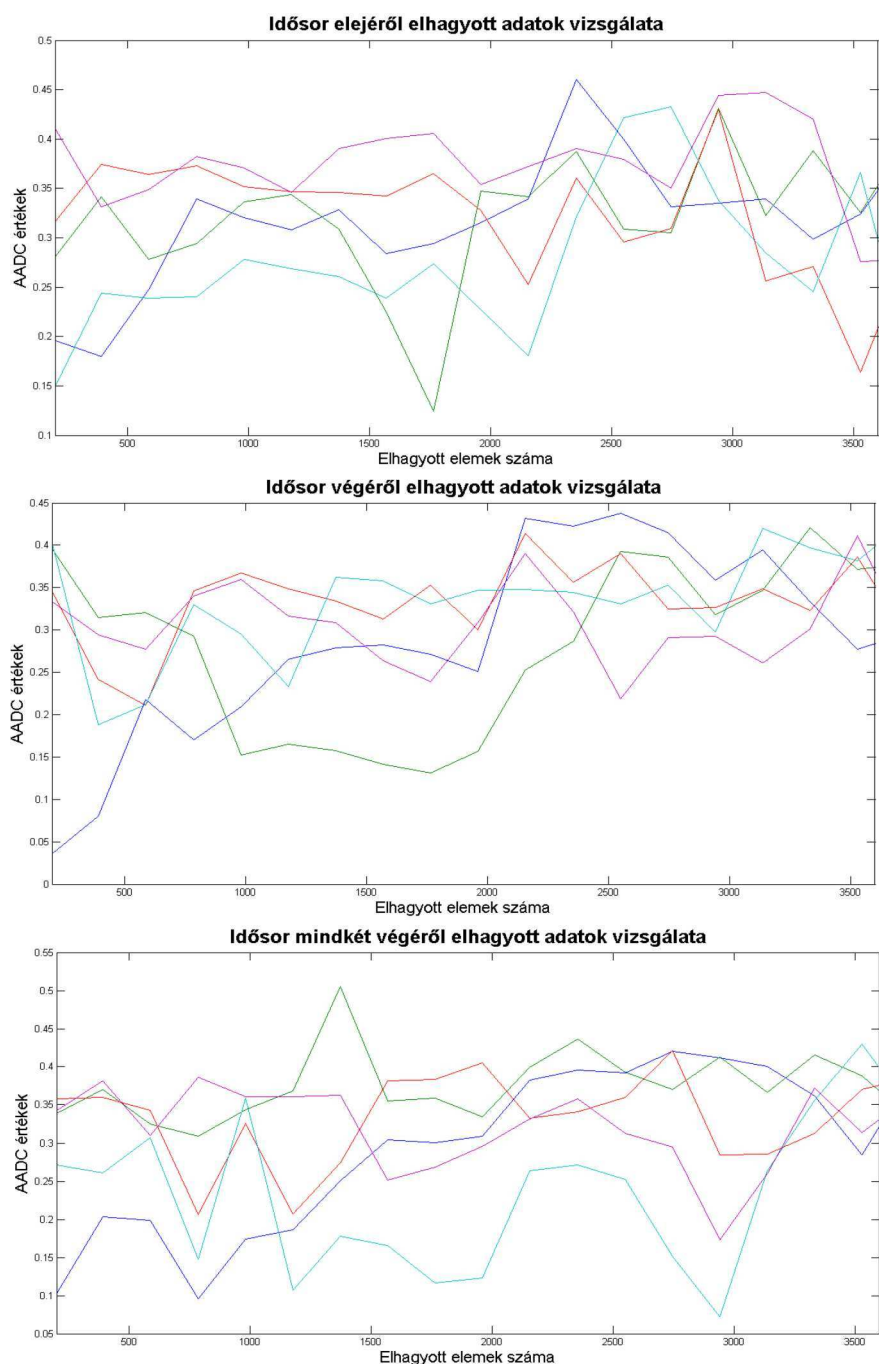
Érdekes eredménynek tartjuk azonban, hogy szinte minden esetben található legalább egy olyan komponens, melynek a teljes idősorra számolt komponensekre való illeszkedése javul az elhagyott elemek számának növelésével.

Kijelenthetjük tehát, hogy adatok elhagyása az idősorokból megváltoztatja a független komponenseket, és ez a változás összességében véve annál nagyobb mértékű, minél több adatot hagyunk el, függetlenül attól, hogy az adatokat honnan hagytuk el. Ennek fontos következménye, hogy az idősor egy szakaszára számított független komponensek eltérhetnek a teljes idősorra

számított komponensektől, ahogy azt a következőkben is látni fogjuk.



3.4. ábra. A neurális ICA architektúra elemszám-robosztusságának vizsgálatának eredményei, különböző adatelhagyási módszerekkel.



3.5. ábra. A FastICA algoritmus elemszám-robosztusságának vizsgálatának eredményei, különböző adatelhagyási módszerekkel.

3.2.5. Pénzügyi sokkok hatásának vizsgálata

A tesztek utolsó csoportja a pénzügyi válságok független komponensekre gyakorolt hatásának a vizsgálata. A vizsgálatba három időszakot vonunk be: az 1998-as orosz válságot, a 2000-es dot-com válságot, valamint a 2008-2009-es gazdasági világválságot (a továbbiakban subprime válság). Az ábrák leolvasásának megkönnyítése céljából pirossal meg fogjuk jelölni a válságok kezdeti időpontját, illetve amennyiben a kezdeti időpont nem egyértelmű, a válság hatásá-

nak megjelenési pontját az idősorokon. A tesztekhez a FastICA algoritmust használjuk, annak pénzügyi területen való – a korábbi tapasztalatok szerint – jobb alkalmazhatósága miatt.

Az orosz válság

Az orosz válság időszakára korlátozódó, 1998.06.18-1999.07.15. közötti idősorok, valamint a különböző módszerekkel elkülönített komponensek a C.7.-C.9. ábrákon láthatóak, pirossal bejelölve 1998.08.01.-e. Amint látható, a legtöbb idősorban a válság a kötvényhozamok esésében nyilvánul meg, a magyar hozamok ugyanakkor átmeneti emelkedést mutatnak.

A FastICA algoritmussal nyert komponensek közül a válság kezdetekor csak kettőben vehe-tő észre szignifikáns trendbeli eltérés, a többi független komponens csak körül-belül egy hónap múltán követi a változásokat. A neurális algoritmussal kapott komponensek esetén hirtelen változások három komponens esetén is megjelennek, míg a többi az előzőeknek megfelelően egy hónappal később mutat eltérést.

Amennyiben a nem ablakozott, 3.2. ábrán látható komponensekkel hasonlítjuk össze az eredményeket, hasonló megállapításokat tehetünk: három komponens mutat hirtelen változást a válság kezdetekor, a többi csak egy hónappal később változik. Megemlítendő, hogy míg a teljes idősorra elvégzett ICA komponensei között jellemzően szerepel egy nulla kis környezetében mozgó, zaj jellegű komponens, az ablakozott idősorok esetén ilyen nem látható.

A dot-com válság

A dot-com válság 1999.10.21-2000.12.14. közötti szakaszának idősorai, valamint a különböző módszerekkel elkülönített komponensek a C.10.-C.12. ábrákon láthatóak, pirossal bejelölve 2000.03.10.-e. Amint látható, a válság nem gyakorol szélsőséges hatást az idősorokra, ennek ellenére a független komponensek közül három mindkét módszer esetén hirtelen változásokat, vagy trendbeli töréspontot tartalmaz a válság kezdetének idején. Ez az eredmény azért is érdekes, mivel a teljes idősor használatával kapott komponensekben ilyen tendencia nem különíthető el, a válság egyértelműen csak egy komponensben jelenik meg.

További érdekesség a komponensek ugrásszerű változása egy hónappal a válság kezdete után, mely a teljes idősor független komponenseiben hasonlóképpen megjelenik.

A subprime válság

A subprime válság idősorai a 2008.03.27-2009.05.21. közötti időszakot ölelik fel, a C.13.-C.15. ábrákon pirossal bejelölve 2008.09.15.-e látható¹⁷. Amint látható, az összes idősoron csökkenő trendeket okozott a válság, a magyart kivéve, melyen átmeneti növekedés volt tapasztalható.

¹⁷A Lehman Brothers Holdings Inc. ezen a napon kért csődvédelmet

Érdekes eredmény, hogy mindkét módszer esetén a trendbeli változások csak egy-egy idő-sorban jelennek meg, a többi komponens hirtelen változásokat tartalmaz a válság kezdetekor, illetve valamivel később. További érdekesség, hogy a teljes idősorral számolt független komponensek között van olyan, amelyik még ez előtt a dátum előtt az addigival ellentétes mozgásba kezd, míg az ablakozott komponenseken ilyen viselkedés nem látható.

Megjegyzendő továbbá, hogy ez az ablak a legnehezebben függetleníthető, a neurális ICA algoritmus által függetlenített komponensek között van 0,7-es korrelációjú pár – a FastICA ebben az esetben is kiváló eredményt nyújt.

3.3. Eredmények összefoglalása

Az eredmények alapján megállapíthatjuk, hogy az ICA rendkívül hatékonyan képes pénzügyi idősorok függetlenítésére. Tesztjeink alapján a módszer kevésbé érzékeny a kis mértékű zajra, illetve arra is, hogy amennyiben hiányzó adatokat tartalmazó idősoraink vannak, azokat milyen módszerrel imputáljuk.

Amennyiben az ICA kimenetének az átlagos keresztkorrelációt vesszük, azaz a célunk az idősorok függetlenítése, úgy a módszer rendkívül robusztus. A vizsgált idősorok esetén az összes futás után alig mérhető keresztkorrelációkat kaptunk. Ezzel kapcsolatban az is megállapítható, hogy a neurális ICA architektúra bár egyszerűbb, jellemzően műszaki területen előforduló jelekre kiváló eredményekkel alkalmazható, pénzügyi idősorok esetén kevésbé kielégítő eredményt nyújt. Ennek oka nyilvánvalóan a költségfüggvény bonyolultságában rejlik, úgy is fogalmazhatunk, hogy az ICA probléma pénzügyi idősorok esetén egy nagybonyolultságú optimalizációs probléma.

Ha viszont az eljárás kimenetének magukat a független komponenseket, mint újabb idősorokat tekintjük – például mert ezek elméleti azonosítása a célunk –, akkor az algoritmus teljesítménye árnyaltabb. Egyrészt az ICA-modell sajátosságai miatt a kapott komponensek tetszőlegesen skálázva, tetszőleges sorrendben jelenhetnek meg, ami megnehezíti az idősorok elemzését. Másrészt két futtatás között kapott komponensek – ha a kiindulási idősorok akár csak minimálisan is eltérnek – nem egyeznek meg, annak ellenére, hogy korrelációjuk a korábban kapott komponensek egyikével igen magas, a többivel pedig alig mérhető. A keresztkorrelációk az idősor egyre nagyobb mértékű változtatásával egyre nőnek, ahogy azt az adatsorok sokkolásával érzékeltettük, bár megjegyzendő, hogy szisztematikus sokkok esetén ez a hatás kevésbé jelentős, ha sokkok mértéke normális szintű. A nagymértékű konstans sokkok nem szeparálható korrelációt visznek a rendszerbe, azonban a sokkok véletlenszerűségének növelésével a szisztematikus eltéréseket a módszer jobban szeparálja.

Elemek elhagyása az idősorból szintén problematikus, ekkor ugyanis a független komponensek nagymértékben megváltoznak, illetve elegendően nagyszámú komponens elhagyásával a függetlenségük sem biztosítható, korrelációjuk megnő. Emiatt ablakozott idősorok vizsgálata

a teljes idősorhoz képest eltérő eredményt nyújt, bár megjegyzendő, hogy sok hasonló tulajdonság figyelhető meg a teljes idősor használatával számított, illetve az ablakozott eset komponensei között.

4. Összefoglalás

A dolgozatban bemutattuk a független komponens analízist, illetve annak pénzügyi adatsorokra való alkalmazásának lehetőségeit.

A módszer elméleti leírása, valamint megvalósítási lehetőségeinek ismertetése után egy általunk választott algoritmust megvalósítottunk, működését különböző egyszerű esetekben ellenőriztük. Emellett használtunk egy szabadon hozzáférhető, széleskörűen használt algoritmust is.

Az ICA gazdasági területen való alkalmazhatóságának vizsgálatát az imputálás hatásának feltárásával kezdtük. A rendelkezésünkre álló, hiányos adatsorainkat különböző módszerekkel imputáltuk, majd meghatároztuk az így nyert komponensek korrelációit, illetve hasonlóságukat általunk definiált mérőszámok alapján. A tesztek során megállapítottuk, hogy az adatpótlási módszer nincs jelentős hatással az ICA eredményeire.

Ezután az ICA által függetlenített idősorok tulajdonságait vizsgáltuk a komponensek és idősorok számának függvényében. Megállapítottuk, hogy a komponensek számának az ICA teljesítménye szempontjából csak neurális ICA esetén van hatása. Megvizsgáltuk a komponensszám rekonstrukciós hibára gyakorolt hatását is, azt állapítottuk meg, hogy a kiindulási adatsorok rekonstrukciója csak akkor tökéletes, ha a komponensek és az idősorok száma megegyezik. Ellenkező esetben az ICA rekonstrukciós teljesítménye romlik, különösen a neurális architektúra esetében. Mindebből következően pénzügyi elemzésekre a FastICA algoritmust tartjuk alkalmasabbnak.

A módszer robusztusságának vizsgálatával folytattuk az elemzést, kitérve mind az adatrobusztusság, mind az elemszám-robusztusság vizsgálatára. Adatrobusztusság tekintetében megállapítottuk, hogy a módszer a komponensek torzulásának tekintetében nem érzékeny az idősorok kismértékű sokkolására, illetve szisztematikus sokkok esetén nagyobb mértékűre sem. Másrészt nagyobb méretű véletlen sokkokat alkalmazva az elkülönített független komponensek a sokkolás nélküli idősorból kapott komponensekkel való keresztkorrelációi megnőnek, illetve azok abszolút eltérése is nagyobb mértékű. Vizsgálataink alapján azonban a keresztkorrelációk nem nőnek tetszőlegesen nagyra, a sokkolt idősorokból számított és az eredeti független komponensek átlagos keresztkorrelációja a sokkolt elemek számának növekedésével egy konkrét értékhez tart.

Az ICA elemszám-változásra való érzékenységevel kapcsolatban elmondható, hogy a komponensek átlagos eltérése a teljes idősorra számoltaktól ugyanabban a nagyságrendben marad, de értéke ingadozik, egyes esetekben a teljes idősorra számoltakhoz nagyon hasonló komponenseket is kaphatunk. Általánosságban azonban elmondható, hogy a különböző hosszúságú idősorokból szétválasztott komponensek nem egyeznek meg egymással, különösen, ha az el-

meszám az ICA hatékonyságának romlását is okozza.

Végezetül az idősorok ablakozásának hatását vizsgáltuk. A korábbiakkal összhangban megállapítottuk, hogy az ablakozott idősorok független komponensei nem egyeznek meg teljesen a teljes idősorra számítottakkal, de azokban azonos tendenciák fellelhetők. Megállapítottuk továbbá, hogy a válságok a független komponensekre is hatással vannak, de változó, hogy milyen mértékben, és hány idősorban okoznak eltéréseket.

Megállapíthatjuk tehát, hogy a FastICA algoritmus kiváló eszköz pénzügyi komponensek függetlenítésére, annak ellenére, hogy az eredményül kapott komponensek feldolgozása nem feltétlenül triviális. A dolgozatban definiáltunk egy lehetséges módszertant a komponensek összehasonlítására, és azok torzulásának elemzésére.

Hivatkozások

- Back, A. D. & Weigend, A. S. (1997), 'A first application of independent component analysis to extracting structure from stock returns', *International Journal of Neural Systems* **8**(4), 473–484.
- Bellini, F. & Salinelli, E. (2003), 'Independent component analysis and immunization: An explanatory study', *International Journal of Theoretical and Applied Finance* pp. 721–738.
- Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**, 307–327.
- Boukong, S., Toch, B., Saad, D. & Lowe, D. (2003), 'Ica for watermarking digital images', *Journal of Machine Learning Research* **4**, 1471–1498.
- Cha, S.-M. & Chan, L.-W. (2000), 'Applying independent component analysis to factor model in finance', *Springer* pp. 538–544.
- 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* **47**, 115–125.
- Cont, R. (2001), 'Empirical properties of asset returns: stylized facts and statistical issues', *Quantitative Finance* **1**, 223–236.
- Genomic Science Program (2012), 'Human genome 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 Financiera* **21**, 37–46.
- Gruber, P., Theis, F., Stadlthanner, K., Lang, E., Tome, A. & Teixeira, A. (2004), Denoising using local ica and kernel-pca, in 'IEEE International Joint Conference on Neural Networks', Vol. 3, pp. 2071–2076.
- Hajdú, O. (2003), *Többváltozós statisztikai számítások*, KSH, Budapest.
- Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.
- Horváth, G., Altrichter, M., Pataki, B., Strausz, G., Takács, G. & Vaylon, J. (2006), *Neurális hálózatok*, Panem, Budapest.
- Hunyadi, L. (2009), *Statisztikai következtetéselemélet közgazdászoknak*, Aula, 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* **13**(4-5), 411–430.
- Kaplan, S. T. & Ulrych, T. J. (2003), Blind deconvolution and ica with a banded mixing matrix, in '4th International Symposium on Independent Component Analysis and Blind Signal Separation', pp. 223–228.
- Kim, J., Choi, J., Yi, J. & Turk, M. (2005), 'Effective representation using ica for face recognition robust to local distortion and partial occlusion', *Journal of Machine Learning Research* **27**(12), 1977–1981.
- Lee, T.-W., Lewicki, M. S., Girolami, M. & Sejnowski, T. J. (1999), 'Blind source separation of more sources than mixtures using overcomplete representations', *IEEE Signal Processing Letters* **6**(4), 87–90.
- Li, H. & Adali, T. (2008), 'A class of complex ica algorithms based on the kurtosis cost function', *IEEE Transactions on Neural Networks* **19**(3), 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-Bizzozero, N. I. & Calhoun, V. (2007), Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica', *Human Brain Mapping* **30**(1), 241–255.
- Medvegyev, P. (2002), *Fejezetek a matematikai analízisből és a valószínűségszámításból*, Aula, Budapest.
- Monostori, Z. (2010), 'Az ötéves, magyar, szuverén, fix kamatozású forintkötvények hozamainak dekompozíciója - kockázati prémiumok bemutatása és becslése', *Budapesti Corvinus Egyetem Szakdolgozat*.
- Plumbey, M. (1993), A hebbian/anti-hebbian network which optimizes information capacity by orthonormalizing the principal subspace, in 'Proc. IEEE Conference on Artificial Neural Networks', Vol. 3, pp. 86–90.
- 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* **E88**(7), 1683–1692.

- Ross, S. (1976), 'The arbitrage theory of capital asset pricing', *Journal of Economic Theory* **13**, 341–360.
- Wang, L. Y. & Karhunen, J. (1996), 'A unified neural bigradient algorithm for robust pca and mca', *International Journal of Neural Systems* **7**(1), 53–67.
- Wu, E. H. & Yu, P. L. H. (2005), *Volatility Modelling of Multivariate Financial Time Series*, Springer, Berlin.

Függelék

A. Transzformáció sűrűsége

Legyenek $\underline{X}$, $\underline{Y}$ valószínűségi vektorváltozók, melyek között fennáll a következő összefüggés:

$$\underline{Y} = g(\underline{X}), \quad (\text{A.1})$$

amelyre létezik és egyetlen az alábbi inverz leképezés:

$$\underline{X} = g^{-1}(\underline{Y}). \quad (\text{A.2})$$

Megmutatható, hogy $\underline{Y}$ változó $p_Y(\mathbf{y})$ valószínűségi sűrűségfüggvénye a $p_X(\mathbf{x})$ sűrűségből a következőképpen számítható ki:

$$p_Y(\mathbf{y}) = \frac{1}{|\det J\mathbf{g}(\mathbf{g}^{-1}(\mathbf{y}))|} p_X(\mathbf{g}^{-1}(\mathbf{y})). \quad (\text{A.3})$$

Legyen $J\mathbf{g}$ a Jacobi-mátrix

$$J\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x})}{\partial x_n} & \frac{\partial g_2(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_n} \end{bmatrix},$$

ahol $g_j(\mathbf{x})$ a $\mathbf{g}(\mathbf{x})$ vektorfüggvény j -edik komponense. Abban az esetben, ha az (A.1) lineáris, nonsinguláris transzformáció, vagyis $\underline{Y} = \mathbf{A}\underline{X}$ és $\underline{X} = \mathbf{A}^{-1}\underline{Y}$, az (A.3) egyenlőség az alábbira egyszerűsödik:

$$p_Y(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} p_X(\mathbf{A}^{-1}\mathbf{y}). \quad (\text{A.4})$$

Ha tehát $\underline{X}$ az (A.4) egyenletben többváltozós normális eloszlású változót jelöl, akkor $\underline{Y}$ is többváltozós normális eloszlású lesz.

A fenti gondolatmenetet folytatva egy transzformált valószínűségi változó entrópiáját is meghatározhatjuk, az eddigi jelöléseket megtartva a következő módon:

$$H(\underline{Y}) = -\mathbb{E} \log p_{\underline{Y}}(\mathbf{y}) = H(\underline{X}) + \mathbb{E} \log |\det J\mathbf{g}(\mathbf{x})|. \quad (\text{A.5})$$

Lineáris, nonszinguláris transzformáció esetén tehát a fenti jelölésekkel a következő összefüggést kapjuk:

$$H(\underline{Y}) = H(\underline{X}) + \log |\det \mathbf{A}|. \quad (\text{A.6})$$

B. Kumulánsok

Jelölje F az Y változó folytonos eloszlásfüggvényét. A valós számokon értelmezett

$$s \mapsto M(s) = \mathbb{E}(e^{sY}) = \int_{\mathbb{R}} e^{sx} F(x) dx \quad (\text{B.1})$$

függvényt az Y változó, illetve az F eloszlás valós momentumgeneráló függvényének nevezzük, ahol az M értelmezési tartománya az olyan s valós számok halmaza, amelyekre az integrál véges (Medvegyev 2002). A momentumgeneráló függvény logaritmusát kumulánsgeneráló függvénynek mondjuk:

$$C = \ln M \quad (\text{B.2})$$

A függvényt Taylor-sorba fejtve a

$$C(s) = \sum_{k=1}^{\infty} c_k \frac{s^k}{k!} \quad (\text{B.3})$$

sorfejtésében szereplő c_k együtthatókat kumulánsoknak nevezzük. Y néhány kumulánása:

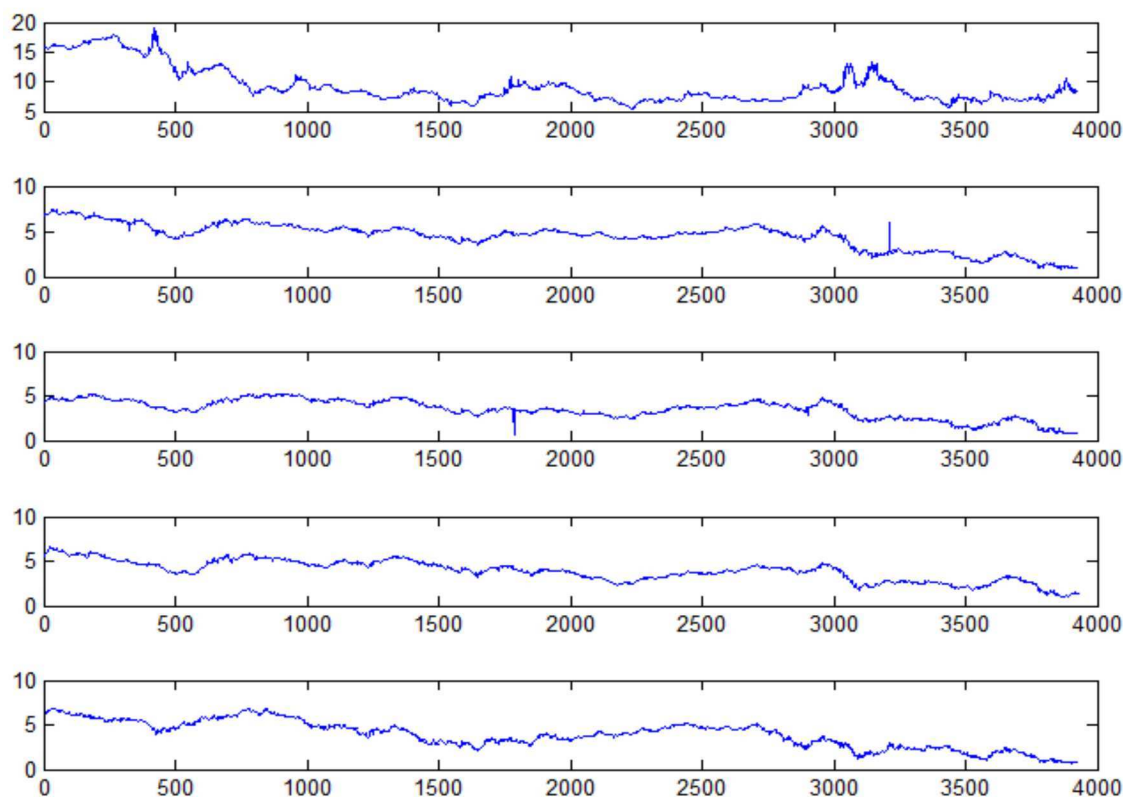
$$\begin{aligned} c_1 &= \mathbb{E}Y \\ c_2 &= \mathbb{E}(Y^2) - (\mathbb{E}Y)^2 \\ c_3 &= \mathbb{E}(Y^3) - 3\mathbb{E}(Y^2)\mathbb{E}Y + 2(\mathbb{E}Y)^3. \end{aligned}$$

Többváltozós esetre a definíció triviálisan kiterjeszthető, illetve a keresztkorreláció analógiájára definiálhatóak az $Y_1, Y_2, \dots, Y_n$ valószínűségi változók keresztkumulánsai is. Jelöljük a keresztkumuláns képzésének operátorát $\text{cum}(\cdot)$ -al, ekkor:

$$\begin{aligned} \text{cum}(Y_i, Y_j) &= \mathbb{E}(Y_i Y_j), \\ \text{cum}(Y_i, Y_j, Y_k) &= \mathbb{E}(Y_i Y_j Y_k), \\ \text{cum}(Y_i, Y_j, Y_k, Y_l) &= \mathbb{E}(Y_i Y_j Y_k Y_l) - \mathbb{E}(Y_i Y_j) \mathbb{E}(Y_k Y_l) - \mathbb{E}(Y_i Y_k) \mathbb{E}(Y_j Y_l) - \mathbb{E}(Y_i Y_l) \mathbb{E}(Y_j Y_k). \end{aligned}$$

C. Ábrák és táblázatok

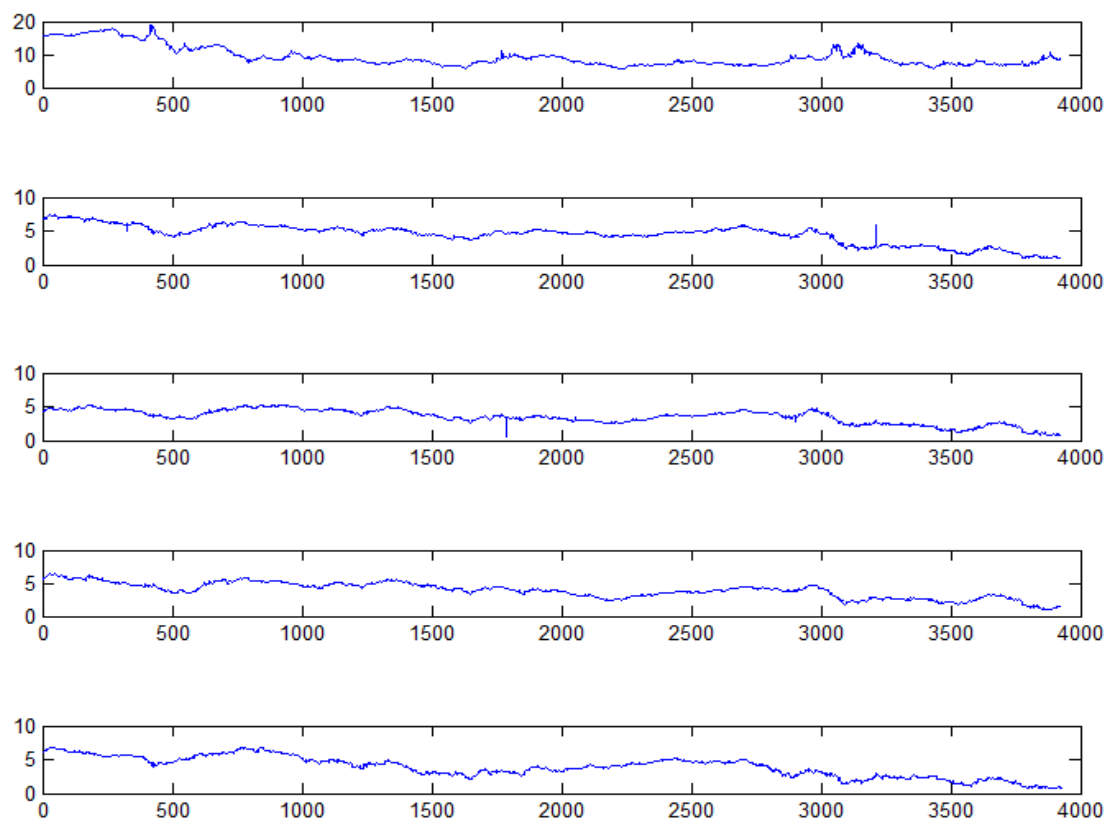
C.1. Adatpótlási módszerek és leíró statisztikáik



C.1. ábra. Konstans módszerrel pótoltt kötvényhozamok idősorai. Az idősorok fentről lefelé: Magyarország, Nagy-Britannia, Németország, Svédország, USA.

C.1. táblázat. Leíró statisztikák a konstans módszerrel pótoltt idősorokra

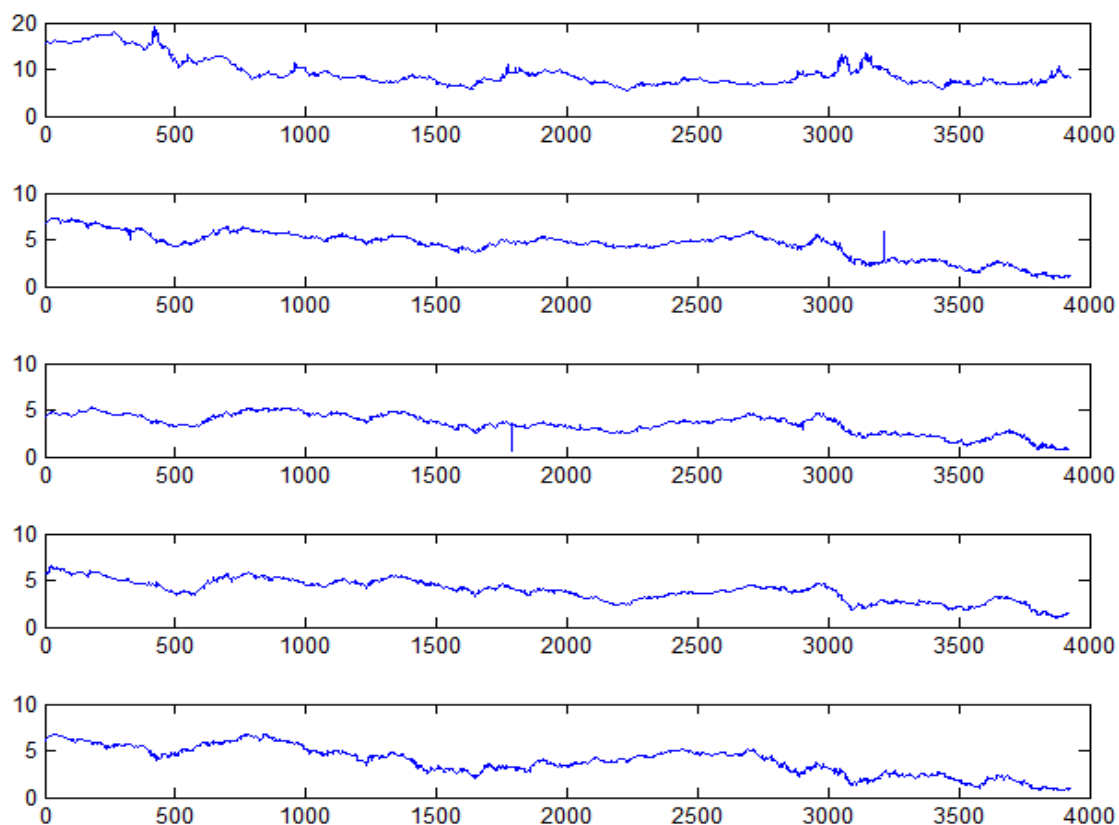
Statisztika	US	GB	D	SE	HU
Átlag	3,909	4,484	3,515	3,906	9,266
Szórás	1,522	1,432	1,059	1,169	2,990
Módusz	4,550	7,040	4,540	3,810	6,850
Medián	4,020	4,757	3,630	3,965	8,280
Ferdeség	-0,113	-0,632	-0,542	-0,234	1,431
Csúcsosság	-0,862	-0,110	-0,368	-0,579	1,046
IQR	2,240	1,301	1,627	1,689	2,680
MAD	1,284	1,096	0,864	0,953	2,257
JB-teszt	130,102	263,425	214,626	90,806	1519,096



C.2. ábra. Lineárisan interpolált kötvényhozamok idősorai. Az idősorok fentről lefelé: Magyarország, Nagy-Britannia, Németország, Svédország, USA.

C.2. táblázat. Leíró statisztikák a lineárisan interpolált idősorokra

Statisztika	US	GB	D	SE	HU
Átlag	3,910	4,484	3,516	3,906	9,266
Szórás	1,522	1,432	1,059	1,169	2,991
Módusz	4,550	7,040	4,730	3,810	6,850
Medián	4,020	4,757	3,630	3,965	8,273
Ferdeség	-0,114	-0,632	-0,544	-0,234	1,430
Csúcsosság	-0,863	-0,111	-0,361	-0,579	1,045
IQR	2,240	1,301	1,616	1,689	2,680
MAD	1,284	1,097	0,863	0,953	2,257
JB-teszt	130,195	263,469	215,089	90,605	1517,767



C.3. ábra. GARCH módszerrel pótoltt kötvényhozamok idősorai. Az idősorok fentről lefelé: Magyarország, Nagy-Britannia, Németország, Svédország, USA.

C.3. táblázat. Leíró statisztikák a GARCH-moddellal pótoltt idősorokra

Statisztika	US	GB	D	SE	HU
Átlag	3,909	4,484	3,515	3,906	9,266
Szórás	1,522	1,432	1,059	1,169	2,991
Módusz	4,550	7,040	4,540	3,810	6,850
Medián	4,020	4,757	3,630	3,965	8,280
Ferdeség	-0,114	-0,632	-0,542	-0,234	1,431
Csúcsosság	-0,862	-0,110	-0,368	-0,580	1,046
IQR	2,240	1,301	1,627	1,689	2,680
MAD	1,284	1,096	0,864	0,953	2,257
JB-teszt	130,132	263,425	214,626	90,951	1518,941

C.2. Adatpótlási módszerek hatása az ICA-ra

C.4. táblázat. A konstans módszerrel pótolts idősorok korrelációja

Komponens	HU	GB	DE	SE	US
HU	1,0000	0,5048	0,3967	0,4899	0,4817
GB	0,5048	1,0000	0,9243	0,9097	0,9157
DE	0,3967	0,9243	1,0000	0,9449	0,8638
SE	0,4899	0,9097	0,9449	1,0000	0,8240
US	0,4817	0,9157	0,8638	0,8240	1,0000

C.5. táblázat. A konstans módszerrel pótolts idősorok komponenseinek korrelációja 5 komponens esetén

Komponens	C_1^{const}	C_2^{const}	C_3^{const}	C_4^{const}	C_5^{const}
C_1^{const}	1	0,0089	0,1350	-0,0649	-0,0243
C_2^{const}	0,0089	1	-0,0045	0,0068	0,0066
C_3^{const}	0,1350	-0,0045	1	-0,0109	0,0233
C_4^{const}	-0,0649	0,0068	-0,0109	1	0,0127
C_5^{const}	-0,0243	0,0066	0,0233	0,0127	1

C.6. táblázat. A lineárisan interpolált idősorok komponenseinek korrelációja 5 komponens esetén

Komponens	C_1^{lin}	C_2^{lin}	C_3^{lin}	C_4^{lin}	C_5^{lin}
C_1^{lin}	1	-0,0060	-0,0228	0,0125	-0,0048
C_2^{lin}	-0,0060	1	-0,0037	-0,0076	0,0076
C_3^{lin}	-0,0228	-0,0037	1	0,0106	0,1181
C_4^{lin}	0,0125	-0,0076	0,0106	1	0,0542
C_5^{lin}	-0,0048	0,0076	0,1181	0,0542	1

C.7. táblázat. A GARCH-modellel pótoltt idősorok komponenseinek korrelációja 5 komponens esetén

Komponens	C_1^{GARCH}	C_2^{GARCH}	C_3^{GARCH}	C_4^{GARCH}	C_5^{GARCH}
C_1^{GARCH}	1	0,0091	-0,0068	-0,0066	-0,0045
C_2^{GARCH}	0,0091	1	0,0645	0,0242	0,1349
C_3^{GARCH}	-0,0068	0,0645	1	0,0127	0,0109
C_4^{GARCH}	-0,0066	0,0242	0,0127	1	-0,0233
C_5^{GARCH}	-0,0045	0,1349	0,0109	-0,0233	1

C.3. Különböző adatpótlási módszerekkel kapott komponensek összehasonlítása

C.8. táblázat. GARCH és Lineáris idősorok komponenseinek korrelációs mátrixa és az abszolút értékben vett maximumok oszloponként

Komponens	C_1^{GARCH}	C_2^{GARCH}	C_3^{GARCH}	C_4^{GARCH}	C_5^{GARCH}
C_1^{lin}	-0,1177	-0,9990	0,0082	0,0090	0,0519
C_2^{lin}	0,9994	0,1265	-0,0309	-0,0039	-0,0047
C_3^{lin}	-0,0100	0,0040	-0,0028	-0,9980	0,0026
C_4^{lin}	-0,0142	-0,0677	-0,0083	-0,0071	0,9966
C_5^{lin}	-0,0163	0,0216	0,9984	0,0061	-0,0126
Abszolút maximum	0,99	0,99	0,99	0,99	0,99

C.9. táblázat. GARCH és Konstans idősorok komponenseinek korrelációs mátrixa és az abszolút értékben vett maximumok oszloponként

Komponens	C_1^{GARCH}	C_2^{GARCH}	C_3^{GARCH}	C_4^{GARCH}	C_5^{GARCH}
C_1^{const}	0,0081	-0,0311	-0,0027	-0,0083	0,9985
C_2^{const}	0,9990	-0,1265	-0,0038	0,0683	-0,0217
C_3^{const}	0,0520	-0,0047	0,0021	0,9966	-0,0126
C_4^{const}	0,0090	-0,0038	-0,9980	-0,0075	0,0062
C_5^{const}	-0,1176	0,9994	-0,0098	-0,0142	-0,0161
Abszolút maximum	0,99	0,99	0,99	0,99	0,99

C.10. táblázat. Lineáris és Konstans idősorok komponenseinek korrelációs mátrixa és az abszolút értékben vett maximumok oszloponként

Komponens	C_1^{const}	C_2^{const}	C_3^{const}	C_4^{const}	C_5^{const}
C_1^{lin}	-0,0233	0,0166	1,0000	0,0061	-0,0125
C_2^{lin}	-0,1254	-1,0000	-0,0166	0,0073	0,0634
C_3^{lin}	-0,0107	-0,0629	-0,0125	-0,0066	1,0000
C_4^{lin}	0,0043	-0,0075	0,0063	1,0000	-0,0074
C_5^{lin}	1,0000	0,1253	-0,0228	0,0040	-0,0107
Abszolút maximum	1	1	1	1	1

C.4. Mesterséges sokkok vizsgálata

C.11. táblázat. Lineáris módszerrel pótoltt idősorok komponenseinek robusztussága konstans sokkolás esetén, véletlen és szisztematikus elemkiválasztással

Véletlenszerű elemkiválasztás				Szisztematikus elemkiválasztás			
Sokk mértéke[%]	1	5	10	Sokk mértéke[%]	1	5	10
1	0,0000	0,0002	0,0009	1	0,0000	0,0000	0,0001
10	0,0001	0,0004	0,0016	10	0,0026	0,0119	0,0211
50	0,0003	0,0018	0,0090	50	0,0000	0,0001	0,0003
100	0,0006	0,0030	0,1990	100	0,0002	0,0009	0,0021
500	0,0020	0,0118	0,2082	500	0,0003	0,0018	0,1883
1000	0,0017	0,2043	0,2070	1000	0,0020	0,0090	0,1920

C.12. táblázat. Lineáris módszerrel pótoltt idősorok komponenseinek robusztussága azonos véletlen sokkolás esetén, véletlen és szisztematikus elemkiválasztással

Véletlenszerű elemkiválasztás				Szisztematikus elemkiválasztás			
Sokk mértéke[%]	1	5	10	Sokk mértéke[%]	1	5	10
1	0,000	0,0001	0,0001	1	0,000	0,000	0,000
10	0,0001	0,0004	0,0009	10	0,0017	0,0018	0,0197
50	0,0002	0,0009	0,0076	50	0,0000	0,0003	0,1979
100	0,0011	0,0023	0,1977	100	0,0000	0,0002	0,0009
500	0,0005	0,1994	0,0242	500	0,0002	0,0008	0,0019
1000	0,0024	0,2180	0,2171	1000	0,0018	0,2016	0,0175

C.13. táblázat. Lineáris módszerrel pótoltt idősorok komponenseinek robusztussága elemenkénti véletlen sokkok esetén, véletlen és szisztematikus elemkiválasztással

Véletlenszerű elemkiválasztás				Szisztematikus elemkiválasztás			
Sokk mértéke[%]	1	5	10	Sokk mértéke[%]	1	5	10
1	0,0000	0,0001	0,0001	1	0,0000	0,0001	0,0001
10	0,0002	0,0003	0,0006	10	0,0017	0,2102	0,2119
50	0,0002	0,0010	0,1979	50	0,0002	0,0010	0,1977
100	0,0002	0,0022	0,2002	100	0,0003	0,0018	0,1979
500	0,0009	0,2000	0,2045	500	0,0005	0,1978	0,0021
1000	0,0018	0,0199	0,2140	1000	0,0022	0,2107	0,0410

C.14. táblázat. Elemenkénti átlagos abszolút eltérések konstans sokkolás esetén, szisztematikus és véletlen elemkiválasztási eljárással

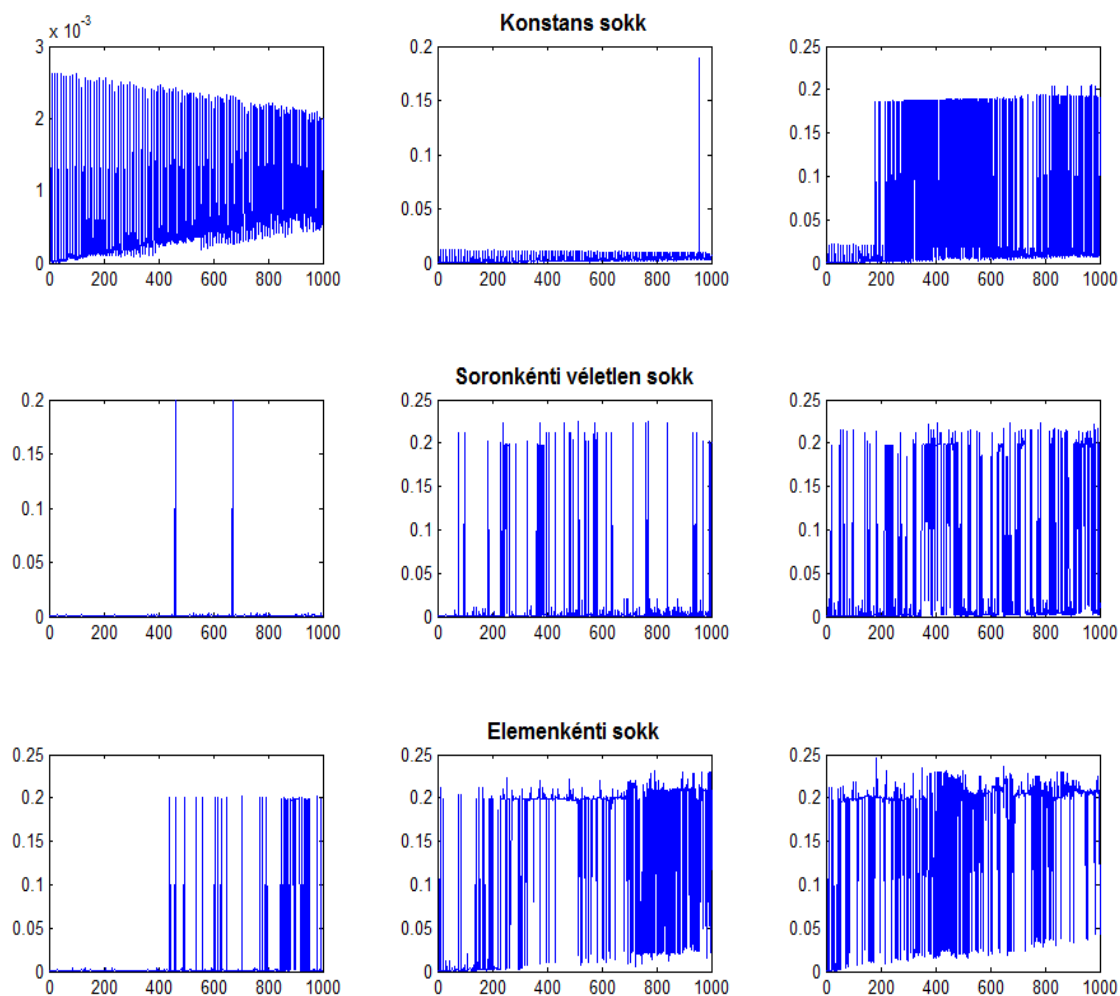
Random elemkiválasztás			Szisztematikus elemkiválasztás		
Sokk mértéke[%]	1	10	Sokk mértéke[%]	1	10
1	0,0003	0,0017	1	0,0001	0,0005
10	0,0014	0,0104	10	0,0113	0,0633
50	0,0052	0,0383	50	0,0007	0,0058
100	0,0086	0,2812	100	0,0017	0,0134

C.15. táblázat. Elemenkénti átlagos abszolút eltérések azonos mértékű véletlen sokkolás esetén, szisztematikus és véletlen elemkiválasztási eljárással

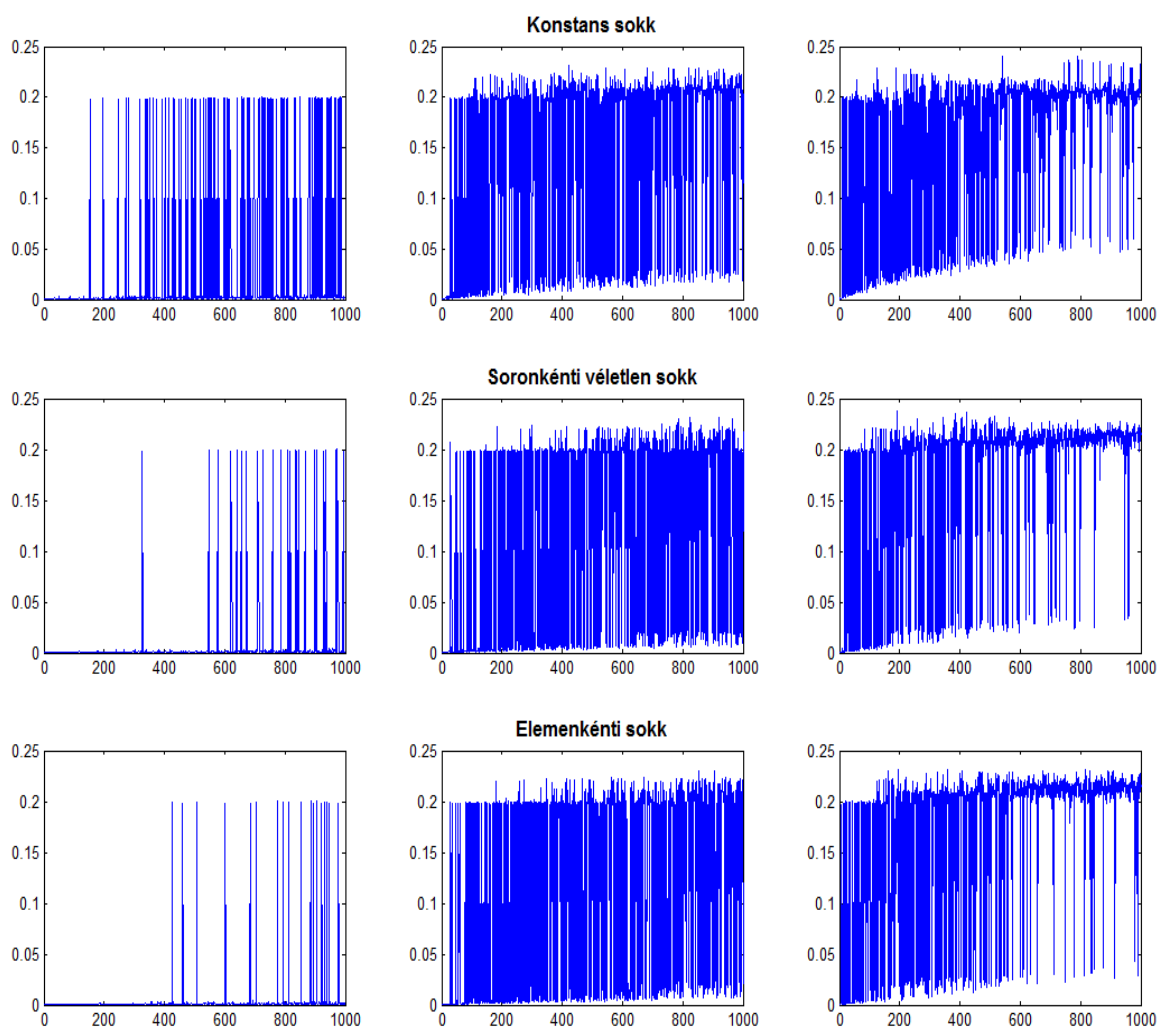
Random elemkiválasztás			Szisztematikus elemkiválasztás		
Sokk mértéke[%]	1	10	Sokk mértéke[%]	1	10
1	0,0002	0,0002	1	0,0000	0,0000
10	0,0011	0,0069	10	0,0113	0,3135
50	0,0030	0,2658	50	0,0005	0,2615
100	0,0050	0,0429	100	0,0005	0,0097

C.16. táblázat. Elemenkénti átlagos abszolút eltérések elemenkénti véletlen sokkolás esetén, szisztematikus és véletlen elemkiválasztási eljárással

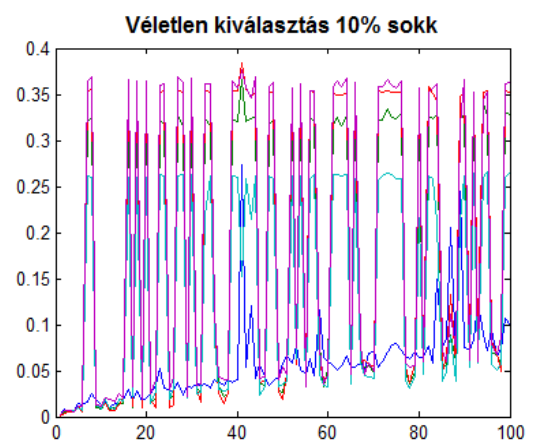
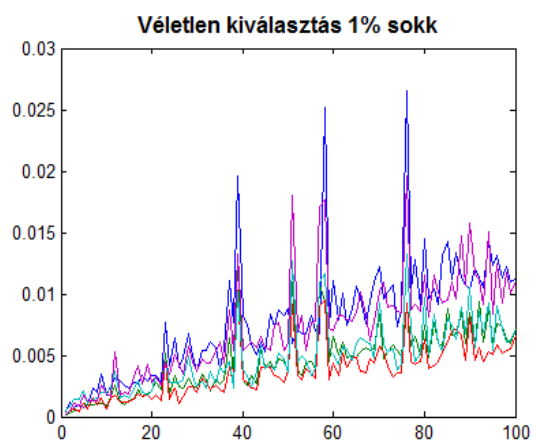
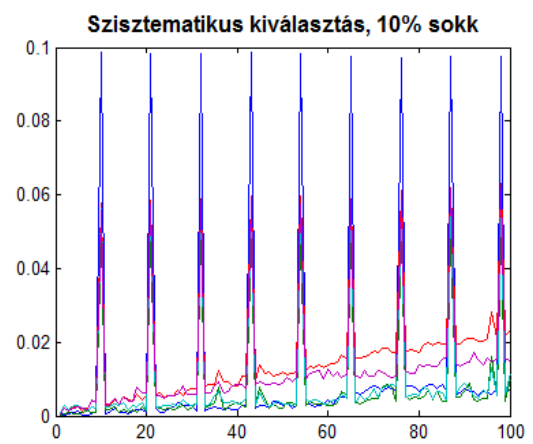
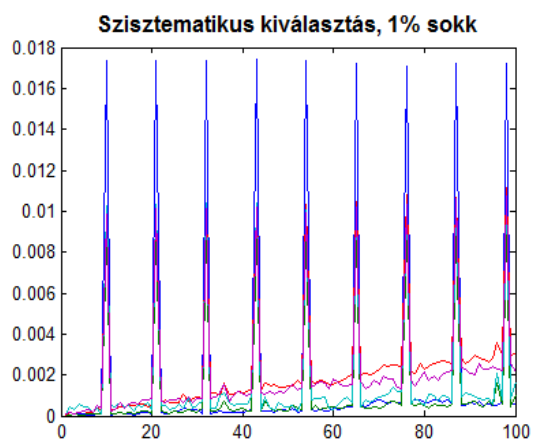
Random elemkiválasztás			Szisztematikus elemkiválasztás		
Sokk mértéke[%]	1	10	Sokk mértéke[%]	1	10
1	0,0001	0,0005	1	0,0001	0,0004
10	0,0010	0,0100	10	0,0101	0,2831
50	0,0027	0,0383	50	0,0020	0,2637
100	0,0053	0,2711	100	0,0037	0,2643



C.4. ábra. Sokkolt és sokkolatlan idősorokra készített komponensek átlagos keresztkorrelációs értékeinek alakulása szisztematikus elemkiválasztás esetén. Az egyes oszlopok a sokkok mértékét mutatják, ez az érték balról jobbra 1, 5, és 10%.

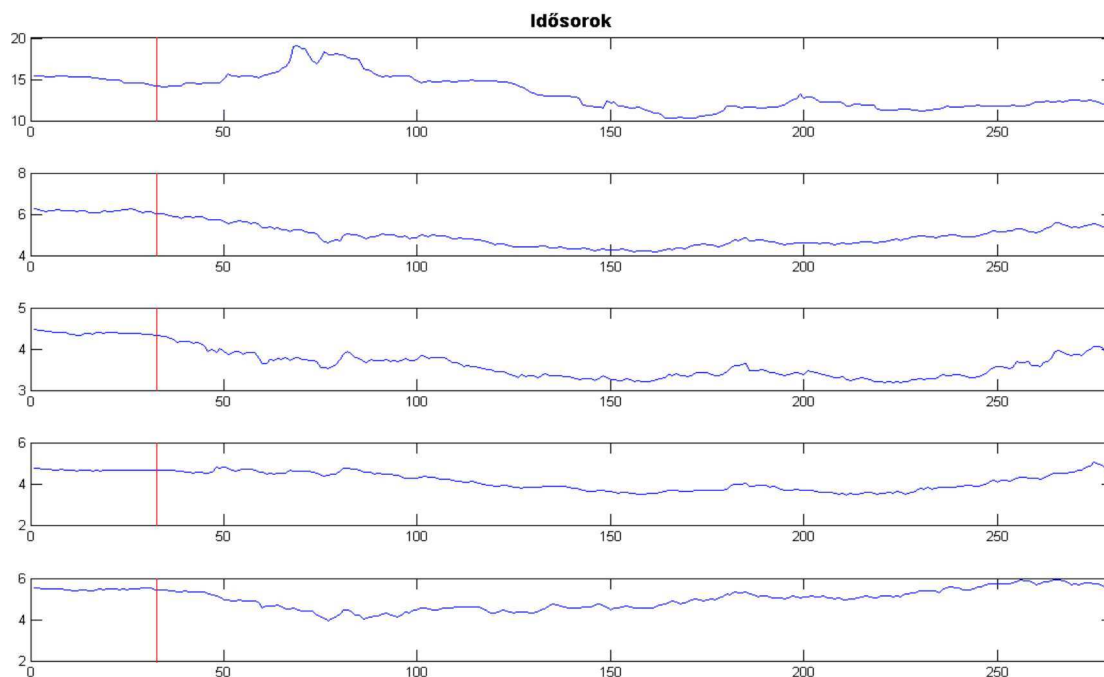


C.5. ábra. Sokkolt és sokkoltatlan idősorokra készített komponensek átlagos keresztkorrelációs értékeinek alakulása véletlen elemkiválasztás esetén. Az egyes oszlopok a sokkok mértékét mutatják, ez az érték balról jobbra 1, 5, és 10%.

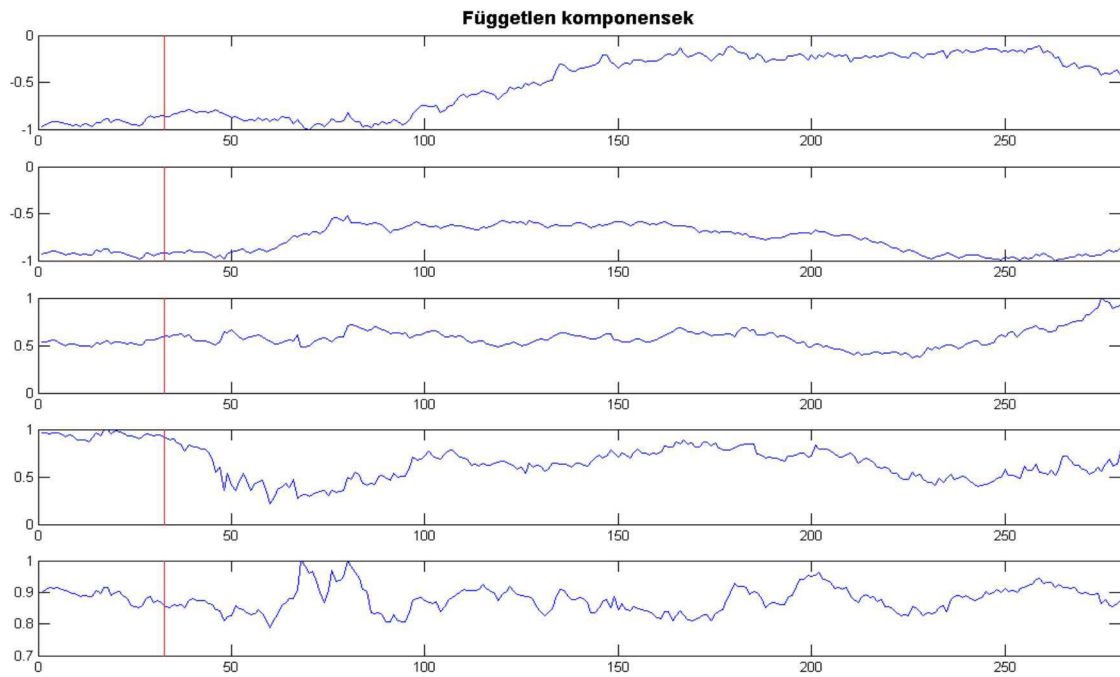


C.6. ábra. Komponensek átlagos abszolút eltérésének alakulása konstans sokkolás esetén.

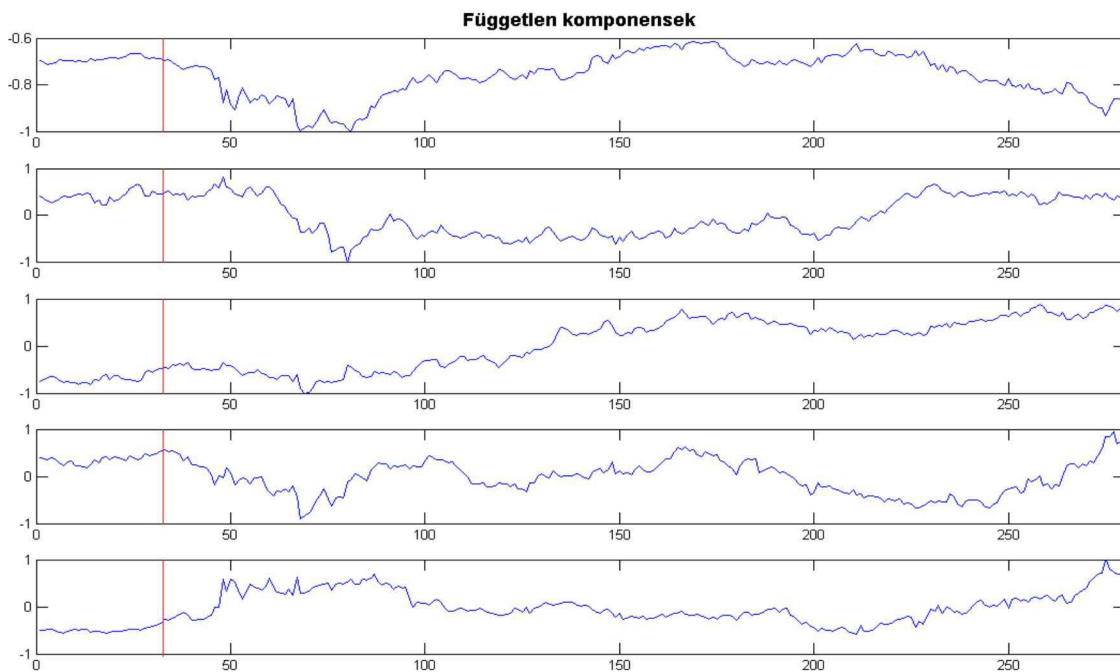
C.5. Pénzügyi sokkok vizsgálata



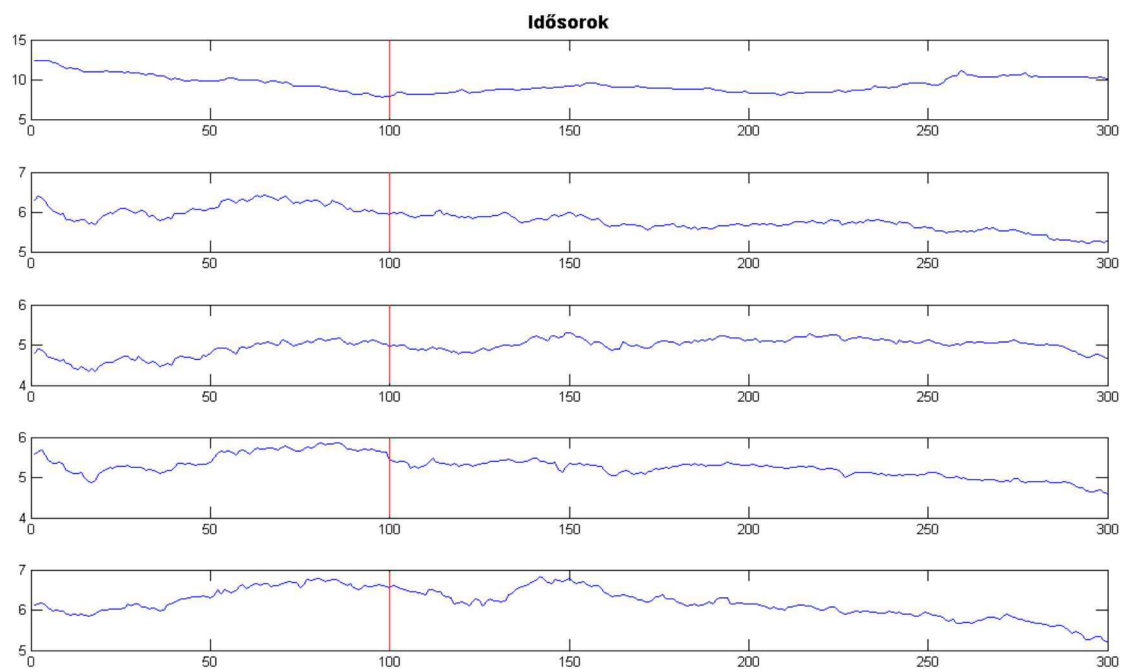
C.7. ábra. Az orosz válság idősorai 1998.06.18-1999.07.15. között, pirossal bejelölve 1998.08.01.-e látható. Az idősorok fentről lefelé: Magyarország, Nagy-Britannia, Németország, Svédország, USA.



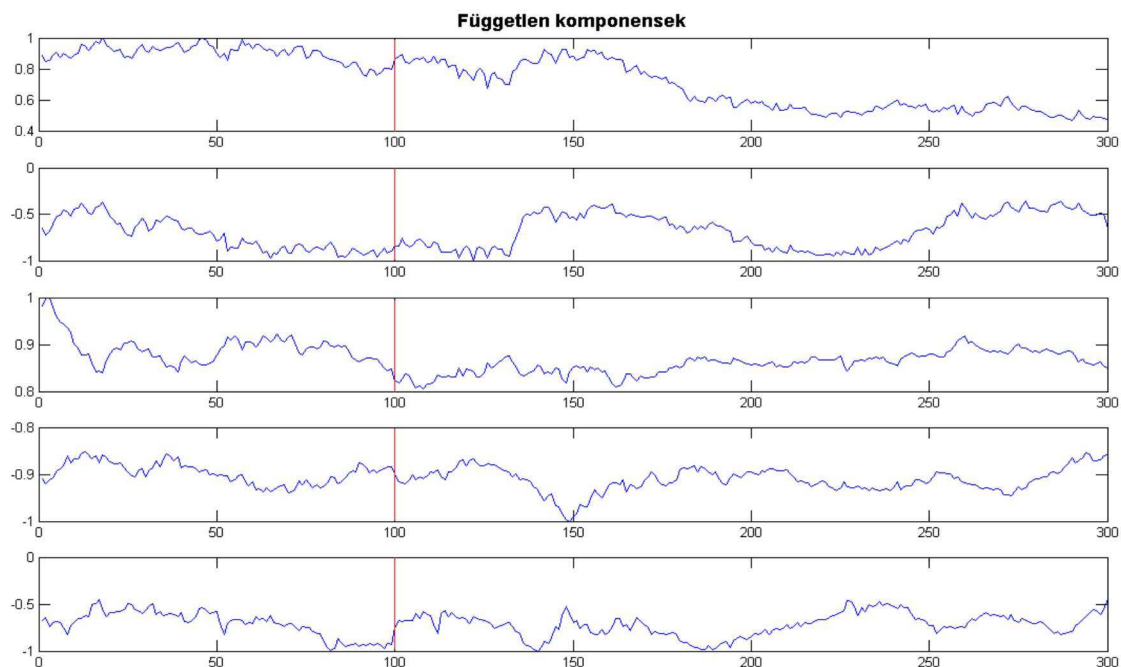
C.8. ábra. Az orosz válság független komponensei FastICA algoritmussal meghatározva, 1998.06.18-1999.07.15. között, pirossal bejelölve 1998.08.01.-e látható.



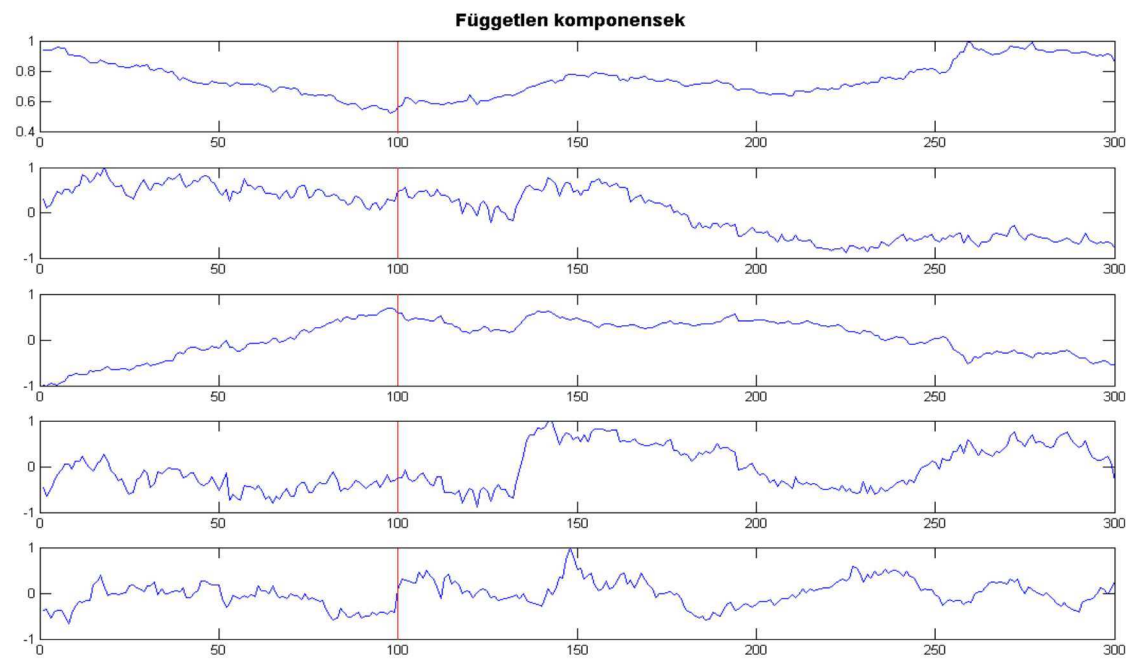
C.9. ábra. Az orosz válság független komponensei neurális ICA algoritmussal meghatározva, 1998.06.18-1999.07.15. között, pirossal bejelölve 1998.08.01.-e látható.



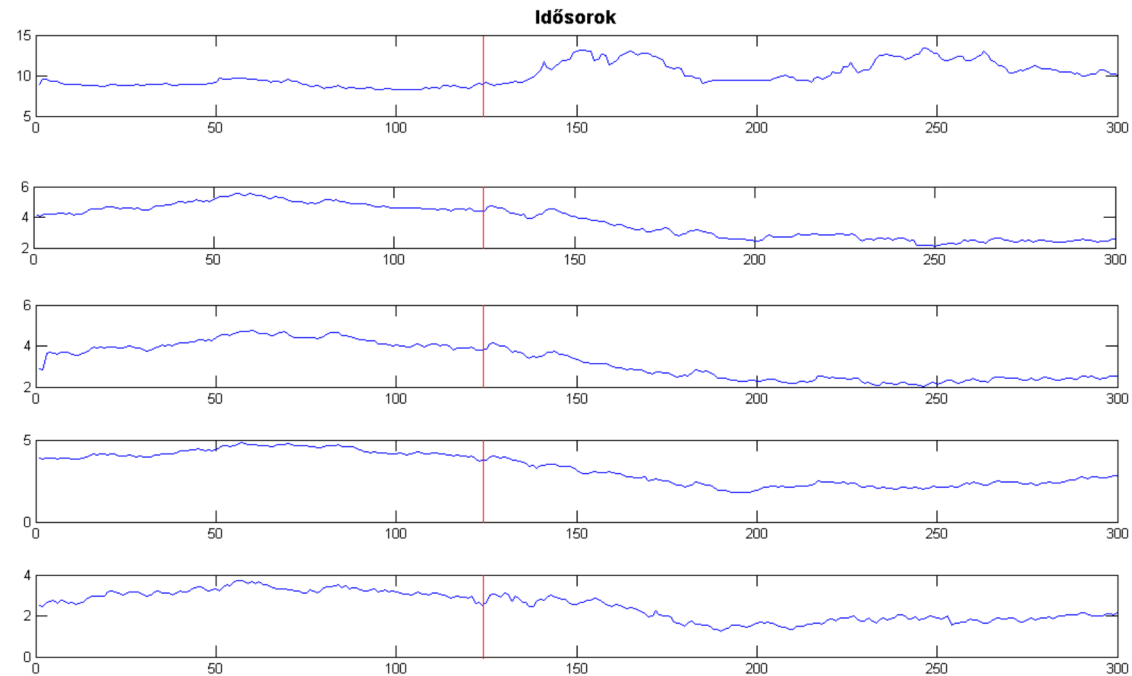
C.10. ábra. Dot-com válság idősorai 1999.10.21-2000.12.14. között, pirossal bejelölve 2000.03.10.-e látható. Az idősorok fentről lefelé: Magyarország, Nagy-Britannia, Németország, Svédország, USA.



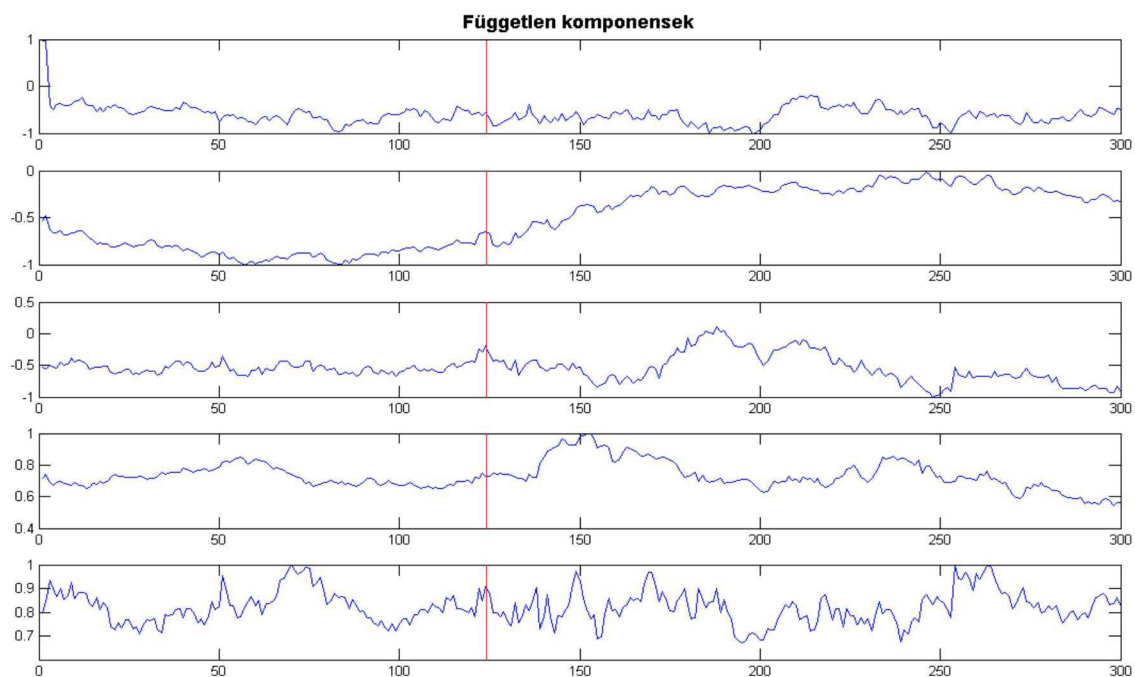
C.11. ábra. Dot-com válság független komponensei FastICA algoritmussal meghatározva, 1999.10.21-2000.12.14. között, pirossal bejelölve 2000.03.10.-e látható.



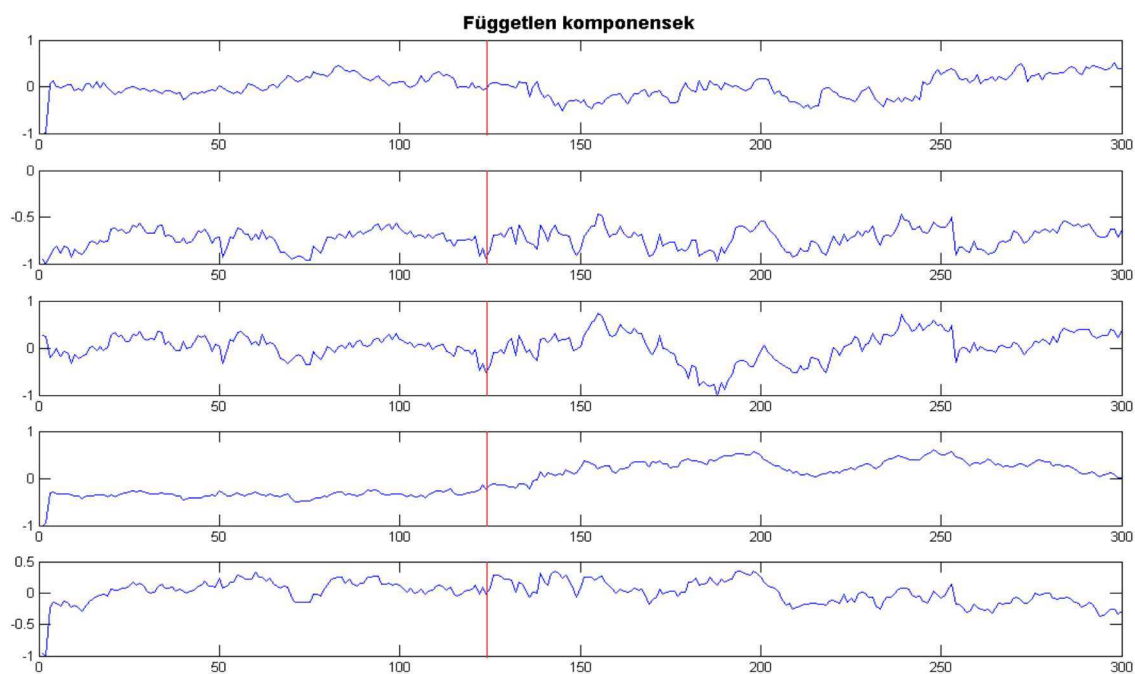
C.12. ábra. Dot-com válság független komponensei neurális ICA algoritmussal meghatározva, 1999.10.21-2000.12.14. között, pirossal bejelölve 2000.03.10.-e látható.



C.13. ábra. Subprime válság idősorai 2008.03.27-2009.05.21. között, pirossal bejelölve 2008.09.15.-e látható. Az idősorok fentről lefelé: Magyarország, Nagy-Britannia, Németország, Svédország, USA.



C.14. ábra. Subprime válság független komponensei FastICA algoritmussal meghatározva, 2008.03.27-2009.05.21. között, pirossal bejelölve 2008.09.15.-e látható.



C.15. ábra. Subprime válság független komponensei neurális ICA algoritmussal meghatározva, 2008.03.27-2009.05.21. között, pirossal bejelölve 2008.09.15.-e látható.