Alkalmazott Matematikai Lapok **37** (2020), 211–224.
DOI: 10.37070/AML.2020.37.2.06

GRÁF ALAPÚ DIMENZIÓREDUKCIÓS HEURISZTIKÁK RÉSZVÉNYPIACI KORRELÁCIÓS MÁTRIXOKRA

GERA IMRE, LONDON ANDRÁS

Az elmúlt években számos tanulmány foglalkozott részvénypiaci hozamok idősoraiból képzett kovariancia-, illetve korrelációs mátrixok vizsgálatával. A korrelációs mátrix becslése során jelentős statisztikai bizonytalanság léphet fel, elsősorban az idősorok véges hossza miatt. Ebben az összefoglaló jellegű cikkben különböző módszereket tárgyalunk a fellépő statisztikai bizonytalanság szűrésére. Bemutatunk egy, a véletlen mátrixok elméletén alapuló, illetve több hierarchikus klaszterezést használó eljárást. A módszerek hatékonyságát a Markowitz-féle portfólió kiválasztási feladaton teszteljük a Budapesti Értéktőzsde historikus részvény idősorain. Az összeállított portfóliókat különböző teljesítménymutatók, valamint a realizált hozam és kockázat segítségével hasonlítjuk össze. Ezen tanulmány elsősorban a szerzők korábban megjelent [10], illetve megjelenés alatt álló [8] munkáit foglalja össze.

1. Bevezetés

A korrelációs mátrixok fontos részét képzik a pénzügyi közgazdaságtannak, elsősorban a portfólió elméletnek és a kockázatmenedzsmentnek. A különböző részvények hozamai közti korrelációt használják például az egyes részvényekbe fektetett tőkearányok meghatározásához úgy, hogy a befektető vállalt kockázata lehetőleg minél kisebb legyen [6]. A korrelációs mátrixokból egyszerű módon képezhetünk gráfokat is. Egy *részvénygráfban* a csúcsok a cégeket (részvényeket), míg a súlyozott élek az árfolyamuk közötti Pearson-korrelációs együtthatót jelentik [5, 12, 17]. Amennyiben gráfként tekintünk ezekre a korrelációs mátrixokra, a gráf alapú adatbányászat, illetve a hálózattudomány széles eszköztára válik elérhetővé [1]. Mindazonáltal a közvetlen gráffá alakítás nem evidens, hiszen a korrelációs mátrixok tényleges információtartalmának meghatározása kulcsszerepet tölt be az alkalmazások területén, különösen a pénzügyi kockázatkezelésben. A korrelációs mátrix múltbéli adatokból való számításához (becsléséhez) jelentős mértékű statisztikai bizonytalanság (zaj) társul a hozamok idősorának véges hossza miatt [25]. Manapság több módszer jelent meg a statisztika, az ökonofizika és a hálózattudomány szakirodalmában a probléma kezelésére, ld. például [4, 9, 22, 24].

Valamennyi módszer azon az elven alapul, hogy meghatározzuk a korrelációs mátrix „információs magját”, ami robusztus a statisztikai bizonytalansággal szemben. Az egyik megközelítés a *véletlen mátrixok elméletén* alapszik. A modellben az empirikus (becsült) korrelációs mátrix és egy *null modell* mátrix sajátértékeit hasonlítjuk össze. A null modell mátrixot általában egy, az empirikussal azonos hosszúságú, véletlen idősorból származtatjuk. Egy hasonló megközelítés, melyet a pénzügyi szakirodalomban alkalmaznak, a főkomponens-analízis [7]. Más szűrési módszerek *hierarchikus klaszterező* eljárásokat alkalmaznak, pl. [12] vagy [22].

Ezen tanulmányban röviden összefoglaljuk az imént említett módszerek alapgondolatait, továbbá megadunk egy új, szintén null modell alapú megközelítést (2. szakasz). Ezután esettanulmányban mutatjuk be a különböző módszerekkel „tisztított” korrelációs mátrixok segítségével meghatározott részvény portfóliók teljesítményét különböző mutatók mentén (3. szakasz). Végül rövid összegzés után megemlítünk néhány potenciális jövőbeni kutatási irányt is (4. szakasz).

2. Korrelációs mátrixok tisztítása

Legyen $X_i \equiv \{x_i(t) : t = 1, 2, \ldots, T\}$ egy idősor, ami egy i elem értékét reprezentálja ($i = 1, 2, \ldots, n$) a $t = 1, 2, \ldots, T$ időpontokban. Speciálisan a részvénypiacot vizsgálva i egy részvény, $x_i(t)$ pedig a $t-1$ és t időpontok közti logaritmikus hozama, azaz

$$x_i(t) = \log \frac{P_i(t)}{P_i(t-1)},$$

ahol $P_i(t)$ az i részvény értéke (ára) a t időpontban. Egy n részvényből álló piacot gyakran vizsgálnak a $\mathbf{C}$ korrelációs mátrixon keresztül, ami statisztikai úton méri a páronkénti függőségeket. A mátrix C_{ij} eleme az i és j részvények közti *Pearson korrelációs együttható*, vagyis

$$C_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i) \cdot \mathrm{Var}(X_j)}},$$

ahol

$$\mathrm{Cov}(X_i, X_j) = \overline{X_i \cdot X_j} - \overline{X_i} \cdot \overline{X_j}$$

az X_i és X_j véletlen változók kovarianciája, $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i) = \sigma_i^2$ az X_i autokovarianciája. Az $\overline{X_i}$ becsült érték az X_i megfigyeléseinek időbeli átlaga, azaz

$$\overline{X_i} = \frac{1}{T} \sum_{t=1}^{T} x_i(t),$$

$$\overline{X_i X_j} = \frac{1}{T} \sum_{t=1}^{T} x_i(t) x_j(t).$$

2.1. Véletlen mátrixok

Egy véletlen mátrix olyan mátrix, melynek elemei véletlenül generált számok valamilyen adott valószínűségi eloszlás szerint [15]. A portfólió elmélet szempontjából a véletlen mátrixok elmélete (Random Matrix Theory, röviden RMT) egy természetes módszertant szolgáltat a korrelációs mátrixok becsléséből adódó statisztikai bizonytalanság kiszűrésére [22]. Legyenek adottak n részvény T hosszú árfolyam idősorai, és tegyük fel, hogy a hozamok független, normális eloszlású véletlen változók 0 várható értékkel és σ^2 varianciával; vagyis adott egy $n \times T$ méretű $\mathbf{W}$ Wishart mátrix. Ekkor határértékben, ha $n \to \infty$, $T \to \infty$, de $Q = T/n$ rögzített, akkor ezen idősorokból képzett $\mathbf{W}\mathbf{W}^\intercal$ korrelációs mátrix sajátértékeinek $\mathcal{P}_{\mathrm{RMT}}(\lambda)$ eloszlása a *Marchenko-Pastur törvény* szerint

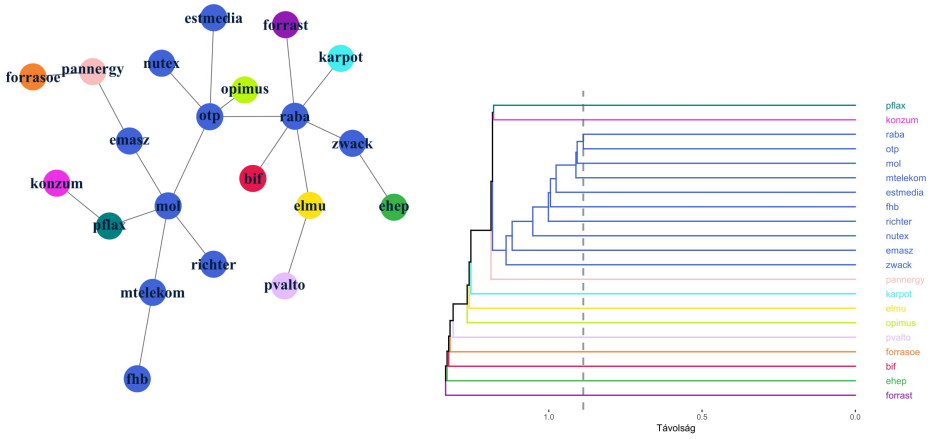
$$\mathcal{P}_{\mathrm{RMT}}(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda - \lambda_{\min})(\lambda_{\max} - \lambda)}}{\lambda},$$

ahol $\lambda_{\min}$ és $\lambda_{\max}$ a mátrix legkisebb, illetve legnagyobb sajátértékei [21], melyek

$$\lambda_{\max,\min} = \sigma^2(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}})$$

alakban adottak, ahol $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, σ^2 pedig $\mathbf{W}$ elemeinek varianciája.

Korábbi tanulmányok rámutattak, hogy részvényárfolyam idősorokból képzett korrelációs mátrixok legnagyobb sajátértéke jelentősen eltér (nagyobb) a véletlen null modellként használatos korrelációs mátrix előbbi $\lambda_{\max}$ sajátértékétől [9, 18]. Elemzők úgy gondolják, hogy a valós adatokból becsült korrelációs mátrix legnagyobb sajátértéke a piac „globális" viselkedését tükrözi [9]. Mivel a Marchenko-Pastur eloszlás csak az $n \to \infty$, $T \to \infty$ esetben teljesül pontosan, ezért az összehasonlításhoz a valós paraméterekkel megegyező n és T értékeket használva szokás véletlen $\mathbf{C}_{\mathrm{RMT}}$ mátrixot generálni és ezt összehasonlítani az eredeti $\mathbf{C}$ korrelációs mátrixszal. Mivel $\mathrm{Trace}(\mathbf{C}) = \lambda_1 + ... + \lambda_n = n$, ezért pl. ha a generált véletlen mátrix elemeinek varianciája $\sigma^2_{\mathrm{RMT}} = 1$, akkor a variancia azon része, amelyet a legnagyobb sajátérték nem magyaráz, a $\sigma_0^2 = 1 - \lambda_{\max}/n$ értékkel becsülhető. Ennek segítségével határozzuk meg a $\mathbf{C}_{\mathrm{RMT}}$ mátrix $\lambda_{\max}$ és $\lambda_{\min}$ értékeit. Az eljárás a $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$ szinguláris érték felbontásával (SVD) folytatódik, ahol $\mathbf{\Lambda}$ a mátrix sajátértékeit csökkenő sorrendben tartalmazó diagonális mátrix, $\mathbf{U}$ pedig az a mátrix, amelynek sorai rendre az ezen sajátértékekhez tartozó sajátvektorok. A zajtisztítás standard módon történik: a $\mathbf{\Lambda}$ mátrixban $\mathbf{C}$ azon sajátértékeit, melyek a kiszámított $\lambda_{\max}$-nál kisebbek, 0-ra állítjuk (legyen a kapott mátrix $\mathbf{\Lambda}'$) és elvégezzük az $\mathbf{U}\mathbf{\Lambda}'\mathbf{U}^\intercal$ szorzást. Végül a kapott mátrix főátlóbeli elemeit visszaállítjuk 1-re.



(a) (b)

1. ábra. Egy részvénygráfon képzett minimális feszítőfa (a) és a hozzá kapcsolódó single-linkage hierarchikus klaszterezés dendrogramja (b).

2.2. Részvény gráfok

Lévén, hogy a $\mathbf{C}$ korrelációs mátrix egy szimmetrikus $n \times n$-es mátrix, tekinthetünk rá egy súlyozott gráf szomszédsági mátrixaként is. Ebben a gráfban a csúcsok a részvényeket jelölik, a súlyozott élek pedig a részvénypárok korrelációs együtthatóit. Az irodalomban $\mathbf{C}$-t gyakran transzformálják egy $\mathbf{D}$ távolságmátrixszá, ahol $D_{ij} = \sqrt{2(1 - C_{ij})}$ [22, 23]. Az így kapott D_{ij} egy ún. ultrametrikus távolság. Az ultrametrikus távolságok ultrametrikus tereket határoznak meg, és a következő axiómákat teljesítik: (i) $D_{ij} = 0 \Leftrightarrow i = j$, (ii) $D_{ij} = D_{ji}$ és (iii) $D_{ij} \leq \max\{D_{ik}, D_{kj}\}, \forall(i, j, k)$; erre egy rövid bizonyítás megtalálható pl. [12]-ben. A módszert korábban többször használták már, mivel a kapott távolságmérték lehetővé teszi gráfalgoritmusok (pl. minimális feszítőfa keresés), illetve hierarchikus klaszterező eljárások alkalmazását [13]. Az ultrametrikus terek alkalmazásaira itt nem térnénk ki ennél részletesebben, az érdeklődő olvasónak a [20] összefoglaló tanulmányt ajánljuk.

Egy egyszerű tisztítási technika a $\mathbf{C}$ (vagy $\mathbf{D}$) értékeinek küszöbölése, ezáltal csak azon élek meghagyása, melyek nagyobbak (kisebbek) egy tetszőlegesen választott küszöbértéknél. Habár a módszer hatékonyan kiküszöböli a leggyengébb korrelációkat, amelyeket vélhetően az idősorok véletlen fluktuációi okoztak, egy nem megfelelően választott küszöbértékkel fontos strukturális jellemzőket dobhatunk el a részvénygráfból.

Egy másik technika, amely nem igényel globális küszöbértéket, az ún. minimális feszítőfa alapú megközelítés. Ez csökkenti a gráfban az élek számát $n \cdot (n-1)/2$-ről $n-1$-re, megtartva a legfontosabb korrelációkat és a gráf összefüggőségét (1a. ábra). Az eljárás szorosan köthető az egyszeres kötésű („single-linkage") agglomeratív hierarchikus klaszterezéshez [12] (1b. ábra). Analóg módon használható az átlagos kötést („average-linkage") használó módszer is. A megközelítés feltételezi, hogy az eredeti korrelációkat jól közelítik a szűrt értékek. Ahhoz, hogy kevesebb információt veszítsünk, használható az ún. maximálisan szűrt síkgráf módszer is [24]. Ez a módszer megtartja a minimális feszítőfa építéséhez használt korrelációkat, illetve néhány további információt is, garantálva, hogy az eredmény egy síkgráf, legfeljebb $3n-6$ éllel.

2.3. Konfigurációs modell és közösségkeresés részvénygráfokban

Részvénygráfok esetén is természetes módon merül fel olyan gráfalapú adatbányászati módszerek használata, mint a közösségkeresés, a közvetlen alkalmazás azonban problematikus lehet. A [11] cikkben a szerzők megmutatták, hogy a korrelációs mátrixot közvetlenül súlyozott gráfként tekintve a modularitás maximalizálás, mint standard közösségkereső eljárás, torzított eredményekhez vezethet (és ugyanez igaz más közösségkereső eljárásokra is). Ennek fő oka, hogy a modularitás függvényben az erősebben korreláló csúcspárok nem feltétlenül kapnak kellően nagy súlyt, ez azonban egy klaszterező eljárásnál kívánatos lenne. A szerzők több, speciálisan korrelációs mátrixokra definiált változatát adták meg a modularitásfüggvénynek. Itt mi egy sokkal egyszerűbb utat választunk. Az eredeti korrelációs mátrixot egy null modell mátrix segítségével tisztítjuk és az így kapott mátrixot megfelelő módon egy távolságmátrixszá alakítjuk. Ezt követően hierarchikus klaszterezést alkalmazunk a távolságmátrixon, mint egyfajta heurisztikát egy modularitás-szerű függvény maximalizálására. Az agglomeratív hierarchikus klaszterezés egy bináris összeolvasztási fát (más néven dendrogramot) épít, amelynek kezdeti elemei (levelei) esetünkben az egyes részvények. A folyamat alulról felfelé haladva minden lépésben valamely távolságfogalom szerint a két legközelebbi adatpontot egy közös csúcsban vonja össze és ezt az összevonást addig ismétli, amíg egyetlen (gyökér) csúcsban egyesül az összes adatpont (részvény). A leggyakrabban használt távolságfogalmak között szerepel két klaszter minimális távolsága (single-linkage), átlagos távolsága (average-linkage) és a maximális távolsága (complete-linkage) [16].

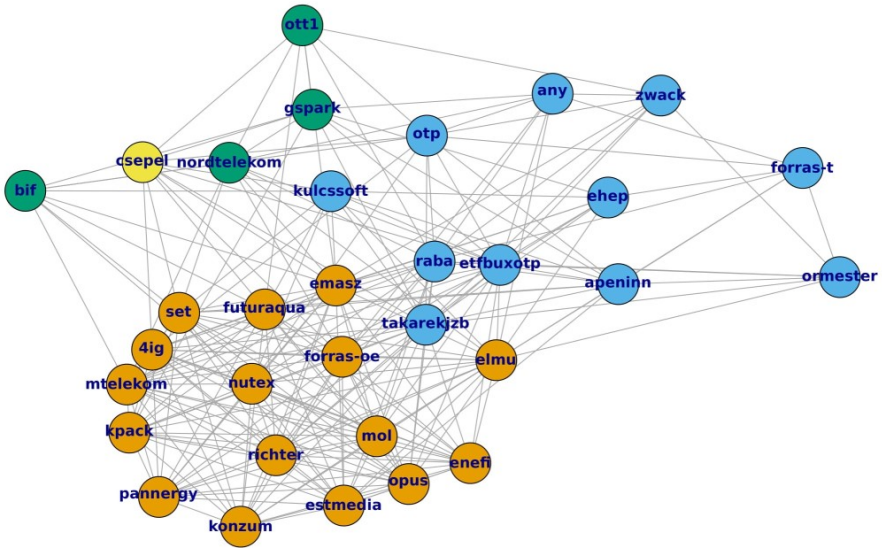
Legyen $\mathbf{C}^0$ egy $n \times n$-es null modell korrelációs mátrix, melynek C^0_{ij} eleme az átlagos korreláció az i és j között valamilyen null modell alatt. Például azt feltételezve, hogy minden részvény korrelálatlan, $\mathbf{C}^0$ az $n \times n$-es egységmátrix lenne. Mi egy konfigurációs modellt használunk null modellként, hogy C^0_{ij}-t generáljuk, a (korrelációs gráfbeli) élek „átdrótozásával", véletlenszerűen és egymástól függetlenül. A feltételezés az, hogy a generált $\mathbf{C}^0$ korrelációs mátrix megtartja minden i részvény súlyozott fokszámát (vagyis *erősségét*), tehát $C^0_i = \sum_j C_{ij}$ amennyi-

re csak lehet állandó, miközben a korrelációs szerkezet véletlenítve van. További részletekért ld. pl. [14].

Ezután egyszerűen a $\mathbf{C}' = |\mathbf{C} - \mathbf{C}^0|$ mátrixot tekintjük a tisztított korrelációs mátrixként. Ez alapján definiáljuk az újraskálázott $\mathbf{D}_c = -\mathbf{C}' + |\min \mathbf{C}'| + |\max \mathbf{C}'|$ távolságmátrixot, ami egy, a korrelációs mátrixhoz kötődő súlyozott gráfként is értelmezhető. Itt a csúcsok között a kisebb távolságok a köztük lévő nagyobb korrelációra utalnak. Ezt követően hierarchikus klaszterezést végzünk a $\mathbf{D}_c$ mátrixon. Ez a módszer tulajdonképpen nem más, mint a modularitás függvény maximalizálására megadott „gyors mohó" („fast-greedy", vagy Leuven [3]) algoritmus. A maximalizálandó függvény megadható

$$M = \sum_{i,j} |C_{ij} - C^0_{ij}| \, \delta_{ij}$$

alakban, ahol $\delta_{ij} = 1$, ha i és j csúcsok azonos klaszterbe kerülnek, különben $\delta_{ij} = 0$. A cél a csúcsok klaszterekbe sorolása úgy, hogy az M érték a lehető legnagyobb legyen. A maximalizáláshoz alkalmazott hierarchikus klaszterezés heurisztika eredményeképpen egy *dendrogramot* kapunk, amelyet egy tetszőleges, a gyökértől számított k-adik szinten elvágva k darab részvényklasztert kapunk (2. ábra).



2. ábra. A $\mathbf{D}_c$ részvénygráf az 1,37-nél nagyobb súlyú éleivel, $k = 4$ klaszterrel.

3. Kísérleti eredmények

Korrelációs (vagy kovariancia) mátrixokat gyakran alkalmaznak a portfólió kiválasztás probléma megoldására. A különböző tisztítási módszerek teljesítményét a tisztított mátrixok segítségével összeállított portfóliók különböző teljesítménymutatóin keresztül mérhetjük. Jelen tanulmányban a Budapesti Értéktőzsdén (BUX) jegyzett részvények záró ár idősorait használva mutatunk be esettanulmányt. További kísérleti eredmények megtalálhatók a szerzők [8, 10] cikkeiben, illetve a bevezetésben hivatkozott publikációkban.

3.1. Adatok

A kísérleteinkhez a Budapesti Értéktőzsde adatai alapján egy részvényhalmaz napi záró árait használtuk fel. Itt a leghosszabban aktív 33 részvényt választottuk ki ($n = 33$, $T = 1962$ rekord, 2011-11-29 és 2019-10-18 között).

3.2. Markowitz-modell

A *Markowitz-féle portfólió kiválasztási probléma* egy olyan optimalizálási feladat, ahol a befektető egy olyan portfóliót szeretne összeállítani a tőzsdei részvényekből (illetve egyéb pénzpiaci termékekből), amely minimális kockázattal és legalább egy adott mértékű várható hozammal bír. A portfóliót egy $\mathbf{p}$ vektorral adjuk meg, melynek az elemei az egyes részvényekbe fektetendő tőkearányokat jelölik. Feltesszük, hogy $\sum_i p_i = 1$. Például a $\mathbf{p} = (0{,}2; 0{,}8)$ azt jelenti, hogy az első részvénybe fektetjük a pénzünk 20%-át, a másodikba pedig a maradék 80%-ot. Az optimális portfóliónak két feltételt kell kielégítenie. Először is a $\sum_i p_i \overline{X}_i$ becsült hozam legyen legalább egy előre adott érték. Másodszor pedig minimális kockázattal kell bírnia, ahol a kockázatot a $\mathbf{p}\boldsymbol{\Sigma}\mathbf{p}^\intercal$ módon számítjuk. Itt a $\boldsymbol{\Sigma}$ a kovarianciamátrixa a figyelembe vett részvényeknek. A negatív p_i súlyok, azaz az ún. *rövidre eladás* („short-selling") is megengedett.

3.3. Egy lehetséges kísérleti módszertan és kiértékelések

Az árfolyam idősorokon a következő mozgóablak módszert alkalmazhatjuk a korreláció (és kovariancia) mérésére és az optimalizálási feladat megoldására. Először meghatározzuk a korrelációs mátrixot a $[t_0, t_0 + \Delta T]$ időablakban, majd végrehajtjuk a különböző tisztítási eljárásokat (ezáltal, visszatranszformálva a korrelációs mátrixokat, újabb kovarianciamátrixokat kapunk), ld. még [22]. Megoldjuk az optimalizálási feladatot mindegyik mátrix esetén (lecserélve az eredeti $\boldsymbol{\Sigma}$ mátrixot), különböző portfólió vektorokat kapva eredményül. Itt ez most hat feladat megoldását jelenti minden t_0 kezdőidőpontra: (1) eredeti Markowitz-modell megoldása, (2) RMT tisztított kovarianciamátrix használata („RMT"), (3-6) hierarchikus klaszterezés alapú tisztítás (i) a $\mathbf{D}$ részvénygráfon („C_Single" és „C_Average"), valamint (ii) a $\mathbf{D}_c$ konfigurációs modell alapú részvénygráfon („Conf_Single" és

„Conf_Average"). A klaszterező eljárások esetében a portfólió kiválasztási stratégiát azzal bővítettük, hogy az optimalizáláshoz klaszterenként egyetlen véletlenszerűen kiválasztott részvényt használhattunk fel. A portfóliók teljesítményét végül a $[t_0+\Delta T, t_0+2\Delta T]$ időintervallum végén értékeltük ki, $t_0 \in \{0, 10, 20, \ldots, T-2\Delta T\}$ és $\Delta T = 100$ paraméterek mellett. Minden $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ portfólióra kiszámoltuk a realizált hozamot a

$$\sum_{i=1}^{n} p_i \frac{P_i(t_0 + 2\Delta T) - P_i(t_0 + \Delta T)}{P_i(t_0 + \Delta T)}$$

képlettel, az előzetes Sharpe hányadost (a becsült hozam és becsült kockázat hányadosa, a $[t_0, t_0 + \Delta T]$ intervallumon számolva) és a kockázati hányadost, ami a 'realizált' ($[t_0 + \Delta T, t_0 + 2\Delta T]$ intervallumon számolt) és becsült kockázat hányadosa. Ezen felül meghatároztuk minden portfólióban az aktív részvények számát is: egy részvényt akkor tekintünk aktívnak, ha az nem szerepel az összeállított portfólióban.

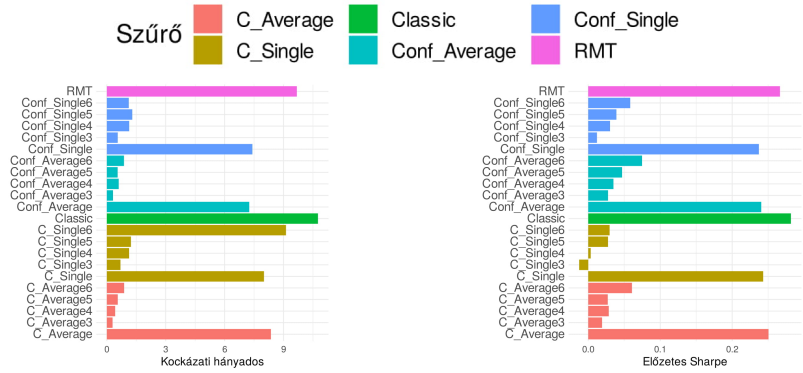
A kísérletek számítógépes megvalósítása R [19] nyelven készült, a 3.2. szekcióban ismertetett kvadratikus programozási feladat megoldását a `quadprog` [2] csomag megoldójával számítottuk ki. Az elvárt minimális hozam értékét dinamikusan, a részvények várható hozamának átlaga és maximuma között 80-20% arányban állítottuk be minden t_0 kezdőpillanatra.

3.4. Eredmények összefoglalása

Az alábbiakban bemutatott eredményeknél és az ábrákon a következő rövidítéseket használjuk: a „Classic" az eredeti Markowitz-modell, az „RMT" a Véletlen Mátrix Elmélet, „C" a hierarchikus klaszterezés, „Conf" a pedig a konfigurációs modell részvénygráfján végrehajtott klaszterezésre utal. Utóbbi kettő esetében a „Single" és „Average" a klaszterezésnél használt távolságfogalmat definiálják. A $k \in \{3, \ldots, 6\}$ szám a jelölés végén arra utal, hogy k klasztert vettünk és mindegyikből egy részvényt használtunk a kovarianciamátrix elkészítéséhez.

A végrehajtott kísérleteink azt mutatják, hogy a bevezetett módszerekkel szűrt kovarianciamátrixok segítségével előállított portfóliók általánosságban javulást mutatnak, főként a becsült és realizált kockázat hányadosában. Ahogyan a 3a. ábra is mutatja, a klaszterezésen alapuló módszerek lényegesen jobb becslést adtak a realizált kockázatra, mint a szűrés nélküli Markowitz modell, különösen akkor, amikor klaszterenként 1-1 részvényt engedtünk csak választani. Ezek közül a konfigurációs modell segítségével végzett szűrések még tovább csökkentették a realizált és becsült kockázat hányadosát. Az előzetes Sharpe hányados (3b. ábra) az eredeti Markowitz-modell esetén adta a legjobb értéket, ehhez az RMT értéke állt legközelebb. Ezt annak tulajdonítjuk, hogy az eredeti modell jelentősen alulbecsülte a kockázatot, ezzel csökkentve a mérőszám nevezőjében szereplő értéket és megnövelve a hányadost. Megfigyelhető, hogy ez a hányados jelentősen csökkent

a klaszterenkénti egy részvény használatának bevezetésével. A realizált hozamok esetében (4a. és 4b. ábra) a klaszterezési eljárások klaszterenként 1-1 részvény felhasználásával magasabb hozamot tudtak produkálni, mint az eredeti modell, ám az alap módszerekkel ezt nem sikerült elérni ezen az adatsoron. Ezek közül az RMT teljesített a legjobban.

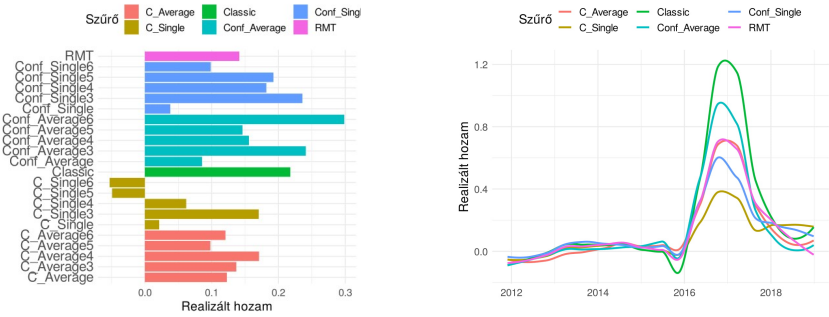


(a) (b)

3. ábra. Átlagos kockázati hányadosok és Sharpe hányadosok a Budapesti Értéktőzsde adatain.

4. Összefoglalás

Ezen tanulmányban korrelációs mátrixok vizsgálatára és gráfalapú adatbányászatra használt módszerek segítségével klaszterezési eljárásokat mutattunk be és hajtottunk végre részvénygráfokon, amelyeket pénzügyi idősorok szűrt korrelációs mátrixaiból készítettünk. Megadtunk egy részvényallokációs stratégiát, amely a kigyűjtött klaszterstruktúrán és a Markowitz portfólió modellen alapszik. Az eredményeink fenti tárgyalása azt mutatja, hogy a korrelációs mátrixok tisztítására használt módszerek képesek a kockázatbecslés tekintetében megbízható portfóliókat összeállítani és az eredeti Markowitz-modellel összevetve kompetitívnek mondhatók a realizált hozamok tekintetében is. A részvénygráfok különböző szűrési procedúrák alapján való definiálása és a klaszter alapú részvénykiválasztási stratégiák számos további kérdést hagynak nyitva a jövőbeli vizsgálatok számára. További kiterjesztési lehetőség különböző hozambecslések alkalmazása (mint pl. a *James-Stein* becslés), illetve az RMT más eredményeinek felhasználása. További vizsgálatok tárgyát képezheti az optimális portfólió méret meghatározása (ugyanis a valóságban általában jelentős tranzakciós költségekkel kell számolni), a befektetési időszak hosszának optimalizálása, valamint a realizált hozamok időbeli alakulása is (ld. pl. 4b. ábra).



(a) (b)

4. ábra. Átlagos és időbeli realizált hozamok a BUX adathalmazon.

1. táblázat. Összefoglaló táblázat a szűrési módszerek eredményeiről. Az oszlopokban az átlagok, mögöttük zárójelben a szórás látható.

Szűrés	k	**Hozam**	**Kock. hányados**	**Pre-Sharpe**	**Aktív részv.**
Classic	–	0.2177 (0.779)	10.7536 (57.305)	0.2816 (0.348)	31.2712 (0.493)
C_Average	3	0.1369 (0.686)	0.2907 (1.158)	0.0192 (0.153)	3.0000 (0.000)
C_Average	4	0.1709 (0.753)	0.4242 (0.948)	0.0285 (0.167)	4.0000 (0.000)
C_Average	5	0.0981 (0.604)	0.5553 (1.410)	0.0271 (0.194)	5.0000 (0.000)
C_Average	6	0.1207 (0.780)	0.8801 (2.869)	0.0607 (0.222)	6.0000 (0.000)
C_Average	–	0.1227 (0.645)	8.3576 (48.182)	0.2505 (0.298)	31.2712 (0.494)
C_Single	3	0.1704 (0.673)	0.6936 (1.361)	-0.0126 (0.164)	3.0000 (0.000)
C_Single	4	0.0619 (0.566)	1.1338 (2.558)	0.0034 (0.171)	4.0000 (0.000)
C_Single	5	-0.0490 (0.656)	1.2277 (2.322)	0.0275 (0.163)	5.0000 (0.000)
C_Single	6	-0.0527 (0.752)	9.1248 (100.466)	0.0297 (0.226)	6.0000 (0.000)
C_Single	–	0.0214 (0.753)	8.0039 (49.410)	0.2430 (0.270)	31.2712 (0.494)
Conf_Average	3	0.2409 (1.195)	0.3130 (0.934)	0.0275 (0.124)	3.0000 (0.000)
Conf_Average	4	0.1558 (0.653)	0.6039 (1.735)	0.0350 (0.131)	4.0000 (0.000)
Conf_Average	5	0.1461 (0.822)	0.5494 (1.212)	0.0470 (0.153)	5.0000 (0.000)
Conf_Average	6	0.2985 (1.108)	0.8718 (2.764)	0.0747 (0.182)	6.0000 (0.000)
Conf_Average	–	0.0856 (0.678)	7.2534 (45.687)	0.2404 (0.276)	31.2712 (0.494)
Conf_Single	3	0.2358 (0.934)	0.5539 (0.819)	0.0121 (0.128)	3.0000 (0.000)
Conf_Single	4	0.1819 (0.827)	1.1377 (3.065)	0.0303 (0.126)	4.0000 (0.000)
Conf_Single	5	0.1925 (1.092)	1.2952 (3.170)	0.0391 (0.127)	5.0000 (0.000)
Conf_Single	6	0.0986 (1.413)	1.1168 (1.744)	0.0583 (0.131)	6.0000 (0.000)
Conf_Single	–	0.0380 (0.761)	7.4137 (46.319)	0.2371 (0.267)	31.2712 (0.494)
RMT	–	0.1414 (0.570)	9.6779 (52.950)	0.2664 (0.314)	31.2712 (0.493)

Köszönetnyilvánítás

A kutatás az EFOP-3.6.1-16-2016-00008 számú projekt támogatásával készült.

Gera Imrét az Innovációs és Technológiai Minisztérium ÚNKP-19-2 kódszámú Új Nemzeti Kiválóság Programja támogatta.

Köszönjük továbbá az anonim bírálók alapos munkáját, javaslataik jelentősen hozzájárultak a cikk minőségének javításához.

Nemzeti Kutatási, Fejlesztési
és Innovációs Hivatal

Hivatkozások

[1] Albert-László Barabási: *Network science*, Cambridge University Press, (2016). DOI: 10.1098/rsta.2012.0375

[2] A. Berwin and Andreas Weingessel: *quadprog: Functions to solve Quadratic Programming Problems*, R package version 1.5-5. 2013. https://cran.r-project.org/web/packages/quadprog/index.html

[3] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre: *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics, P10008 (2008). DOI: 10.1088/1742-5468/2008/10/P10008

[4] Joël Bun, Jean-Philippe Bouchaud and Marc Potters: *Cleaning large correlation matrices: tools from random matrix theory*, Physics Reports, Vol. **666**, pp. 1-109 (2017). DOI: 10.1016/j.physrep.2016.10.00

[5] K. Tse Chi, Jing Liu and Francis C.M. Lau: *A network perspective of the stock market*, Journal of Empirical Finance, Vol. **17** No. **4**, pp. 659-667 (2010). DOI: 10.1016/j.jempfin.2010.04.008

[6] Edwin J. Elton, Martin J. Gruber, Stephen J. Brown and William N. Goetzmann: *Modern portfolio theory and investment analysis*, John Wiley & Sons, (2009).

[7] Robert F. Engle, Victor K. Ng and Michael Rothschild: *Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills*, Journal of Econometrics, Vol. **45** No. **1-2**, pp. 213-237 (1990). DOI: 10.1016/0304-4076(90)90099-F

[8] Imre Gera and András London: *Portfolio selection based on a configuration model and hierarchical clustering for asset graphs*, Proceedings of the MATCOS'19, (2019). (megjelenés alatt)

[9] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud and Marc Potters: *Noise dressing of financial correlation matrices*, Physical Review Letters, Vol. **83** No. **7**, p. 1467 (1999). DOI: 10.1103/PhysRevLett.83.1467

[10] András London, Imre Gera and Balázs Bánhelyi: *Markowitz Portfolio Selection Using Various Estimators of Expected Returns and Filtering Techniques for Correlation Matrices*, Acta Polytechnica Hungarica, Vol. **15** No. **1**, pp. 217-229 (2018). DOI: 10.12700/APH.15.1.2018.1.13

[11] Mel MacMahon and Diego Garlaschelli: *Community detection for correlation matrices*, Physical Review E, Vol. **5**, p. 21006 (2013). DOI: 10.1103/PhysRevX.5.021006

[12] Rosario N. Mantegna: *Hierarchical structure in financial markets*, The European Physical Journal B, Vol. **11** No. **1**, pp. 193-197 (1999). DOI: 10.1007/s100510050929

[13] K.V. Mardia, J.T. Kent and J.M. Bibby: *Multivariate Analysis*, Academic Press, London-New York-Toronto-Sydney-San Francisco, Vol. **15**, p. 518 (1979).

[14] Naoki Masuda, Sadamori Kojaku and Yukie Sano: *Configuration model for correlation matrices preserving the node strength*, Physical Review E, Vol. **98** No. **1**, p. 12312 (2018). DOI: 10.1103/PhysRevE.98.012312

[15] Madan Lal Mehta: *Random matrices*, Academic Press, Vol. **142** (2004).

[16] Frank Nielsen: *Hierarchical Clustering*, pp. 195-211, Febr. (2016). ISBN: 9783-319-21902-8. DOI: 10.1007/978-3-319-21903-5_8

[17] J-P Onnela, Kimmo Kaski and János Kertész: *Clustering and information in correlation based financial networks*, The European Physical Journal B, Vol. **38** No. **2**, pp. 353-362 (2004). DOI: 10.1140/epjb/e2004-00128-7

[18] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A. Nunes Amaral and H. Eugene Stanley: *Universal and nonuniversal properties of cross correlations in financial time series*, Physical Review Letters, Vol. **83** No. **7**, p. 1471 (1999). DOI: 10.1103/PhysRevLett.83.1471

[19] R. Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, (2019). https://www.R-project.org/

[20] Rammal Rammal, Gérard Toulouse and Miguel Angel Virasoro: Ultrametricity for physicists, Reviews of Modern Physics, Vol. **58** No. **3**, p. 765 (1986). DOI: 10.1103/RevModPhys.58.765

[21] Anirvan M. Sengupta and Partha P. Mitra: *Distributions of singular values for some random matrices*, Physical Review E, Vol. **60** No. **3**, p. 3389 (1999). DOI: 10.1103/PhysRevE.60.3389

[22] Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, Rosario N. Mantegna: *Cluster analysis for portfolio optimization*, Journal of Economic Dynamics and Control, Vol. **32** No. **1**, pp. 235-258 (2008). DOI: 10.1016/j.jedc.2007.01.034

[23] Michele Tumminello, Fabrizio Lillo and Rosario N. Mantegna: *Hierarchically nested factor model from multivariate data*, EPL (Europhysics Letters), Vol. **78** No. **3**, p. 30006 (2007). DOI: 10.1209/0295-5075/78/30006

[24] M. Tumminello, T. Aste, T. Di Matteo and R.N. Mantegna: *A tool for filtering information in complex systems*, Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol. **102** No. **30**, pp. 10421-10426 (2005). DOI: 10.1073/pnas.0500298102

[25] Michel Verleysen and Damien François: *The curse of dimensionality in data mining and time series prediction*, International Work-Conference on Artificial Neural Networks, Springer, pp. 758-770 (2005). DOI: 10.1007/11494669_93

Gera Imre 1997-ben született Orosházán. 2015 óta a Szegedi Tudományegyetem programtervező informatikus hallgatója, ahol 2018-ban alapképzésen szerzett diplomát, azóta mesterképzésen folytatja tanulmányait. Egyetemi évei alatt elsősorban a Számítógépes Optimalizálás Tanszék kutatásaiba kapcsolódott be. Kétszer is indult a Tudományos Diákköri Konferencián, ahol helyi fordulóban egyszer első, egyszer második helyezést ért el, valamint az országos versenyen első, illetve különdíjas lett. 2018 óta az Új Nemzeti Kiválóság Program ösztöndíjasa.

Gera Imre

Szegedi Tudományegyetem,
Informatikai Intézet,
6720 Árpád tér 2.
gerai@inf.u-szeged.hu

London András 1989-ben született Szegeden. Alkalmazott matematikus MSc diplomát 2012-ben, PhD fokozatot (Informatikai tudományok) 2018-ban szerzett. Jelenleg a Szegedi Tudományegyetem Informatikai Intézet Számítógépes Optimalizálás Tanszékének adjunktusa, illetve részállásban a Poznań-i Közgazdasági Egyetem Operációkutatás Tanszékének adjunktusa. Fő érdeklődési területe a matematikai modellezés, gráf alapú adatbányászat és komplex hálózatok vizsgálata. Referált cikkeinek száma 15, hivatkozási száma több, mint 300.

London András

Szegedi Tudományegyetem,
Informatikai Intézet,
6720 Árpád tér 2.
london@inf.u-szeged.hu

GRAPH-BASED DIMENSION REDUCTION HEURISTICS TO STOCK CORRELATION MATRICES

Imre Gera, András London

Many studies have dealt with the investigation of covariance- and correlation matrices defined by stock price time series over the past few years. Most of these studies highlighted that the estimation of the correlation matrix is associated with a significant amount of statistical uncertainty (or sometimes called noise) and proposed several methods to filter it out. In this survey-kind paper we present different methods found in the literature and propose a novel approach too. Namely, we present a method using the results of random matrix theory, and other methods based on hierarchical clustering procedures. To measure and compare the performance of the methods we utilize the Markowitz portfolio selection problem and perform experiments on the historical stock time series data of the Budapest Stock Exchange. The created portfolios are compared based on several performance indices, realized returns and risk measures. This paper is mainly considered as an overview of papers published [10] and under publishing [8].

Keywords: Correlation matrices, asset graphs, portfolio optimization.

Mathematics Subject Classification (2000): I.6 [Simulation and Modeling]: Applications; G.1.6 [Optimization]: Nonlinear programming