A főkomponens analízis bemutatása

Sári Zoltán
Neumann János Informatikai Kar
Óbudai Egyetem
Budapest
Email: sari.zoltan.tamas@gmail.com

2015.11.30.

Kivonat

A statisztikai elemzési és az adatbányászati feladatok egyik legfontosabb feltétele az adatok hatékony előfeldolgozása. Az előfeldolgozás kritikus, mind az alkalmazott elemzési eljárások futási ideje, mind az elemzési cél elérésének hatékonysága szempontjából. A tevékenység széles terület, amely számos különböző stratégiából és ezekhez kapcsolódó olyan módszerekből áll, amelyek szoros kölcsönhatásban állnak egymással. A folyamat az összes olyan tevékenységet lefedi, amely az elemzés bemeneteként szolgáló nyers adatoknak a hatékony modellezésre alkalmassá tételét eredményezi. Kiemeltek ezek közül az adatredukciós lépések.

Az adatbányászati alkalmazásokban a nyers adathalmaz mérete általában nagy, mind a rekordok számosságát, mind az objektumokat leíró jellemzők (attribútumok) dimenzióját tekintve. Az elemzés során az irreleváns és redundáns attribútumokkal járó adattömeg jelentősen lassítja a feldolgozást. Alapvető fontosságú az adathalmaz attribútumai számának (dimenzió) redukálása úgy, hogy a megfigyelésekben rejlő információ ne csökkenjen lényegesen.

Dimenziócsökkentés fontos eszköze a főkomponens analízis, amely egy többváltozós statisztikai eljárás, a faktoranalízis egy részterülete. Segítségével kiküszöbölhetők a lényegtelen jellemzők, mérsékelhető a zaj és a kezelhető a dimenzió probléma (dimenzió átok). Az eljárás a redukció mellett megtartja az adathalmaz karakterisztikáját és lényeges információtartalmát annak érdekében, hogy a redukálást követően az adathalmazban rejlő minták továbbra is észrevehetőek és felismerhetőek legyenek. A főkomponens elemzés emellett megkönnyíti a modell interpretálását és az eredmények ábrázolását. A módszer legfőbb alkalmazási területe az adatbányászat, az eljárást elsősorban a feltáró adatelemzésben és a prediktív modellek előállításában használják.

A dolgozatot három fő fejezet alkotja. Elsőként általánosan kerülnek bemutatásra az adatredukciós technikák. A következő fejezet a dimneziócsökkentés speciális eseteit tárgyalja. Majd a dolgozat fókuszát jelentő főkomponens analízis kerül részletesen ismertetésre. A főkomponens elemzés bemutatását követően a dolgozat kitér a technika értékelésére, előnyeinek és hátrányainak bemutatására, végül a módszer alkalmazását egy példán keresztül mutatja be.

1. Adatredukció

Az adatok redukálásának módszereivel az adathalmaz olyan csökkentett reprezentációja kapható meg, amely méretében sokkal kisebb, mégis megőrzi az eredeti adatok integritását. A redukált adathalmazon végzett adatbányászat sokkal hatékonyabb, gyorsabb, mégis az eredetivel közel azonos eredményt szolgáltat.

Alapvető elvárás, hogy a redukált adatmennyiségből levont statisztikai következtetések érvényesek maradjanak az eredeti adathalmazon is. Ennek megfelelően az adatredukció problémáját egy nagy adathalmazt egy kisebb méretűvel való helyettesítése jelenti úgy, hogy emellett a redukált adatok valamilyen szempont szerint

hűen reprezentálják az eredeti, nagy adathalmazt. Az adathalmaz csökkentésre több módszer is lehetőséget ad, amelyek vonatkozhatnak a rekordszám (vertikális) illetve az attribútumok számának a csökkentésére (horizontális).

	vertikális rekordok számosságát érintő	**horizontális** attribútumok számosságát érintő
részhalmaz kiválasztása	mintavételezés: rekordok egy reprezentatív részhalmazának kiválasztása	eliminálás: egyes attribútumok elhagyása
összevonás	aggregálás: több rekord egyesítése	dimenziócsökentés: egymással összefüggő attribútumok összevonása

1. táblázat. Az adatredukció alapvető módszerei

További redukálási lehetőséget jelentenek az egyes attribútumok értékkészletének szűkítésére irányuló módszerek:

- **kisebb méretű reprezentációk**: az adatok helyettesítése parametrizált modellel (pl. regresszió)
- **fogalmi helyettesítés**: attribútumok értékeinek értékkészletük intervallumaival vagy magasabb szintű fogalmakkal helyettesítése (diszkretizáció, klaszterezés)
- **adattömörítés**: információveszteség nélküli kódolási mechanizmusok

2. Attribútumok redukálása

Az adatok feldolgozását jelentősen lassítják az irreleváns és redundáns attribútumok. A dimenziócsökkentési eljárások az ilyen attribútumok eltávolítására, összevonására irányulnak. A dimenziócsökkentés módszerei lehetővé teszik, hogy az adatmátrix méretét csökkentve kisebb költséggel legyen értékelhető a statisztikai sokaság. A feldolgozási hatékonyság növelése mellett a dimenziószám csökkentésének további előnye, hogy megkönnyíti az adatok megjelenítését, megértését és csökkenti a tárolási költségeket.

A feladat egy olyan minimális attribútumhalmaz keresése, amelyben az adatosztályok valószínűségi eloszlása a lehető legjobban közelíti az összes attribútum alapján kapható eredeti eloszlást. Kisebb adathalmazok elemzésekor, az adatok közötti összefüggések ismeretében célra vezetőek a manuális, felhasználó döntésén alapuló redukálás. Nagy attribútum halmaz esetében azonban a manuális módszer veszélye, hogy kizárásra kerülnek az elemzés szempontjából releváns jellemzők és megmaradnak a lényegtelen attribútumok, amelynek főként a tanuló és neurális algoritmusok látják kárát. A nagy, 10-nél magasabb dimenzió számú adathalmazok redukálására számos algoritmus áll rendelkezésre. Az eljárások a heurisztikus módszerek és statisztikai módszerek (faktoranalízis) köré csoportosíthatók.

2.1. Heurisztikus módszerek

Mivel az optimális részhalmaz kiválasztási probléma NP-teljes (n elemszámú attribútum részhalmazainak száma 2^n), ezért a kimerítő keresés helyett hatékony alkalmazási lehetőséget a csökkentett keresési téren alkalmazott heurisztikus eljárások kínálnak. A heurisztikus algoritmusok mohók, az adott pillanatban legjobbnak tűnő lépést választják, jellemzőjük a lokálisan optimális választás előnyben részesítése.

A heurisztikus módszerek esetében a legjobb ill. legrosszabb attribútumok kiválasztása az attribútumok relevanciáját mérő metrikák és az azokra vonatkozó küszöbértékek alkalmazását igénylik. A metrikák jellemzően a statisztikai szignifikanciát mérő tesztekre épülnek, amelyek az attribútumok függetlenségét feltételezik. Használatos továbbá a döntési fáknál is alkalmazott, az entrópiára épülő információnyereség mérése.

Az attribútumok részhalmazának kiválasztására irányuló alapvető heurisztikus módszerek:

- **előrelépéses kiválasztás**: üres attribútumhalmazból kiindulva mindig a legjobb, még ki nem választott attribútum hozzáadása
- **visszalépéses eliminálás**: teljes attribútumhalmazból kiindulva mindig a legrosszabb attribútum elhagyása
- **előzőek kombinációja**: meglévő legrosszabb attribútum elhagyása, maradékok közül a legjobb hozzáadása
- **döntési fa alapú módszerek**: a fában meg nem jelenő, nem releváns attribútumok eliminálása

2.2. Faktoranalízis

A heurisztikus eljárásokkal ellentétben a statisztikai adatredukciós eljárások figyelembe veszik az attribútumok közötti összefüggéseket, a korreláció lehetőségét. Az adatredukálás során alkalmazható többváltozós statisztika módszer a faktorelemzés. Míg a heurisztikus dimenziócsökkentéskor az eredeti attribútumok részhalmazának kiválasztása a cél, addig a faktoranalízis a korrelációkra támaszkodva kombinálja az attribútumok tartalmát kisebb számú új változó létrehozásával. Az eredeti adatok ezután erre a szűkebb halmazra vetíthetők.

Az attribútumok lineáris kapcsolata a kovariancia (1) elemzés segítségével vizsgálható meg. Az elemzés eredményeként kapott kovariancia mátrix alapján meghatározható, hogy mely attribútumok változnak azonos irányba.

$$cov(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n-1)} \tag{1}$$

A kovarianciák $[-1, 1]$ tartományba normalizálásával (standardizálás) kapott korreláció (2) elemzés segítségével pedig meghatározható az összefüggés szorossága.

$$r(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{\sigma_x \sigma_y} \tag{2}$$

A statisztikai módszerek a heurisztikus eljárásokkal ellentétben a redukálás során nem csak elhagynak attribútumokat, hanem azokat kevesebb számú változóval helyettesítve a releváns információtartalom megtartásával is operálnak. A kovarianciák alapján ugyanis közvetlenül nem megfigyelhető háttérváltozók (latens változók), azaz faktorok határozhatóak meg.

A faktorelemzés feltárja az eredeti változók egymással szorosabb korrelációban levő csoportjait, így ezek a változók egy faktorhoz tartozónak tekinthetők. A feltárt faktorok lineáris kombinálásával az eredeti adathalmaz több attribútuma csoportosítható és helyettesíthető néhány új, kevesebb számú attribútum képzésével. Így a nagyszámú eredeti változók néhány faktorban, mint új változóban összesíthetők. (Például egy sportoló futó, gyaloglási, ugró, dobó és lökő teljesítményének összevonása egy általános atlétikai képesség faktorban.) A faktoranalízis segítségével elérhető az eredeti változók számának csökkentése amellett, hogy az eredeti adatok leírása a lehető legkevesebb információveszteséggel járjon.

A faktoranalízis lépései:

1. **korrelációelemzés**: a változók egymással való kapcsolatának meghatározása
2. **faktorextrakció**: a faktorok azonosítása a változók közötti korrelációs vizsgálat alapján
3. **faktorsúlyok becslése**: a változók és a faktor közti korreláció meghatározása
4. **interpretáció**: a faktorok értelmezése
5. **faktor rotációja**: megfelelő elforgatással markánsabb jelentést adható a faktoroknak, az eredeti változók korábbi viszonylag nagy faktorterhelései még nagyobbak, a kis faktorterhelések pedig még kisebbek lesznek, így az egyes faktorok könnyebben értelmezhetők

6. **faktorpontok meghatározása**: az egyes esetek jellemzése a faktorpontok segítségével (a faktorok elnevezésekor törekedni kell arra, hogy az elnevezések tükrözzék a legnagyobb faktorsúlyú tételeket)

A faktoranalízis főbb módszerei a főfaktoranalízis, a maximum likelihood-faktoranalízis és a főkomponens elemzés.

3. Főkomponens elemzés

3.1. A módszer bemutatása

A főkomponens-analízis (Principal Component Analysis, PCA) a legelterjedtebb módszer a faktorsúlyok becslésére. A PCA folytonos értékkészletű attribútumok esetén a lineáris algebra módszereit alkalmazza és az eredeti adatok jóval kisebb altérre vetítésével ér el tömörítést.

Az eljárás célja az eredeti, kölcsönös kapcsolatban álló attribútumok transzformációja kevesebb számú, új, olyan független változóvá, amelyek legjobban tükrözik az eredeti adatok varianciáját (3). Lényege az eredeti attribútumok kovariancia (korreláció) struktúrájának a változók minél kevesebb számú lineáris kombinációjával való leírása.

$$\sigma^2(\mathbf{x}) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{(n-1)} \tag{3}$$

A dimenziócsökkentés az eredeti attribútumok új, korrelálatlan változókba (faktorok) történő lineáris transzformálásával érhető el. Az ortogonális transzformáció az adathalmaz korreláltatható változóit lineárisan korrelálatlan változók értékkészletévé alakítja át.

Az eljárás fő lépései:

1. **normalizálás**: eredeti adatok normalizálása
2. **kovariancia struktúra meghatározása**: eredeti adatok variabilitásának jellemzése
3. **transzformáció**: új változók, a faktorok meghatározása lineáris transzformációval
4. **rendezés**: főkomponensek sorba rendezése az adathalmaz varianciájának jellemzésének erőssége szempontjából
5. **eliminálás**:variancia leírása szempontjából gyenge komponensek elhagyása
6. **adathalmaz származtatása**: az eredeti adatok kifejezése a domináns komponensek lineáris kombinálásával

Az első két lépés a kovarianciamátrix előállítását jelenti. Az eljárás harmadik lépése során a PCA olyan faktorokat (komponenseket) tár fel, amelyek:

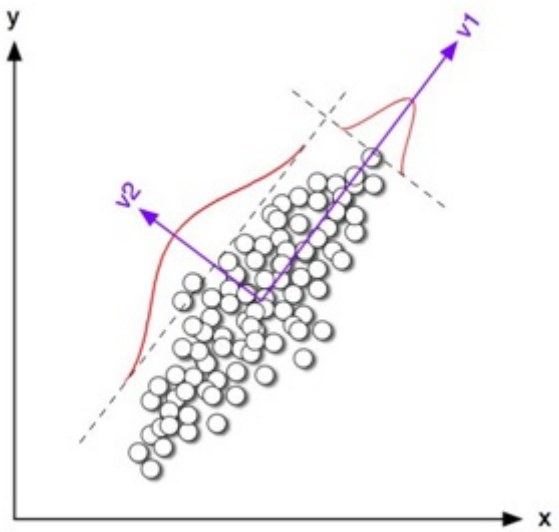
- az eredeti attribútumok lineáris kombinációi és
- ortogonálisak, páronként merőlegesek egymásra és
- adatokban fellelhető ingadozást maximálisan kifejezik és
- minden komponenspár kovarianciája 0

A faktorok előállítását követően a komponensek rendezésre kerülnek úgy, hogy közülük az első néhány az eredeti változók varianciájának jelentős részét megmagyarázza. Az első komponens a lehető legnagyobb varianciát hordozza. A második komponens merőleges az elsőre és ezen kényszerfeltétel mellett a fennmaradó

varianciából a legtöbbet lefedi. Az egymást követő további komponensek a teljes variancia egyre kisebb hányadát magyarázzák.

A komponensek közül csak az első néhány kerül megtartásra, ezek lesznek a főkomponensek, amelyek az eredeti normalizált adatok egy bázisát alkotják. A főkomponensek gyakorlatilag új koordináta tengelyeket jelölnek ki az adatok számára és a fontos információt adnak azok szórásáról.

A PCA a leglényegesebb tulajdonságait keresi meg az adatoknak. Az eljárás működése az adat belső struktúrájának feltárásaként fogható fel, oly módon, hogy az a legjobban magyarázza az adathalmaz szóródását. A főkomponens-analízis a magas dimenziós adattérről egy alacsonyabb dimenziójú képet szolgáltat, a leginformatívabb nézőpontra vetítve az objektumot. Az eredeti adatok az első néhány komponens lineáris kombinációjaként állnak elő.



1. ábra. Példa egy adathalmaz főkomponenseire

A főkomponens-analízis felfogható úgy is, mint egy n dimenziós ellipszoid illesztése az adatokra, ahol az ellipszoid mindegyik tengelye egy főkomponens. Ha az ellipszoid néhány tengelyének hossza alacsony, akkor a tengely menti variancia is alacsony, így a tengely és a hozzá tartozó főkomponens elhagyása az adathalmaz reprezentációjából csak egy ezzel arányosan kis információ mennyiség elvesztését eredményezi.

1. Az ellipszoid tengelyeinek megtalálásához, elsőként ki kell vonni minden változó átlagát az adathalmazból, hogy az adatokat az origó köré igazodjanak.

2. Ezt az adatok kovariancia mátrixának és saját értékeinek, illetve a kovariancia mátrixhoz tartozó sajátvektorok kiszámítása követi.

3. Ezután a sajátvektor halmaz merőlegesítésével (ortogonalizálása) és normalizálásával megkaphatók az egységvektorok. A kapott, kölcsönösen merőleges egységvektorok képezik az ellipszoid adatra illesztett tengelyeit.

4. A varianciák azon hányada, melyet az egyes sajátvektorok képviselnek kiszámítható a sajátvektorhoz tartozó sajátérték és az összes sajátérték összegének hányadosaként.

3.2. Matematikai háttér

Az alábbiakban a főkomponens elemzés formalizált feladata és matematikai eszköztára kerül bemutatásra.

Adott $\mathbf{D}$ nxm-es adatmátrix, amelynek m sora az adatokat, n oszlop pedig az attribútumokat azonosítja (m db n-komponensű adatvektor).

Az eljárás azzal a feltételezéssel él, hogy az attribútumok várható értéke 0, ezért a feltétel teljesülése érdekében minden attribútumból ki kell vonni a saját átlagát.

A $\mathbf{D}$ adatmátrixot $\mathbf{S} = [s_{ij}]$ az adatmátrix varianciáját leíró nxn-es kovarianciamátrix jellemzi, ahol s_{ij} az i-dik és j-dik attribútum kovarianciája. A kovarianciamátrix a változók 0 várható érték feltétel teljesülése esetén megkapható az alábbi formában.

$$\mathbf{S} = \mathbf{D}^T\mathbf{D} \tag{4}$$

A feladat egy olyan $\mathbf{U}$ nxn-es vetítési mátrix megkeresése, amire teljesül

$$\mathbf{D}' = \mathbf{DU} \tag{5}$$

ahol

$\mathbf{D}'$ a transzformált adatok mátrixa

$\mathbf{U}$ vetítési mátrixra igaz, hogy inverze egyenlő annak transzponáltjával, azaz $\mathbf{U}^{-1} = \mathbf{U}^T$

$\mathbf{U}$ oszlopvektorai korrelátlatlanok

$\mathbf{U}$ ortonormális vetítési mátrix

A $\mathbf{U}$ vetítési mátrix tekinthető úgy, mint egy geometriai művelet, amely eltolja és/vagy elforgatja az adatokat és azt a (hiper)síkot teszi láthatóvá, amelynek tengelyei mentén a legnagyobb a variancia. Ez a mátrix tartalmazza a főkomponenseket. A vetítés nem áll meg egy síknál, több tengely is létrejön, tulajdonképpen az összes attribútumot újra generálja.

Gyakorlatban a főkomponensek az adathalmaz attribútumainak $\mathbf{S}$ kovariancia mátrixából számíthatók ki. Az $\mathbf{U}$ transzformáció megkapható $\mathbf{S}$ kovarianciamátrix karakterisztikus értékeiből, a 6. egyenlet megoldásait képező saját értékek és saját vektorok meghatározásával.

$$\mathbf{Su} = \lambda\mathbf{u} \tag{6}$$

Az $\mathbf{u}$ sajátvektor olyan vektor, amellyel egy négyzetes mátrixot szorozva önmaga skalárszorosa adódik. Ez a skalár a λ saját érték.

A saját értékek és saját vektorok a 6. egyenlet átalakításával kapott 7. egyenlet megoldásával határozhatók meg, amely egy n ismeretlenes, n egyenletből álló lineáris homogén egyenletrendszert reprezentál.

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{u} = 0 \tag{7}$$

Az egyenletrendszernek akkor létezik a triviálistól különböző megoldása, ha az együtthatómátrix determinánsa 0 (8).

$$det(\mathbf{S} - \lambda\mathbf{I}) = 0 \tag{8}$$

Az egyenletrendszer $\lambda_1, ..., \lambda_n$ gyökei $\mathbf{S}$ mátrix saját értékei. A saját értékeket a 7. egyenletbe helyettesítve, annak triviálistól különböző megoldásvektorai pedig $\mathbf{S}$ mátrix saját vektorai. A sajátvektor kijelöl egy pontot az n-dimenziós térben, az origó és a pont összekötése pedig egy irányvektort ad. Egy adott sajátvektor tulajdonságát az irányultsága határozza meg a hosszától (és irányától) függetlenül, így egy saját vektort skalárral szorozása ugyanazt a saját vektort adja.

Az n db n dimenziós saját érték csökkenő sorrendbe rendezendő. $\lambda_1, ..., \lambda_n$ saját értékek mindegyike nemnegatív ($\mathbf{S}$ pozitív szemidefinit) és rendezhető úgy, hogy $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ teljesüljön.

A saját vektoroknak a hozzájuk tartozó saját értékek szerint csökkenő sorrendbe rendezése a $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_n]$ **S** sajátvektoraiból álló mátrix, ahol i-dik sajátvektor tartozik i-dik legnagyobb saját értékhez.

A legnagyobb saját értékhez tartozó saját vektor lesz az első főkomponens, a második a második, stb. Ekkor egy ortogonális bázist adódik, melynek az első sajátvektora a legnagyobb szórás irányába mutat az n-dimenziós térben. Ez lesz a keresett transzformáció.

Állítások:

- $\mathbf{D}' = \mathbf{DU}$ teljesíti a feltételeket
- minden új attribútum az eredetiek lineáris kombinációja: az i-dik attribútumot előállító lineáris kombináció súlyai az i-dik sajátvektor komponensei, $\mathbf{D}'$ j-dik oszlopát $\mathbf{Du}_j$ adja
- i-edik új attribútum varianciája λ_i
- a régi és az új attirbútumok varianciájának összege megegyezik

Az új jellemzők a főkomponenseknek. A főkomponensek határozzák meg az adatok vetítésének tengelyeit. Az esetek többségében az első néhány k db főkomponens hordozza a lényeges információkat, a többi elhagyása némi információ vesztéssel jár ugyan, de jelentős veszteség nélkül lényegesen kevesebb adattal jellemezhető az adathalmaz (akkor lehetne biztonságosan eltávolítani egy-egy irányt, ha szórása/saját értéke közelít a 0-hoz).

Végül elvégzendő **D** adatmátrix új dimenziókba vetítése a kiválasztott főkomponensek segítségével, amelynek kiindulóegyenlete a 5. egyenlet. A szorzást elvégezve $\mathbf{D}'$ egy hasonló felosztású adatmátrix, mint a kiinduló **D** adatmátrix, csak kevesebb sor dimenzióval.

Ha nem kerül minden főkomponens felhasználásra, akkor az így kapott közelítés hibája (kumulatív sajátérték-százalék) megkapható a kihagyott főkomponensekhez tartozó sajátértékek szummájaként.

$$\sum_{i=1}^{k} \lambda_i \tag{9}$$

3.3. Értékelés

A PCA egy lineáris leképezés, ami teljes mértékben az adatok szórására támaszkodik. Néhány esetben az adatok szórása azonban nem lényeges információ. Rosszabb esetben elképzelhető, hogy a releváns attribútum teljesen el van rejtve és csak a speciális kivetítést alkalmazva, diszkrimináns elemzés segítségével található meg.

Ennek ellenére a PCA a gyakorlatban eredményesen használható számos területen, mint a dimenziócsökkentés, tömörítés, mintafelismerés és a képfeldolgozás. A módszer a többváltozós kutatások alapjául szolgál, de azokon a területeken alkalmazható leginkább, ahol nagy mennyiségű adattal dolgoznak. A módszer legfőbb alkalmazási területe az adatbányászat, az eljárást elsősorban a feltáró adatelemzésben és a prediktív modellek előállításában használják. A faktorelemzés módszerét alkalmazzák a pszichometriában, a viselkedés- és társadalomtudományokban, használja a szociológia, a marketing, a termékmenedzsment és az operációkutatás is.

Előnyök:

- segít megtalálni az adatokat legjobban jellemző mintázatokat
- mintázatok segítségével a zajok nagy része kiküszöbölhető
- az adatok az információtartalom lényeges csökkenése nélkül, nagy mértékben tömöríthetők
- a főkomponensek és a modellváltozók korrelációs vizsgálatával meghatározhatók, hogy mik a releváns és mik a nem releváns változók

Hátrányok

- faktoranalízis csak annyira lehet jó, amennyire az adatok minősége engedi
- csak intervallum és arány mérési skálán értelmezett (folytonos) változókra alkalmazható (nominális és ordinális típusuaknál nem)
- az eljárás csak lineáris tulajdonságokat képes megtalálni
- kizárólag a szórásra támaszkodik, akkor is, ha az nem lényeges információ
- főkomponensek értelmezése nehézkes, ugyanaz a faktor többféleképpen interpretálható
- eljárás érzékeny az eredeti adatok relatív skálázására
- magas dimenziószám esetén a sajátértékek, saját vektorok meghatározása során numerikus módszereket kell használni, amelyek hibaterjedéssel járnak

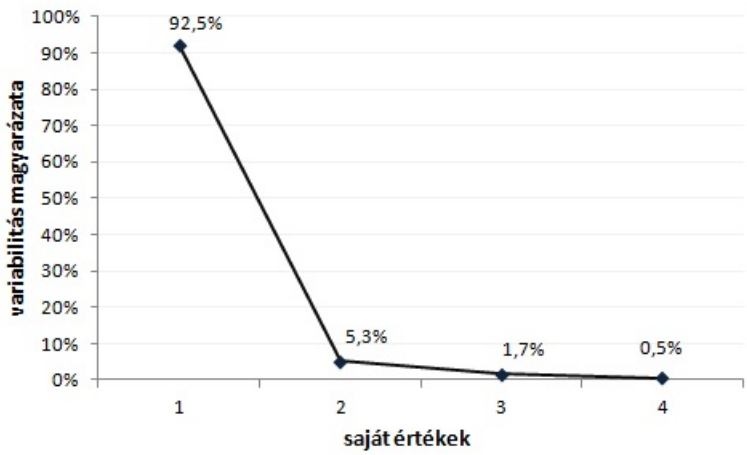
3.4. Az eljárás bemutatása egy példán keresztül

A főkomponens analízis alkalmazásának bemutatása a publikusan elérhető[3] Iris adathalmazra támaszkodik. Az adatkészlet 150 íriszvirágról (nőszirom) tartalmaz információt, 50 egyedről három íriszfajtából. Ez a három fajta a nőszirom (Setosa), a foltos nőszirom (Versicolor), és a virginaiai nőszirom (Virginica).

Minden egyes virágot öt attribútum ír le:

1. a csészelevél hossza centiméterben
2. a csészelevél szélessége centitnéterben
3. a sziromlevél hossza centiméterben
4. a sziromlevél szélessége centiméterben
5. a fajta (Setosa, Versicolor, Virginica)

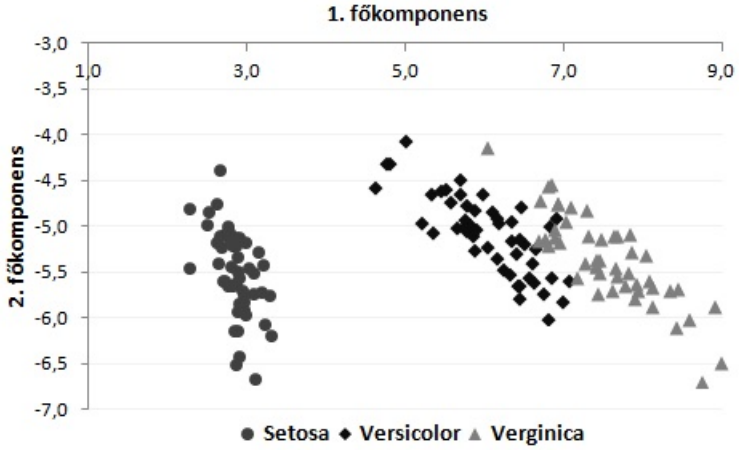
A 2. ábra bemutatja, hogy a kovarianciamátrix egyes saját értékei milyen arányban magyarázzák a teljes szórást.



2. ábra. A minta adathalmaz főkomponenseinek magyarázó ereje

A írisz adatai esetébenaz első főkomponens magyarázza a szórás legnagyobb részét (92,5%), a második csak 5,3%-ot, az utolsó kettő pedig együttesen csak 2,2%-ot. Az első két főkomponens megtartásával az adatsor variabilitásának 97.8%-a megőrizhető.

A 3. ábra az első két főkomponensre vetített adathalmazt szemlélteti, amelyen az egyes fajok már az első főkomponens szerint jól láthatóan elkülönülnek egymástól.



3. ábra. A minta adathalmaz levetítése az első két főkomponensre

Hivatkozások

[1] https://hu.wikipedia.org/wiki/Faktoranal%C3%ADzis

[2] https://hu.wikipedia.org/wiki/F%C5%91komponens-anal%C3%ADzis

[3] http://faculty.smu.edu/tfomby/eco5385/data/Iris.xls

[4] J. Abonyi "Adatbányászat, a hatékonyság eszköze" *Computerbooks*, 2006

[5] J. Han, M. Kamber "Adatbányászat. Koncepciók és technikák" *Panem*, 2004

[6] P.N. Tan, M. Steinbach, V. Kumar "Adatbányászat alapvetés" *Panem*, 2012