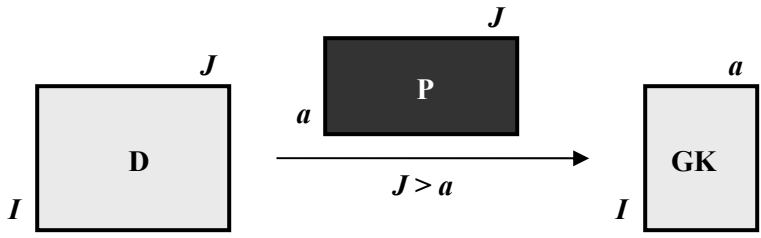
Dr. Pósa Mihály[*], **Szebenyi Anna**[**] és **Dr. Gaál Ferenc**[***]

A főkomponens-elemzés alkalmazása a kémiában

1. Bevezetés

A főkomponens-elemzés (*Principial Component Analysis*, PCA) a molekulaszerkezet - hatás - kvantitatív kutatásának (*Quantitative Structure Activity Relationships* - QSAR) az egyik alapmódszere.[1] A PCA olyan matematikai eljárás, melynek alapvető célja az adatmátrix D rendűségének csökkentése (az adatmátrix változóinak csökkentése).[1-3]



1. ábra
A főkomponens elemzés alkalmazása a D *adatmátrixon*

A főkomponens-elemzés alkalmazásakor a megfigyelt *I* objektumok (molekulák, különböző minták, stb, a D mátrix sorai) a *J* dimenziós vektortérből (a D mátrix oszlopainak a száma, tulajdonképpen az *I* objektumok megfigyelt tulajdonságaik) a P főkomponens mátrixszal (vetítési mátrix, főkomponens-együtthatók mátrixa, *loading*) egy alacsonyabb dimenziójú (*a*, *a<J, a* a „lényeges" főkomponensek száma) vek-

[*] *Dr. Pósa Mihály, egyetemi docens, Újvidéki Egyetem, Orvostudományi Kar, Gyógyszerészettudományi Tanszék, Újvidék*

[**] *Szebenyi Anna, PhD hallgató, tanársegéd, Újvidéki Egyetem, Orvostudományi Kar, Gyógyszerészettudományi Tanszék, Újvidék*

[***] *Dr. Gaál Ferenc akadémikus, a Vajdasági Tudományos és Művészeti Akadémia rendes tagja, nyugalmazott egyetemi tanár, Újvidéki Egyetem, Természettudományi Kar, Kémia Intézet, Újvidék*

tortérbe, a főkomponens-térbe, (GK mátrixba) vetítődnek át. A GK mátrixot általában főkomponens-érték mátrixnak nevezik (*score*).[1,4]

2. A kovarianciamátrix sajátértéke és sajátvektor - a kovarianciamátrix diagonizálása

Ismeretes, hogy a négyzetes és szimmetrikus mátrixok ortogonálisan diagonizálhatók. Mivel a D adatmátrix nem szimmetrikus (általában $I \neq J$), azt először szimmetrikussá kell tenni. Viszont a D adatmátrix kovarianciamátrixa C négyzetes és szimmetrikus mátrix, ahol a D^T mátrix a D adatmátrix transzponált alakját jelöli:

$$\mathbf{C} = \mathbf{D}^\mathbf{T}\mathbf{D} \qquad (1),$$

A D^TD kovarianciamátrix ortogonális diagonalizációját a következő mátrixegyenlettel ábrázolható:

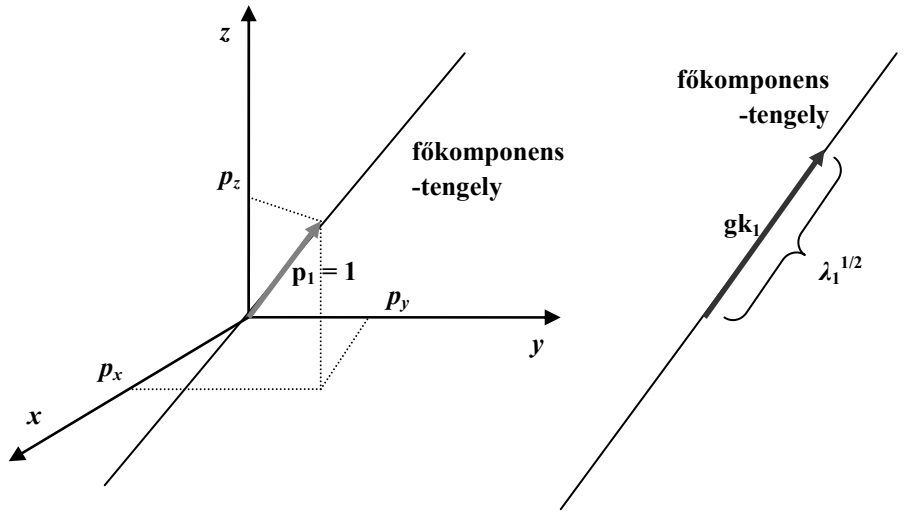
$$\mathbf{D}^\mathbf{T}\mathbf{D} = \mathbf{P}\Lambda\mathbf{P}^\mathbf{T} \qquad (2).$$

A felső mátrixegyenletben P jelöli a D^TD kovarianciamátrix ortonormált sajátvektorainak (*eigenvectors*) mátrixát (a sajátvektorok a P mátrix oszlopvektorai), míg a Λ a D^TD mátrix sajátértékeinek λ (*eigenvalues*) a diagonális (átlós) mátrixa. Ha a D^TD kovarianciamátrix nem elfajult (reguláris, nem szinguláris), akkor *J* darab sajátvektora, illetve sajátértéke van, így a P és Λ mátrixok *J* rendű négyzetes mátrixok. Abban az esetben, ha a D^TD kovarianciamátrix szinguláris (a D^TD mátrix determinánsa nulla), *r* rangú mátrix és *r* számú sajátértéke, illetve sajátvektora van. A főkomponens-elemzés elméletében a P mátrix, tehát a D^TD sajátvektorainak a mátrixa, a főkomponens-vektorok p mátrixa. A főkomponens-vektorok p a P mátrix oszlopvektorai a *J* vagy *r* dimenziós vektortérben. Helytálló a P mátrixnak a főkomponens-együtthatók mátrixának az elnevezése is, mivel a főkomponens-vektorok koordinátasorai (koordinátaegyütthatói) meghatározzák a p vektorok helyzetét (irányát) a vizsgált objektumok tulajdonságainak vektorterében. A p vektorok valójában egységvektorok, amelyek egymásra merőlegesek és irányuk meghatározza a főkomponens-tengelyt.[1]

A D^TD kovarianciamátrix ortogonális diagonalizációjaként kapott P mátrix után következik a főkomponens-érték mátrix meghatározása:

$$\mathbf{gk}_{[I\times1]} = \mathbf{D}_{[I\times J]}\mathbf{p}_{[J\times1]} \qquad (3).$$

A 3. egyenlet szerint, minden p vektor az *I* objektumokat a D adatmátrix *J* dimenziós vektorteréből az általa (p) meghatározott főtengelyre vetíti (amely kolineáris a p egységvektorral, és mindegyik főtengely merőleges minden más főtengellyel). Tehát, a főkomponens-érték vektorok gk az *I* objektumok koordinátái a p által meghatározott főtengelyen. A főkomponens-érték vektorok ortogonálisak, de nem egységvektorok, hanem értékük egyenlő a velük komplementer sajátérték négyzetgyökével[«] (2 ábra).[1,3]



2. ábra

A főkomponens-vektor $\mathbf{p}_1$ *egységvektor az x,y,z változok (a megfigyelt objektumok tulajdonságaik) terében, azaz a jelen példában az* R^3 *vektortérben. A* $\mathbf{p}_1$ *koordinátai meghatározzák a főtengely irányát, amelyre vetítődnek az* R^3 *térből*

« $\mathbf{gk}_i\,\mathbf{gk}_j = \mathbf{Dp}_i\,\mathbf{Dp}_j = (\mathbf{Dp}_i)^T\,\mathbf{Dp}_j = \mathbf{p}_i^T\mathbf{D}^T\mathbf{Dp}_j$
$\mathbf{D}^T\mathbf{Dp} = \lambda\mathbf{p}$
$\mathbf{p}_i^T\mathbf{D}^T\mathbf{Dp}_j = \mathbf{p}_i^T\,\lambda_j\,\mathbf{p}_j = \lambda_j\,\mathbf{p}_i\,\mathbf{p}_j$
$\mathbf{p}_i^T\mathbf{D}^T\mathbf{Dp}_i = \mathbf{p}_i^T\,\lambda_i\,\mathbf{p}_i = \lambda_i\,\mathbf{p}_i\,\mathbf{p}_i$, $(\mathbf{p}_i\,\mathbf{p}_i = 0) \rightarrow \mathbf{gk}_i\,\mathbf{gk}_i = \lambda_i \rightarrow \|\mathrm{gk}_i\| = \lambda_i^{1/2}$.

a tanulmányozott objektumok. Az I objektumok főtengelybéli értékei meghatározzák a főkomponens-érték vektort gk_1

A főkomponens-érték mátrixot GK a 3-as kifejezés értelmében a következő mátrixegyenlet határozza meg:

$$\mathbf{GK} = \mathbf{DP} \quad (4).$$

A főkomponens-térben az *I* objektumok értékei alapján a következő kovarianciamátrixot lehet definiálni (főkomponensek kovarianciája):

$$\mathbf{C} = \frac{1}{I-1}\mathbf{GK}^{\mathbf{T}}\mathbf{GK} \quad (5).$$

Ha az 5-ös kifejezésben a GK helyett a 4-es mátrixegyenlet szerinti DP helyettesítjük, majd figyelembe vesszük a 2-es kifejezést, akkor a kovarianciamátrixra a következőt kapjuk[«]:

$$\frac{1}{I-1}\mathbf{GK}^{\mathbf{T}}\mathbf{GK} = \frac{1}{I-1}(\mathbf{DP})^{\mathbf{T}}\mathbf{DP} = \mathbf{P}^{\mathbf{T}}\mathbf{D}^{\mathbf{T}}\mathbf{DP} = \frac{1}{I-1}\mathbf{P}^{\mathbf{T}}\mathbf{P}\Lambda\mathbf{P}^{\mathbf{T}}\mathbf{P} = \frac{1}{I-1}\mathbf{\Lambda} \quad (6).$$

A 6-os kifejezés értelmében a főkomponensek kovarianciamátrixában az átlós elemek a főkomponensérték-vektorok, mint az *I* objektumok főkomponens-változóinak a szórásnégyzetei $s^2(\mathrm{gk}_i)$, míg a nem átlós elemek nullával egyenlők:[1]

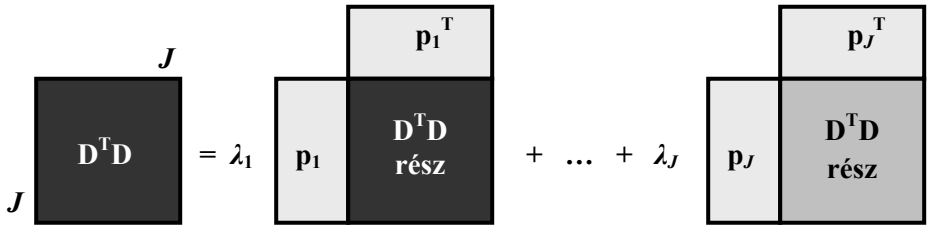
$$\begin{bmatrix} s^2(\mathbf{gk}_1) & & & 0 \\ & \cdots & & \\ & & s^2(\mathbf{gk}_i) & \\ & & & \cdots \\ 0 & & & s^2(\mathbf{gk}_J) \end{bmatrix} = \frac{1}{I-1}\begin{bmatrix} \lambda_1 & & & 0 \\ & \cdots & & \\ & & \lambda_i & \\ & & & \cdots \\ 0 & & & \lambda_J \end{bmatrix}$$

[«] Mivel a **P** ortonormált mátrix, ezért $\mathbf{P}^{\mathbf{T}} = \mathbf{P}^{-1}$, azaz $\mathbf{P}^{-1}\mathbf{P} = \mathbf{P}\mathbf{P}^{-1} = \mathbf{I}$; **I** pedig az egységmátrix.

Bebizonyították, hogy a $s^2(gk_i)$ szórásnégyzetek az eredeti adatmátrix változóinak (a D mátrix oszlopainak) a szórásnégyzeteiből erednek.[1] A D^TD kovarianciamátrix kifejezhető a főkomponens-vektorok p sajátértékeikkel súlyzott dijádikus szorzatuk összegével (a D^TD kovariancia-mátrix spektrális felbontása, 3 ábra):[1,4]

$$\mathbf{D^TD} = \lambda_1\mathbf{p}_1\mathbf{p}_1^\mathbf{T} + \lambda_1\mathbf{p}_2\mathbf{p}_2^\mathbf{T} + \cdots \lambda_i\mathbf{p}_i\mathbf{p}_i^\mathbf{T} \cdots + \lambda_J\mathbf{p}_J\mathbf{p}_j^\mathbf{T} \qquad (7).$$

Mivel a sajátértékek közül a λ_1-nek van a legnagyobb értéke, a 7-es spektrális felbontásban az első dijádikus szorzat ad magyarázatot a legnagyobb mértékben az adatmátrix kovarianciamátrixára. Ezek szerint az eredeti adatok varianciájából legtöbbet az első főkomponens magyaráz.



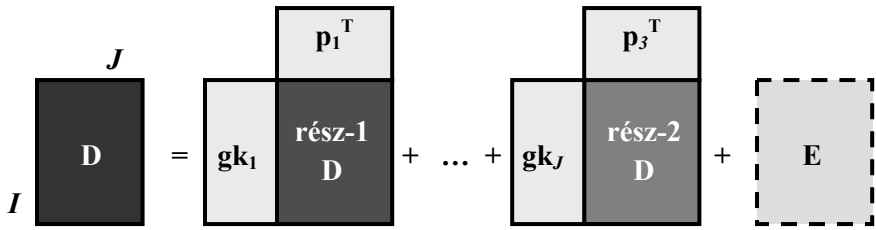
3. ábra
A D^TD *adatmátrix spektrális felbontása Falco mátrixszorzási sémája szerint*

A főkomponens-érték mátrixban GK a főkomponensérték-vektorok gk (oszlopvektorok) úgy sorakoznak, hogy balról jobbra haladva csökken az általuk megmagyarázott D^TD mátrix varianciájának a hányada. Így az első főkomponens-érték vektor gk_1 felöleli a D^TD mátrix varianciájának a legnagyobb részét, míg a második főkomponens-érték vektor gk_2 mutatja a D^TD mátrix varianciájának azt a részét, amelyet az első főkomponens-vektor nem utal. A harmadik főkomponens-érték vektor gk_3 pedig a D^TD mátrix azon varianciáját mutatja, amelyet az első kettő főkomponens-érték vektor nem utalt, stb. *Wold* szerint azok a főkomponensérték-vektorok amelyek sajátértékei alacsonyak, elhagyhatók a GK mátrixból, mivel ezek a főkomponens-érték vektorok az eredeti

adatmátrixban azt a szórást mutatják, amely a zajból vagy mérési hibából származik:[1,4]

$$\mathbf{D} = \mathbf{gk_1p_1}^T + \mathbf{gk_2p_2}^T + \mathbf{gk_3p_3}^T + \mathbf{E} = \mathbf{D_{rész1}} + \mathbf{D_{rész2}} + \mathbf{D_{rész3}} + \mathbf{E} \quad (8).$$

A 8-as kifejezésben az E a hibamátrix, amely tulajdonképpen az elhagyott kis sajátértékű főkomponens-érték vektorokban foglaltatott (4. ábra).



4. ábra
A Wold elmélet Falco-féle mátrix sémája

3. Alakfelismerés (pattern recognition) a főkomponens-érték vektorok terében

A két- (gk_1-gk_2) vagy három dimenziós (gk_1-gk_2-gk_3) főkomponens-értek vektortérben a hasonló tulajdonságú molekulák (minták) csoportokat alkotnak, így például a homológ vegyületek lineáris kongenerikus csoportokat.[1,5] *Pósa* és társai vizsgálták a hőmérséklet hatását az epesavak retenciós kapacitására reverz fázisú nagyhatékonyságú folyadék kromatográfiában (*reverse phase high performance liquid chromatography* RPHPLC). A kísérlet célja az volt, hogy kivizsgálják vajon a hőmérsékletváltozás a kromatográfiában hordoz-e olyan többlet információt, amely specifikus az epesavak szerkezetére, különösen a szteroidvázat illetően, ami főkomponens-elemzéssel leellenőrizhető. A kísérletben az adatmátrix D a következő volt: az epesav molekulák képviselték az objektumokat (a mátrix sorai), míg a különböző hőmérséklet értékekhez tartozó retenciós kapacitások reprezentálták a mátrix oszlopait (4. ábra). A D adatmátrix kovarianciamátrixa határozta meg a főkom-

ponenseket. Az első két főkomponens az eredeti adatmátrix D varianciájának a 99%-ra adott magyarázatot.[5]

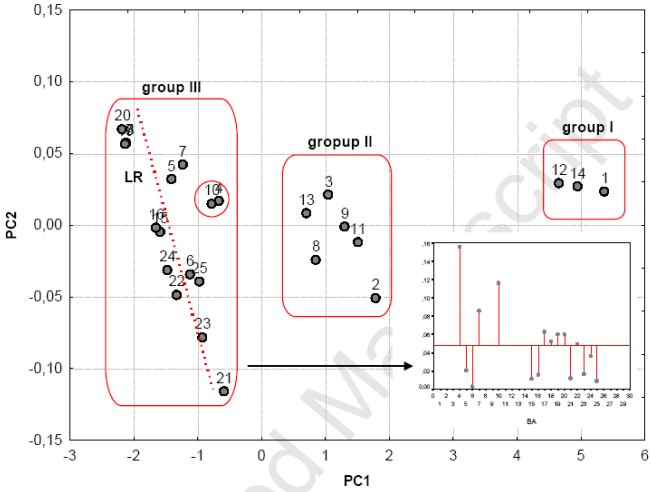
bile acids		20	25	30	35	40	45
		[°C]					
1	L	29.32	25.25	20.73	16.86	14.18	11.53
2	D	15.09	13.28	10.87	9.34	7.71	6.43
3	CD	12.83	10.87	8.88	7.45	6.33	5.27
4	C	6.02	5.21	4.34	3.69	3.19	2.74
5	UD	3.24	2.74	2.36	2.05	1.78	1.50
6	HD	4.22	3.49	2.96	2.83	2.51	2.03
7	HC	3.92	3.40	2.83	2.41	2.07	1.77
8	G-CD	11.41	10.23	8.63	7.23	5.91	5.00
9	T-D	13.67	11.67	9.54	8.05	6.84	5.79
10	G-C	5.58	4.83	4.04	3.46	2.98	2.56
11	G-D	14.54	12.15	10.11	8.53	7.24	6.13
12	T-L	27.43	22.37	18.43	15.38	12.91	10.8
13	T-CD	11.53	9.52	7.97	6.73	5.72	4.71
14	G-C	28.69	23.23	19.17	16.05	13.45	10.9
15	12-OxC	2.41	2.08	1.82	1.77	1.54	1.33
16	7-OxC	2.18	1.89	1.66	1.62	1.43	1.23
17	7,12-dOxC	0.50	0.47	0.44	0.42	0.39	0.36
18	3,7-dOxC	0.56	0.51	0.46	0.45	0.41	0.35
19	3,12-dOxC	0.50	0.47	0.43	0.42	0.39	0.34
20	3,7,12-tOxC	0.40	0.35	0.33	0.32	0.28	0.25
21	12-OxD	5.91	5.03	4.35	4.23	3.65	3.13
22	3,12-dOxD	3.22	2.81	2.47	2.42	2.13	1.85
23	7-OxCD	4.72	4.02	3.51	3.40	2.93	2.53
24	3,7-dOxCD	2.73	2.38	2.12	2.06	1.83	1.59
25	6-OxHD	4.75	3.95	3.38	3.23	2.73	2.32

4. ábra

Az RPHPLC-beli kisérletben használt adatmátrix (Pósa, M., Pilipović, A., Lalić, M., Popović, J. Talanta doi: 10.1016/j.talanta.2010.11.050 (in press))

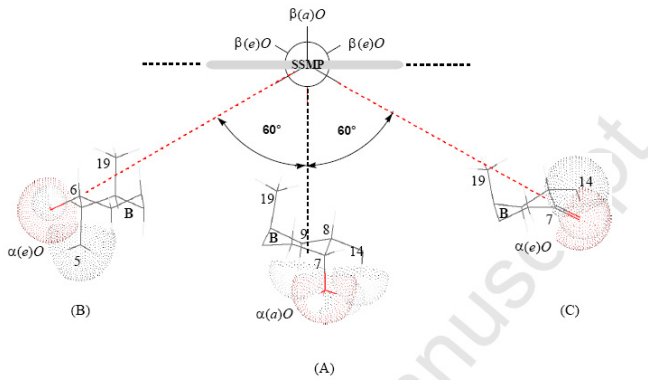
Az első két főkomponensérték-vektor síkjában (5. ábra) az epesavak három csoportot képeznek. Az első csoportra jellemző, hogy az epesav molekulák szteroidrendszerükben csak egy hidroxid csoportot tartalmaznak (C3-as OH), míg a második csoportban kettőt (C3-as OH, C7-es vagy a C12-es OH). A harmadik csoportba olyan epesav molekulák tartoznak, melyek szteroidváza oxo csoportot is tartalmaz, vagy az oxo csoporttal azonos térbeli helyzetű hidroxid csoportot (6. ábra). A három hidroxid csoportot tartalmazó epesavak a *Cook*-féle távolságaik

alapján nem tartoznak egyik csoportba sem, hanem különálló szigetet alkotnak.[5]



5. ábra

Az első két főkomponens-érték vektor síkjában történő csoportosulások (Pósa, M., Pilipović, A., Lalić, M., Popović, J. Talanta doi: doi.1016/j.talanta.2010.11.050 (in press))

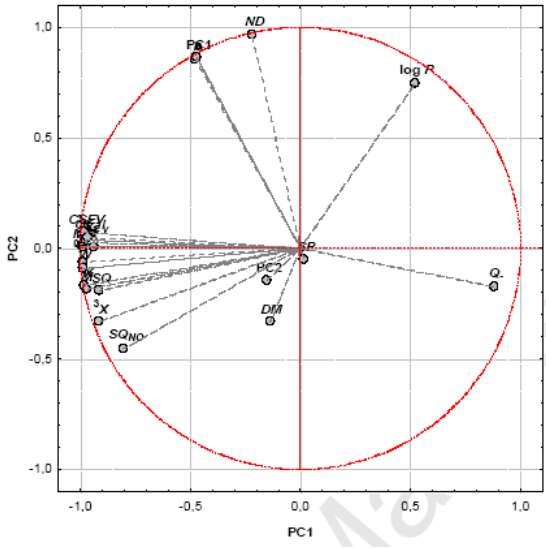


6. ábra

Az 5-ös ábra epesav molekulák III csoportjának a szteroidvázuk közös sajátossága Newman-projekcióban ábrázolva (Pósa, M., Pilipović, A., Lalić, M., Popović, J. Talanta doi: 10.1016/j.talanta.2010.11.050 (in press))

4. Az adatmátrix D *változóinak (az objektumok tulajdonságainak) korrelációja a főkomponens-vektorok (főkomponens-együtthatók) alapján*

Mivel a főkomponens-vektorok p együtthatói az eredeti vektortérben értelmezett koordinátaértékek, ezért ha az első két főkomponens-vektort egymás függvényében ábrázoljuk, akkor az így kapott síkban minden eredeti tulajdonságnak (a D mátrix oszlopváltozójának – tulajdonságvektorának) egy rádiuszvektor felel meg. A rádiuszvektorok egymással bezárt szögeinek koszinusza arányos a rádiuszvektorokat reprezentáló változók közötti korrelációval. Minél kisebb a két helyvektor által bezárt szög, annál nagyobb a vektorokhoz rendelt tulajdonságok közötti korreláció. Az ortogonális rádiuszvektorok közötti korreláció nulla (7. ábra). Az adatmátrixból, a további prediktív (előrejelző) modell kidolgozásakor azok a tulajdonságvektorok közül, amelyek rádiuszvektorai a főkomponens-vektorok síkjában egymással éles szöget zárnak be, többségük elhagyható.[1,3-5]



7. ábra

Az adatmátrix tulajdonságvektorainak korrelációja az első kettő főkomponens-vektor síkjában(Pósa, M., Pilipović, A., Lalić, M., Popović, J. Talanta doi: 10.1016/j.talanta.2010.11.050 (in press))

5. Összegzés

A főkomponens-elemzés olyan többváltozós adatelemzés, amelynél egy időben kapunk információt mind az adatmátrix objektumainak (molekulák) egymás közötti hasonlóságairól, mind a tulajdonságvektorok korrelációjáról.

Felhasznált irodalom:

1. Pósa, M., 2010. Osnovne metode u hemometriji. Medicinski fakultet, Novi Sad.
2. Otto, M., 2007. Chemometrics, Willey-VCH, Weinheim.
3. King, F.D., 2002. Medicinal Chemistry Principles and Practice, R.S.C., Cambridge.
4. Horvai, Gy., 2004. Chemometrics. Budapest, Nemzeti Tankönyvkiadó.
5. Pósa, M., Pilipović, A., Lalić, M., Popović, J., 2011. Determination and importance of temperature dependence of retention coefficient (RPHPLC) in QSAR model of nitrazepams' partition coefficient in bile acid micelles. Talanta doi: 10.1016/j.talanta.2010.11.050 (in press)