

rajko.robert_14_22

MTA Doktora Pályázat
Értekezés

GÖRBEILLESZTÉS NÉLKÜLI KOMPONENSPROFIL-KINYERÉS
AZ ANALITIKAI KÉMIÁBAN

Rajkó Róbert

Szeged
2022-23

Tartalomjegyzék

| | |
|--|----|
| 1. Bevezetés | 4 |
| 2. Célkitűzés | 6 |
| 3. Irodalmi áttekintés | 8 |
| 4. Mérés adatok modellezése: elméleti fejlesztések és alkalmazások | 17 |
| 4.1 A megengedett megoldás-tartományok analitikus meghatározása számítógépes geometriai módszerekkel [P1]. | 17 |
| 4.2 Természetes dualitás a görbeillesztés nélküli komponensprofil-kinyerés során minimális kényszerfeltételeket alkalmazva [P2] | 22 |
| 4.3 A megengedett megoldás-tartományok határainak számítása és a forgatási bizonytalanság mértékének meghatározása [P3] | 24 |
| 4.4 Különböző Borgen-normák alkalmazhatóságának vizsgálata [P4] | 27 |
| 4.5 Elméleti fejtegetések kétkomponensű rendszerek megengedett megoldás-tartomány határainak értelmezéséhez és számításához [P5] | 32 |
| 4.6 Ismert-profil kényszerfeltétel hatása a forgatási bizonytalanság csökkenésére [P6] | 36 |
| 4.7 Újabb fejlesztések a görbeillesztés nélküli komponensprofil-kinyerésben az ismert-profil és unimodalitás kényszerfeltételek alkalmazásával [P7] | 39 |
| 4.8 Adat-alapú egyértelműen meghatározható profilbecslés definíciója és kimutatása [P8] | 45 |

| | |
|--|----|
| 4.9 Újabb keletű görbeillesztés nélküli komponensprofil- kinyerő módszerek áttekintése [P9] | 49 |
| 4.10 Két- és háromkomponensű három-utas adattömbök nemnegatív felbonthatóságának egyértelműsége [P10] | 51 |
| 4.11 Megengedett megoldás-tartományok analitikus megadása és értelmezése kétkomponensű három-utas adattömbök esetén [P11] | 55 |
| 4.12 Az L_0 , L_1 és L_2 normák vizsgálata ritkás profil- mátrixokra történő nemnegatív többváltozós görbefelbontás során [P12] | 59 |
| 5. Mérési adatok kiértékelése: gyakorlati fejlesztések és alkalmazások | 62 |
| 5.1 Alternatív érvényesítési eljárás a túlillesztés elkerülésére [P13] | 62 |
| 5.2 Az MCR-ALS algoritmus néhány meglepő gyakorlati tulajdonságáról [P14] | 65 |
| 5.3 Kritikai megjegyzések NIR hiperspektrális adatok kiértékelése során alkalmazott minimális térfogat kritériumra vonatkozóan [P15] | 67 |
| 5.4 HPLC/FT-IR mérések során az eluens-hatás elimináció kemometriai megvalósítása [P16] | 70 |
| 5.5 A meloxikám-mannit binér rendszer fiziko-kémiai jellemzése [P17] | 72 |
| 5.6 Görbeillesztés nélküli komponensprofil-kinyerő módszer alkalmazása meloxikám-mannit binér rendszerek kioldódási vizsgálata során [P18] | 73 |
| 5.7 A lokális rang-vesztés értelmezése és vizsgálata [P19] | 76 |

| | |
|---|-----|
| 5.8 Dualitás alapú általánosított rang-eltüntető algoritmus koncentráció meghatározáshoz, realisztikus hibabecsléssel [P20] | 80 |
| 6. Az új tudományos eredmények tézisszerűen rendezett összefoglalása | 84 |
| 7. Az eredmények más hasznosításának lehetőségei, újabb kutatási irányok, tervek | 96 |
| 8. Irodalomjegyzék | 105 |

1. Bevezetés

A *kemometria* a többváltozós/többdimenziós kémiai vagy hasonló jellegű mérési adatok kiértékelésére szolgáló módszerek összességét jelenti [1-3]. Egy kicsit részletesebben, a kemometria a kémiai és analitikai kémiai információtartalom minél teljesebb (közel optimális) kinyeréséhez szükséges matematikai statisztikai, lineáris algebrai, konvex geometriai, információtechnológiai és számítástechnikai, ill. formál logikai módszerek és eljárások célirányos alkalmazásait foglalja magába. A kemometria tehát a kémia, a matematika és az informatika *mérési adat*-központú határtudománya.

A kemometria nem tartozik az elméleti kémia tudományterülethez, mivel nem egy elvont modell irányából közelít (pl. a Schrödinger-egyenlet mint posztulátum alkalmazása a kvantumkémiaiában [4]), hanem a mérési adatok felől. A két diszciplína közötti kapcsolat azonban megteremthető, hiszen az elméleti vagy félempirikus úton meghatározott molekula-deszkriptorok és valamilyen mérhető (kémiai, biológiai, klinikai) hatás, aktivitás stb. közötti empirikus modellalkotásban kemometriai módszereket alkalmazhatunk. Az összevont tudományterület elnevezése kemoinformatika (kémiai informatika) lett [5]. A *matematikai kémia* elnevezés/diszciplína összefoglalhatja a kemometria és az elméleti kémia módszereit, bár eddig inkább az utóbbi szinonimájaként alkalmazták [6].

A kemometria tehát adatközpontú tudományterület, mégis vannak olyan elméleti jellegű kemometriai levezetések, amikor az adatokat egyes tulajdonságaik alapján általánosítjuk és így konkrét mérési adattól függetlenül végezhetjük elméleti vizsgálódásainkat. Nagyon gyakran számítógéppel szimulált „mérési” adatokkal (is) dolgozunk. Az elméleti kemometria keretében, a molekulamodellezés elnevezés analógiájára, a továbbiakban adatmodellezésnek nevezzük az előbb vázolt tevékenységet (az informatikában használt adatmodell fogalom eltér az előzőekben leírtaktól, hiszen ott egy logikai adatbázishoz rendelik és programozási technikákkal hozzák kapcsolatba [7]).

Jelen értekezés szerzője még egyetemi hallgatóként találkozott a kemometriával: diplomadolgozatom [8], tudományos diákkörös dolgozatom [9], későbbi doktori disszertációim [10,11], ill. habilitációs anyagom [12,13] is e tudományterületről vett problémákkal és feladatokkal foglalkozott.

Tudományos tevékenységem során kezdetben konkrét egy és többváltozós mérési adatok kiértékelésével, ill. az ehhez kapcsolódó kalibráció matematikai statisztikai elemzésével foglalkoztam; később a változószelekcióval kapcsolatos fejlesztéseket, ill. kvantitatív szerkezet-hatás összefüggés (QSAR) vizsgálatokat végeztem. Az utóbbi időben a táblázatba rendezhető kétdimenziójú (mátrix), ill. magasabb dimenziójú (tenzor) adattömbökkel kapcsolatos rejtett-változójú adatkiértékelő, valamint adatmodellező eljárások elméleti vizsgálatát és gyakorlati alkalmazhatóságuk tanulmányozását végzem.

A továbbiakban az értekezés címének megfelelően, a görbeillesztés nélküli komponensprofil-kinyerés (mint a kemometria egyik meghatározó területe) témakörében megjelent publikációim eredményeit mutatom be.

2. Célkitűzés

Az értekezés címének megfelelően, a görbeillesztés nélküli komponensprofil-kinyerés (mint a kemometria egyik meghatározó területe) témakörében megjelent publikációimat kívánom összefoglalni önálló, jól meghatározott tudományos kérdés (görbeillesztés nélküli komponensprofil-kinyerés) megoldásával foglalkozó tudományos dolgozataimban megjelent eredményekkel.

A mérési adatok modellezése kapcsán első körben elméleti fejlesztéseket tűztem ki magamnak, amely Lawton és Sylvestre 1970-ben, ill. Borgen 1985-86-ban publikált eredményeinek tovább gondolásával indult a megengedett megoldás-tartományok számítógépes geometriai módszerekkel történő analitikus meghatározásával.

Azután Henry cikkének eredményét vizsgáltam meg és a természetes dualitás igazolását szándékoztam elvégezni minimális kényszer-feltételeket alkalmazva.

Tauler cikkére reagálva a megengedett megoldás-tartományok határainak számítását és a forgatási bizonytalanság meghatározását kívántam tüzetesen elemezni. Felismerve, hogy az LS módszer L_1 normán, a Borgen módszer viszont az első-sajátvektor-értékei-1 normán alapul, a Borgen normák általános megfogalmazását és vizsgálatát gondoltam megvalósítani.

Abdollahi és mtsai kétkomponensű rendszereket vizsgált numerikus módszerekkel (*mcrbands* szkript és rácsmódszer), viszont én felismertem, hogy a feladat elvégezhető analitikus levezetésekkel is.

Ismert-profil kényszerfeltétel alkalmazását határoztuk el fajtaalapú részecske-keraj-optimalizálás alkalmazásával.

Tovább lépve, nemcsak az ismert-profil, hanem az unimodalitás kényszerfeltételek alkalmazása és a Borgen módszer dualitáson alapuló kidolgozása mellett döntötünk.

A Borgen módszer alkalmazása néha megmagyarázhatatlan hibaüzenettel végződött, melynek kivizsgálását kívántuk elvégezni és ennek kapcsán az adat-alapú egyértelműen meghatározható profilbecslés definícióját és kimutatását kidolgozni.

Az MCR és SMCR területen dolgozó néhány nemzetközi kiválósággal karöltve egy áttekintő (review) cikk megírását határoztuk el.

Két- és háromkomponensű három-utas adattömbök nemnegatív egyértelmű felbontóságának boncolgatását választottuk következő feladatként.

Tovább lépve, kétkomponensű három-utas adattömbökre analitikus megoldástartományt meghatározó módszer kifejlesztését céloztuk meg.

Az elméleti blokk lezárásaként, ritkás profil-mátrixokra történő nemnegatív többváltozós görbefelbontás során az L_0 , L_1 és L_2 normák szerepét kívántuk tisztázni.

A mérési adatok kiértékeléséhez kapcsolódó gyakorlati fejlesztéseket és alkalmazásokat a PLS regresszióhoz, a túlillesztés elkerülésére, alternatív érvényesítési eljárás kidolgozásával indítottuk.

A következőkben az MCR-ALS algoritmus néhány meglepő gyakorlati tulajdonságainak összegyűjtését és elemzését határoztam el.

Lopes et al. cikkével kapcsolatban kritikai megjegyzéseimnek publikálását döntöttem el.

HPLC/FT-IR mérések során az eluens-hatás kemometriai eliminációját céloztuk meg.

A továbbiakban a meloxikám (ME) és a mannit (MA) mint hordozóanyag fizikai és olvadék keverékének kioldódási viselkedésének jellemzését választottuk.

Folytatásként a meloxikám-mannit binér rendszerek kioldódási vizsgálata során a görbeillesztés nélküli komponensprofil-kinyerő módszer alkalmazhatóságának vizsgálata mellett döntöttünk.

Érdekes újdonságnak ígérkezett a lokális rang kényszerfeltétel vizsgálata és a lokális rang-vesztés fogalmának bevezetése és értelemzése.

Befejezésként, kalibrációs mintákból hiányzó zavaró komponensek jelenléte esetén koncentráció meghatározáshoz dualitás alapú általánosított rang-eltüntető algoritmus kifejlesztése vetődött fel, realiztikus hibabecsléssel, mivel pl. ez utóbbival az akkor rendelkezésre álló GRAM módszerek nem rendelkeztek.

3. Irodalmi áttekintés

Az analitikai kémiai mérések során jelentős esetben kapunk mátrixba vagy általánosan többdimenziós adattömbbe rendezhető mérési adatokat. Ezek kiértékelésére, rendszerezésére, elméleti vizsgálatára alakult ki az 1970-es évektől kezdve a kemometria mint kémiai diszciplína, melynek létrejötté Wold, Kowalski és Massart kutatók nevéhez kötődik [14,15]. Említésre méltó, hogy Magyarországon az 1980-as évek elejétől szerveznek Kemometria elnevezésű szemináriumokat, szimpóziumokat, majd 1997-ben Conferentia Chemometrica '97 néven az első hazánkban rendezett nemzetközi konferencia is sikeresen megvalósult.

Wallace már 1960-ban megfogalmazta azt az adatmátrixokra kifejtett alapvetést, ami a kemometria alapkövét jelentette: "Ha az elegyben lévő minden komponensnek különböző a spektruma és a kifejlődő (elúciós) koncentráció profilja, akkor az adatmátrix *algebrai rangja* (*pszeudorangja* zaj esetén) és a rendszerben lévő *komponensek száma* között a kapcsolat egyértelmű. Az egyetlen feltétel a Bouguer-Lambert-Beer-törvény érvényesülése." Kezdetben tehát a lineáris algebra, de főként a mátrixaritmetika vizsgálata jelentette a fő irányvonalat, mennyire lehet ezen elveket analitikai kémiai mérések adatainak kiértékelésére használni.

A legfontosabb alapmódszer a *szinguláris érték felbontás* (Singular Value Decomposition, SVD):

$$\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{V}^T \quad (1)$$

$I \times J$ $I \times I$ $I \times J$ $J \times J$

ahol \mathbf{A} az adatmátrix: pl. egy spektroszkópiai mérés során az adott mintákra kapott, több hullámhosszon mért abszorbanancia értékek; \mathbf{U} egy ortonormált mátrix ($\mathbf{U}^T \mathbf{U} = \mathbf{I}$, ahol \mathbf{I} az egységmátrix, a főátlóban csupa 1-es, azon kívül csak nulla szerepel); \mathbf{U} az \mathbf{A} mátrix baloldali szinguláris-vektor mátrixa (ortonormált); \mathbf{D} egy olyan (általában téglalap alakú) mátrix ahol a főátlón kívüli összes elem zérus, a főátlóban lévő elemek pedig a szinguláris értékek, amelyek csökkenő értékben vannak rendezve; \mathbf{V} egy újabb ortonormált mátrix: az \mathbf{A} mátrix jobboldali szingu-

lárís-vektor mátrixa. A T a tranzponálás jele: a sorok és oszlopok szerepének felcserélése, azaz a mátrix elemeinek (esetleg a meghosszabbított) főátlóra való tükrözése. Az \mathbf{A} mátrix alatti első index, azaz I pl. a minták számát jelöli (I db sorból áll az \mathbf{A} mátrix), míg a második index, azaz J pl. az alkalmazott hullámhosszak számát jelenti (J db oszlopból áll az \mathbf{A} mátrix); az $I \times J$ jelölés az \mathbf{A} mátrix méretére utal. Általában a nagyon kicsi szinguláris értékhez tartozó szinguláris vektorokat elhagyjuk, így dimenzióredukciót érhetünk el. Ekkor a következő közelítést kapjuk:

$$\mathbf{A}_{I \times J} = \mathbf{U}_{I \times N}^* \mathbf{D}_{N \times N}^* \mathbf{V}_{N \times J}^{*T} + \mathbf{R}_{I \times J} \quad (2)$$

ahol \mathbf{R} a reziduális (maradék) mátrix (gyakran a mérési hibához (zajhoz) rendelhető, az értékelhető jelinformációtól elkülönült rész); az N a benntartott komponensek száma lesz ($N+1$ -től I -ig, ill. J -ig a szinguláris vektorokat és értékeket elhagyjuk), a $*$ jelölés különbözteti meg a redukálás után kapott kisebb méretű mátrixokat. A $\mathbf{D}_{N \times N}^*$ jelölés egy $N \times N$ méretű (tehát négyzetes) diagonális mátrixot jelent (a főátlón kívüli összes elem zérus).

A következő lépésben, a *főkomponens elemzés* (Principal Component Analysis, PCA) eredménye alakítható ki az SVD felhasználásával. A főkomponens elemzés alap gondolata az, hogy az eredeti \mathbf{A} mátrixot két mátrix szorzatával adjuk meg. Végtelen sok ilyen felbontás létezik. Olyan megkötésekkel, mint ortogonalitás és normálás, a forgatástól eltekintve egyértelmű felbontást kapunk:

$$\mathbf{A}_{I \times J} = \left(\mathbf{U}_{I \times I} \mathbf{D}_{I \times J} \right)_{I \times J} \mathbf{V}_{J \times J}^T = \mathbf{X}_{I \times J} \mathbf{V}_{J \times J}^T = \mathbf{X}_{I \times N}^* \mathbf{V}_{N \times J}^{*T} + \mathbf{R}_{I \times J} \quad (3)$$

ahol az \mathbf{X} mátrixot a „score”, a \mathbf{V} mátrixot a „loading” jelzőkkel látják el a PCA szakirodalomban, mivel a \mathbf{V} -ben lévő oszlopvektorok által kifeszített ortogonális altérben az \mathbf{X} -ben lévő sorok mint koordináták jelenítik meg az eredeti \mathbf{A} mátrix sorvektorait egy-egy pontként (absztrakt leképezés): ezt az absztrakt teret sor-tér vagy V-tér vagy spektrum-tér elnevezésekkel illelhetjük. Az \mathbf{A} adatmátrix tranzponáltjára is felírhatjuk az SVD és az ebből származtatható PCA felbontásokat:

$$\mathbf{A}^T_{J \times I} = \left(\mathbf{V}_{J \times J} \mathbf{D}^T_{I \times J} \right)_{I \times J} \mathbf{U}_{I \times I}^T = \mathbf{Y}_{J \times I} \mathbf{U}_{I \times I}^T = \mathbf{Y}_{J \times N}^* \mathbf{U}_{N \times I}^{*T} + \mathbf{R}^T_{J \times I} \quad (4)$$

ahol az \mathbf{Y} mátrixot a „score”, az \mathbf{U} mátrixot a „loading” jelzőkkel látják el a PCA szakirodalomban az előzőeknek megfelelően, mivel az \mathbf{U} -ban lévő oszlopvektorok által kifeszített ortogonális altérben az \mathbf{Y} -ban lévő sorok mint koordináták jelenítik meg az \mathbf{A}^T mátrix sorvektorait (az eredeti \mathbf{A} mátrix oszlopvektorait) egy-egy pontként (absztrakt leképezés): ezt az absztrakt teret oszlop-tér vagy U -tér vagy minta-tér elnevezésekkel illelhetjük.

Minden adatredukciós eljárás egyik fontos lépése az adatok előkészítése (data preprocessing). Itt csak utalunk a centrálásra, a skálázásra, az egységnyi hosszra történő normálásra [16]. Ha olyan változócsoporthoz vannak, melyeket más és más skálán mértek, akkor blokkskálázást kell alkalmazni (amikor nem a teljes mátrixot, hanem csak egy részét, egy blokkot skálázunk). Főkomponens elemzés a skálázott adatmátrixból vagy a kovariancia és korrelációs mátrixokból kiindulva történhet. Kovariancia mátrixot akkor célszerű elemezni, ha az eredeti változók azonos nagyságrendbe esnek és skálájuk is közel azonos. Ha az egyes mátrix elemek mérési hibája elhanyagolható a változók mint oszlopok varianciájához képest, akkor a korrelációs mátrixot célszerűbb elemezni. Másképpen fogalmazva, korrelációs mátrixot elemzünk, ha az adatok nagyságrendje és/vagy skálája különböző, azaz szinte mindig. Fontos tudnunk még, hogy kovariancia-mátrix elemzése esetén forgatásinvariáns, de nem skálainvariáns eredményt kapunk, míg korrelációs mátrixok esetén a helyzet éppen fordított lesz.

Első lépésben meg kell vizsgálni az adatokat [16]: hiánytalanok-e? A hiányzó adatok nem feltétlenül gátolják a főkomponens elemzést. Általában a kiindulási mátrixban lévő 5% adathiány nem okoz gondot, nem változtatja meg az elemzés eredményét. Vannak módszerek, pl. a parciális legkisebb négyzetek módszere (PLS), mely 20%-ot is elvisel és a meglévő adatokból becsülni képes a hiányzókat. Nem hiányozhat azonban egy sorból vagy egy oszlopból az elemeknek több, mint a fele. Mesterséges ideghálók alkalmazásakor a hiányzó elemek száma akár az 50%-ot is elérheti, de ekkor is csak véletlenszerű elrendezésben. Természetesen a hiányzó adatok helyére nem tanácsos nullákat írni. A hiányzó helyekre a sorok vagy az oszlopok átlagát írhatjuk, vagy ha ezek valamiért nem tűnnek célravezetőnek, akkor

generálhatunk véletlen számot is az üres helyre, az oszlop vagy a sor tartományán (terjedelmén) belül. Ha az adatok strukturáltak, azaz az adatok egy része eltér az oszlop többi elemétől akkor részátlagokat alkalmazhatunk, pl. nappali és éjjeli adatok jelentősen eltérhetnek környezetvédelmi alkalmazásokban, értelemszerűen ilyenkor eltérő nappali és éjszakai átlagokat számolunk a hiányzó helyekre. Sose feledjük el azonban, hogy a hiányzó adatok pótlásakor torzítást viszünk a kiindulási mátrixba. Mindig ellenőrizni kell az elemzés eredményeit, nem vittünk-e be kiugró értéket a hiányzók helyettesítésével? Másrészt pedig ez a torzítás lehet igazolt, ha egy-két hiányzó adat miatt nem mondunk le egész oszlopokban (vagy) sorokban lévő információról. A hiányzó adatok becslésére dolgozták ki, pl. a sokszoros ráfogás (multiple imputation) módszerét [17]. Tulajdonságokat (oszlopokat) akkor távolíthatunk el a mátrixból, ha nagyon erősen korreláltak más oszlopokkal (vagyis ha redundánsak, pl. egyik oszlop kétszerese egy másiknak) vagy ha állandók.

Tehát a főkomponens elemzéssel az adatokat úgy adjuk vissza kevesebb dimenzióban, hogy új, nem korrelált változók keletkeznek [16]. Az adatok kevesebb dimenzióba történő leképezése alapvető fontosságú, hiszen $N > 3$ dimenziós adatkészlet az ember mintázatfelismerő képessége számára felfoghatatlan, míg 1, 2 vagy 3 dimenzióban az emberi agy mintázatfelismerő képessége nagyon jó, kapcsolatokat, hasonlóságokat könnyen észreveszünk így.

Általában nem szokás/érdemes az összes főkomponenst használni [16]. Például a tiszta alkotókat el kell választani a zaj komponensektől egy spektrumban vagy egy kromatogramban. Sajnos nincs arra egyértelmű szabály, hogy hány főkomponenssel célszerű az eredeti adatmátrixot visszaadni, reprezentálni. Sokat segíthet a tapasztalat, az intuíció, a hibaszintek és az adott probléma ismerete. Többféle heurisztikus és statisztikus kritérium létezik, hogy eldönthessük, mennyi főkomponensre van szükség:

- a megmagyarázott variancia bizonyos százaléka,
- a sajátérték egyenlő 1 kritérium,
- a hegyomlás ábra,

- a keresztellenőrzés.

A megmagyarázott variancia bizonyos százalékát mint heurisztikus kritériumot használjuk. Akkor lehet alkalmazni, ha elegendő tapasztalatot gyűjtöttünk hasonló adatok elemzésekor. A (kumulált) variancia adott részaránya számítható, mint a db fontos sajátérték (λ) összegének és az összes N sajátérték összegének hányadosa:

$$\frac{\sum_{i=1}^a \lambda_i}{\sum_{i=1}^N \lambda_i}$$

Ha az összes lehetséges főkomponenst használnánk, a teljes variancia 100%-át magyaráznánk meg [16]. Rendszerint csak annyi főkomponenst veszünk figyelembe, ahánnyal a variancia bizonyos százalékát, pl. 90 %-át tudjuk megmagyarázni.

A $\lambda_i = 1$ (a sajátérték egyenlő egy) kritérium azon alapszik, hogy a standardizált adatok átlagos sajátértéke éppen egységnyi. Ez azért van, mert a standardizált adatok esetén a korrelációs mátrix összes sajátértékének összege pontosan a lineárisan független változók számával, N -nel egyezik meg (ez éppen a korrelációs mátrix nyoma, $\text{Spur}(\mathbf{A})$, azaz az N db diagonálisban szereplő 1-esek összege). Csak azokat a komponenseket tekintjük fontosnak, melyek sajátértéke egynél nagyobb.

A hegyomlás (scree) próba alkalmazásakor a reziduumokat vagy még gyakrabban a sajátértékeket ábrázolják a főkomponensek sorszáma függvényében [16]. Így kapják meg a hegyomlás ábrát. (Nevét onnan kapta, hogy rendszerint a hegyomlás keresztmetszeti ábrájára hasonlít.) A reziduális variancia görbéjén a kellő számú főkomponens elérése esetén törés látható: ahol a függvény megtörik, ott leolvassuk a szükséges főkomponens számot.

A főkomponens szám kiválasztásának negyedik módszere a keresztellenőrzés [18]. A keresztellenőrzés alapötlete az, hogy az adatok egy részét nem használjuk fel a főkomponens elemzés során, hanem becsülni próbáljuk a főkomponens elemzés segítségével, majd összehasonlítjuk a becsült és a mért adatokat. Szoktak egyszerre teljes sorokat vagy oszlopokat is elhagyni. A becsült és mért adatok különbségének négyzetösszegét s^2 -nek nevezzük. Ezt az eljárást aztán többször ismételjük, míg minden adatelemet egyszer (és csak egyszer) hagyunk ki. Így az s^2 -ben minden adatelem hozzájárulása összegződik, s ezáltal a vizsgált modell prediktív erejére

lesz jellemző. A PCA során a keresztellenőrzést $a=0$ -val kezdjük és minden további dimenzióra kiszámítjuk az s^2 -et. Ábrázoljuk az egyetlen, a kettő, a három stb. főkomponens segítségével kapott, előrejelzett s értékeket a főkomponensek sorszámanak függvényében. A szignifikáns főkomponens számot a reziduális hiba minimuma adja. Nagy adatkészletek esetén az egy elem kihagyás (leave one out) módszerét felválthatja a több elem (oszlop) kihagyásának módszere (leave more out). A főkomponens elemzés eredményeinek láthatóvá tételét rendszerint a főkomponens-együttható (score) és főkomponens (loading) ábrákkal oldhatjuk meg [16]. Néha az adatok már egyetlen főkomponens segítségével magyarázhatók. A kereskedelemben kapható szoftverek két vagy háromdimenziós ábrázolásokat is lehetővé tesznek. A főkomponens-együttható ábrából a tulajdonságváltozók hasonlóságaira, korrelációira lehet következtetni. Itt természetesen nem az új változók, azaz nem a főkomponensek közötti, korrelációról van szó, hiszen azok ortogonálisak, hanem az eredeti tulajdonságváltozók és a főkomponensek közötti korrelációról. Ha a korrelációs mátrixot elemezzük, akkor a főkomponens-együtthatók közvetlenül a korrelációs koefficienseket adják meg. A tulajdonságváltozók korrelációját a főkomponens-együttható vektorok által bezárt szög koszinusza adja. Minél kisebb a szög, annál nagyobb a tulajdonságok közötti korreláció. A nem korrelált tulajdonságok egymással derékszöget zárnak be (ortogonálisak). Ha a változók nagyon korreláltak, akkor elegendő az egyiket használni a korrelált változók közül. A főkomponens-együttható abszolút értékben vett nagysága adja meg az adott tulajdonság fontosságát a kiindulási adatok főkomponensekkel való reprodukálása során. A koordináta rendszer origója közelében lévő főkomponens-együtthatók nem fontos tulajdonságokat reprezentálnak. Meg kell említenünk egy további bizonytalansági tényezőt. Nevezetesen, a főkomponens elemzés numerikusan azonos eredményt szolgáltat, ha a főkomponens-együtthatókat -1 -el beszorozzuk, csak előjelkonvenció dönti el milyen előjelet használunk. Az ellentétes előjeleknek azonban meg kell maradniuk. Nagyon fontos, hogy a főkomponens-együtthatók a -1 -el történő szorzásra invariánsak, hiszen kvalitatív vizsgálatkor csak az együtváltozás ténye az érdekes (azaz legyen korreláció) az iránya nem. Mintázatok felderítésénél

gyakran előfordul, különösen kezdő elemzőnél, hogy egy tulajdonságváltozó kiugrónak mutatkozik, holott az adott tulajdonságváltozóra vonatkozó pont az eredeti tulajdonságváltozó -1 -el történő szorzásával már a domináns csoportba kerül. A főkomponens-együtthetők centrálisan az origóra szimmetrikusak. Ezért a főkomponens-együtthető ábrákat gondosan elemezni kell, ha felmerül a gyanú, hogy egy kiugró pont az origóra történő tükrözés esetén nem marad kiugró érték, akkor az adott ponthoz tartozó tulajdonságváltozó -1 -el történő beszorzása után is el kell végezni az elemzést.

Hasonlóan a főkomponens-együtthető ábrákhoz, a főkomponens-ábrák is hordozhatnak jellegzetes mintázatot [16]. A skálájuk azonban eltér, míg a főkomponens-együtthető értéke -1 és $+1$ közé esik, a főkomponensek nem szoríthatók határok közé. A főkomponensek és főkomponens-együtthetők együttes értelmezése úgy is lehetséges, ha az együtthetőket és a főkomponenseket azonos mértékűre skálázzuk, és együttesen ábrázoljuk őket (biplot).

A PCA jelöléseket megtartva, megfelelő transzformáció után fizikailag és/vagy kémiailag értelmezhető megoldást kaphatunk az absztrakt helyett:

$$\mathbf{A}_{I \times J} = \begin{pmatrix} \mathbf{X}^* & \mathbf{T} \end{pmatrix}_{\substack{I \times N & N \times N}} \begin{pmatrix} \mathbf{T}^{-1} & \mathbf{V}^* \mathbf{T} \end{pmatrix}_{\substack{N \times N & N \times J}} + \mathbf{R}_{I \times J} = \mathbf{C}_{I \times N} \mathbf{S}^T_{N \times J} + \mathbf{R}_{I \times J} \quad (5)$$

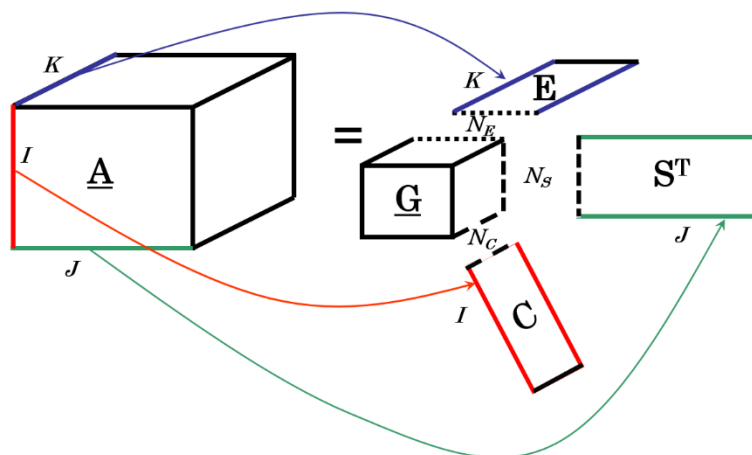
ahol \mathbf{T} a transzformációs mátrix (négyzetes mátrix, amelynek létezik az inverze (\mathbf{T}^{-1}), ami a reciprokképzés általánosítása mátrixokra), \mathbf{C} a koncentráció mátrix, \mathbf{S} az egységnyi koncentrációhoz tartozó tiszta komponensek spektrumait tartalmazó mátrix.

Mint az kikövetkeztethető, a \mathbf{T} transzformációs mátrix nem mindig választható meg egyértelműen: ezt a bizonytalanságot hívják forgatási (vagy transzformációs) bizonytalanságnak (rotational ambiguity, RA) [19]. Ebben benne van a skála invariancia is, azaz amikor a transzformációs mátrix diagonális mátrixra egyszerűsödik. A bizonytalanságok feloldásához kalibrációs mintákat kell alkalmaznunk.

A fentiekben vázolt feladat megoldására (\mathbf{C} és \mathbf{S} együttes becslése az \mathbf{A} adatmátrix dekompozíciójával, felbontásával) két markáns irányvonal alakult ki:

- az összes **C** és **S** megengedett megoldás feltérképezése (Self-modeling Curve Resolution – SMCR: görbeillesztés nélküli komponensprofil-kinyerés) [20],
- egyetlen **C** és **S** megengedett megoldás megtalálása (Multivariate Curve Resolution – MCR: többváltozós görbefelbontás) [21].

Mindkét irányvonal kényszerfeltételeket (constraints) használ a megengedett megoldás-halmaz (set of feasible solutions, SFS) vagy más kifejezéssel megengedett megoldás-tartomány (feasible region, FR) esetleges szűkítésére. A legkézenfekvőbb (természetes) kényszerfeltétel a nemnegativitás: a koncentráció sohasem lehet negatív és a legtöbb spektrum (jól beállított alapvonallal mérve!) szintén nem tartalmazhat negatív értéket, de pl. kivétel a cirkuláris dikroizmus (CD) spektroszkópia.



1. ábra Egy 3-dimenziós adattömb Tucker3 felbontásának illusztrálása

Összetettebb, több információt tartalmazó multilineáris/többutas (multilinear/multi-way) adatok nyerhetők a kapcsolt (kötőjeles) technikák révén [22], pl. HPLC-UV-Vis, HPLC-EEM, HPLC-FTIR, GC-MS(-nMS), LC-MS(-nMS) stb. Legalább két 'út' szükséges, az egyik lehet egy elválasztó 'út' (pl. kromatográfia), a másik pedig egy érzékelő 'út' (pl. UV spektroszkópia).

Ezen adatok hagyományos, többváltozós statisztikai módszerekkel történő feldolgozása, kiértékelése szinte lehetetlen, ill. alkalmazásuk során sok többlet információ elvész, vagy fel sem használjuk őket. A multilineáris/többutas módszerek előnye, hogy a kötőjeles technikákból nyert hatalmas mennyiségű összetett adattömböt képesek feldolgozni, egyértelmű dekompozíciós megoldást nyújtva viszonylag

enyhe feltételek mellett [23,24]. Számos multilineáris módszer említhető, mint például a Tucker3, PARAFAC, PARATUCK, PARALIND stb. Az általános Tucker3 dekompozíció az 1. ábrán látható.

ahol \mathbf{A} egy $I \times J \times K$ méretű adattömb (adatkocka), \mathbf{G} a Tucker mag-tömb (az egyes utak közötti kapcsolatrendszer megadó, minimális méretű tömb), \mathbf{C} a minták szerinti koncentráció (összetétel) mátrix, \mathbf{S} az egységnyi koncentrációhoz tartozó tiszta komponensek spektrumait tartalmazó mátrix, \mathbf{E} az egységnyi koncentrációhoz tartozó tiszta komponensek elúciós profiljait tartalmazó mátrix. A képletszerű összefüggést az adattömb k (ami 1-től K -ig fut) mátrixszeleteire tudjuk felírni:

$$\mathbf{A}_k = \mathbf{C} \left[\sum_{n_E=1}^{N_E} E_{k,n_E} \mathbf{G}_{n_E} \right] \mathbf{S}^T \quad (6)$$

$I \times J$ $I \times N_C$ $N_C \times N_S$ $N_S \times J$

Érdekesség, hogy a CANDECOMP/PARAFAC (CP) [21,22] dekompozíció speciális esete a Tucker felbontásnak, amikor is a \mathbf{G} Tucker mag-tömb szuperdiagonális és $N_C = N_S = N_E = N$. A CP probléma egyértelmű megoldásához a Kruskal-féle (enyhe) feltételek elegendőek [26].

4. Mérési adatok modellezése: elméleti fejlesztések és alkalmazások

4.1 A megengedett megoldás-tartományok analitikus meghatározása számítógépes geometriai módszerekkel [P1]

1970-ben Lawton és Sylvestre [20] vezette be a módszert és az elnevezést (SMCR) kétkomponensű rendszerek felbontására (erre figyelmemet Felinger Attila hívta fel az ACE-CC2003 konferencián). A módszer annyiban különbözik a paraméterillesztést felhasználó módszerektől, hogy itt nem kell a csúcsalak leírására egy konkrét függvényt felhasználni, hanem egyetlen elv (a Bogouert-Lambert-Beer törvény vagy annak analogonja) alkalmazásával ún. bilineáris adatmátrixot kapunk és ezt az SVD vagy a PCA segítségével ortogonális összetevőkre (ún. főkomponensekre) bonthatjuk. Háromkomponensű rendszerekre Borgen és munkatársai [28,29] általánosították Lawton és Sylvestre (LS) módszerét, sajnos azonban a cikkek alapján sokáig nem tudták rekonstruálni, ill. maradéktalanul értelmezni az algoritmust, ezért is kezdték el a közelítő eljárások fejlesztését (multivariate curve resolution – alternating least squares, MCR-ALS [21]). Mi ún. számítógépes geometriai módszereket [30,31] alkalmaztunk a lineáris programozás helyett [32,33], így egy igen gyors Matlab implementációt sikerült készítenünk.

Egy d -dimenziós politóp (pl. két dimenzióban poligon, három dimenzióban poliéder) nulla dimenziós oldala a csúcsa (vertex), egy dimenziós oldala az éle (edge), $(d-1)$ dimenziós oldala a lapja (facet) lesz, ahol az oldal (face) az általános határon lévő elem. Minden konvex politóp ábrázolható zárt félterek (closed half-space) véges halmazából származó metszetekként, ill. csúcsainak (vertex) konvex burkaként (convex hull). Az előbbit féltér-reprezentációnak (H-representation), az utóbbit csúcs-reprezentációnak (V-representation) nevezzük. A féltér-, ill. csúcs-reprezentációk egymásba történő átalakításait csúcs-felsorolásnak (vertex enumeration), ill. lap-felsorolásnak (facet enumeration) nevezzük. A d -dimenziós konvex

politóp összes csúcsának számbavétele és a ponthalmaz összes lapjának meghatározása alapvető probléma a matematikai programozásban és a számítógépes geometriában.

A Fourier-Motzkin eliminációs módszer (FMEM) [34] az egyik számítógépes geometriai eljárás ami a fentebbi feladatokat megoldja, azaz egy lineáris egyenlőtlenség-rendszerből a redundáns egyenlőtlenségeket kiszűri. A redundáns egyenlőtlenségek észlelése történhet lineáris programozással vagy a csúcsok és szélső irányok listájának kezelésével a dupla-deszkripció (double description (DD)) módszerrel [31].

Valós elemű mátrixok egy (\mathbf{F}, \mathbf{P}) párja ún. dupla-deszkripció pár (DD pair), ha a következő kapcsolat fennáll: $\mathbf{F} \mathbf{x} \geq \mathbf{0}$ akkor és csak akkor, ha

$$\mathbf{x} = \mathbf{P} \boldsymbol{\lambda} \text{ valamilyen } \boldsymbol{\lambda} \geq \mathbf{0} \text{ esetén.}$$

Egy DD pár mindkét (lap-, ill. csúcs-) felsorolást, azaz deskripciót tartalmazza. Nevezetesen, a $\mathcal{P} \in \mathbb{R}^n$ halmaz egyrészt megadható az \mathbf{F} mátrix által, azaz féltér-reprezentációként:

$$\mathcal{P} = \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{F} \mathbf{x} \geq \mathbf{0} \right\} \quad (7)$$

másrészt a \mathbf{P} mátrix által, azaz csúcs-reprezentációként:

$$\mathcal{P} = \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{P} \boldsymbol{\lambda} \text{ és } \boldsymbol{\lambda} \geq \mathbf{0} \right\} \quad (8)$$

Egy C programozási nyelvű DD módszer implementáció, amely egy az origót is csúcsként tartalmazó konvex politóp (pointed polyhedron) összes csúcsát és alkotóját, vagyis szélső sugarait (extremal ray) meghatározza, elérhető az Interneten keresztül [35].

Az általunk Borgen grafikonnak (Borgen plot) elnevezett ábra elkészítéséhez először az adatmátrixunk normalizálását végeztük el az egyes sorok, ill. oszlopok összegeit felhasználva (azaz a normalizálás után kapott felülvonással jelölt új mátrix sorai, ill. oszlopai L_1 normája 1 lesz):

$$\bar{\mathbf{A}}_S = \mathbf{A} ./ \left(\mathbf{A} \mathbf{1} \right) \quad (9)$$

$$\bar{\mathbf{A}}_C = \mathbf{A}_{I \times J} ./ \left(\mathbf{1}_{I \times I} \mathbf{A}_{I \times J} \right) \quad (10)$$

Mivel Matlab [36] környezetben történt a fejlesztés, ezért a ./ műveleti jel a mátrixok elemenkénti osztását jelenti. A továbbiakban $N=3$ komponensre mutatom be a Borgen grafikon elkészítésének lényegi lépéseit. A sorok szerint normalizált adatmátrix PCA felbontása:

$$\bar{\mathbf{A}}_S = \mathbf{X}^*_{I \times J} \mathbf{V}^{*T}_{3 \times J} \geq \mathbf{0}_{I \times J} \quad (11)$$

Az \mathbf{X}^* mátrix sorai által meghatározott minden pont a \mathbf{V}^* mátrix oszlopai által kifesztett absztrakt térben egy síkon fog elhelyezkedni az alkalmazott normalizálás következtében:

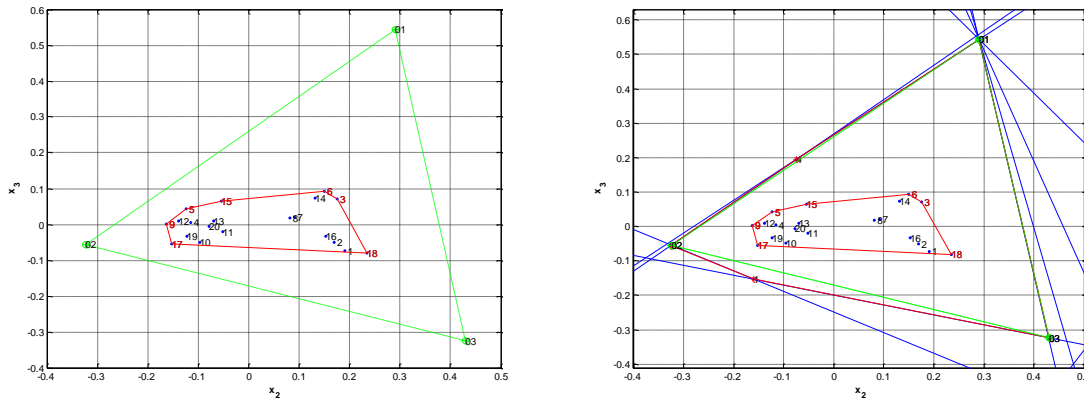
$$\bar{\mathbf{A}}_S \mathbf{1}_{I \times J \times 1} = \mathbf{X}^*_{I \times 3} \mathbf{V}^{*T}_{3 \times J \times 1} \mathbf{1}_{I \times 1} = \mathbf{1}_{I \times 1} \Rightarrow \mathbf{x}^{*T}_{1 \times 3} \mathbf{V}^{*T}_{3 \times J \times 1} \mathbf{1}_{I \times 1} = 1 \Rightarrow \left(\mathbf{1}_{1 \times J \times 3} \mathbf{V}^* \right)_{3 \times 1} \mathbf{x}^*_{3 \times 1} = 1 \Rightarrow \mathbf{a}_{1 \times 3} \mathbf{x}^*_{3 \times 1} = b \quad (12)$$

A normalizálás okozta dimenzió-csökkenés miatt pl. a \mathbf{v}_1^* jobboldali szinguláris vektor által meghatározott irányt elhagyhatjuk, így csak az \mathbf{x}_2^* és \mathbf{x}_3^* vektorok elemeinek megfelelő koordináták számítanak. Bevethetjük az első számítógépes geometriai eszközt, vagyis az absztrakt 2-dimenziós síkon lévő I db adatpont konvex burkának meghatározását, amit a *convhull* Matlab [36] függvénnyel végeztünk. Megkaptuk a belső poligont (inner polygon), amit egy megengedett megoldás szimplexe, vagyis három komponens esetén háromszög egyetlen oldala sem metszhet, legfeljebb érintheti. A 2. ábrán a szaggatott zöld háromszög a példaadat előállításához használt ismert három spektrum absztrakt leképezése, amely egy tagja a megengedett megoldás-halmaznak.

A 11. egyenletben is nyomatékosított nemnegativitás miatt J db féltérrel, azaz 2-dimenzióban félsíkot is származtathatunk:

$$\mathbf{X}^*_{I \times 3} \mathbf{V}^{*T}_{3 \times J} \geq \mathbf{0}_{I \times J} \Rightarrow \mathbf{x}^{*T}_{1 \times 3} \mathbf{V}^{*T}_{3 \times J} \geq \mathbf{0}_{1 \times J} \Rightarrow \mathbf{x}^{*T}_{1 \times 3} \mathbf{v}_j^*_{3 \times 1} \geq 0 \Rightarrow \mathbf{v}_j^{*T}_{1 \times 3} \mathbf{x}^*_{3 \times 1} \geq 0 \quad (13)$$

A fenti egyenlet utolsó egyenlőtlenségében a szerepek felcserélődnek, azaz \mathbf{x}^* elemei a változók és \mathbf{v}_j^* elemei azok együtthatói.



2. ábra A baloldalon a belső poligon, a jobb oldalon a külső poligon ábrázolása

A feladat most az, hogy a redundáns félsíkokat kizárjuk a másik számítógépes geometriai eszköz, azaz a DD segítségével, amit a Matlabban a *cddmex* [37] révén valósítottunk meg. A nem redundáns félsíkok a külső poligont határozzák meg (2. ábra jobb oldal). A kék egyenesek a J db félsíkot jelképezik, piros átfedéssel láthatjuk a nem redundáns elemeket. A megengedett megoldás-tartományok a belső és a külső poligonok között keresendők, tehát nemnegatív profilok absztrakt pontjai a külső poligonon belül, amelyek egy háromszög csúcsaként a belső poligont körbeveszik, azaz lineáris konvex kompozícióként az adatpontok előállíthatók.

A Borgen által bevezetett forgatott érintők algoritmusával megszerkeszthető a határoló görbe, aminek a külső poligonnal való metszete határozza meg a komponensek megengedett megoldás-tartományait (3. ábra). A P_r kezdőpont legyen az L_r külső poligon oldalán. Az L_l érintő a P_r ponttól a belső poligon P_l pontját érintve metszi a külső poligon L_s oldalát a P_s pontban. Az új L_m érintő a P_s pontból indul érintve a belső poligon P_m pontját. Egy másik L_k érintőt indítunk a P_r pontból a belső poligon ellentétes részéhez. Az L_k érintő a belső poligont a P_k pontban érinti és a P_l pontban metszi az L_m érintőt. A P_r pontot szisztematikusan csúsztatva a külső poligonon, a határoló görbe összes pontját meghatározhatjuk.

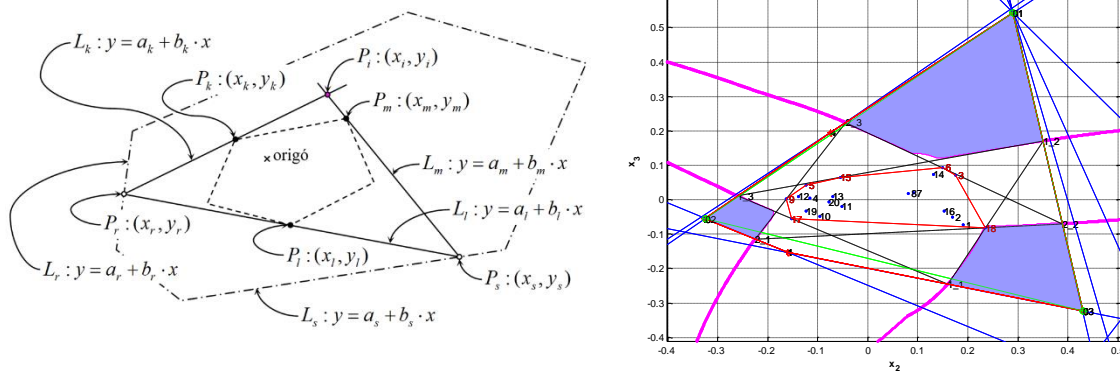
Borgen a forgatott érintők algoritmus egy változatát is bevezette a forgatott szimplex algoritmust, ami annyiban különbözik az előzőtől, hogy egyszerre csak egy oldalon forgatjuk az érintőket kvázi szimplexek sorozatát létrehozva, határesetben pedig kialakul az általunk Borgen háromszögnek nevezett alakzat. A Borgen há-

romszögek metszetei a megengedett megoldás-tartományok konvex határait jelölik ki. A 3. ábra jobb oldalán láthatjuk a két Borgen háromszöget szaggatott vonallal jelölve.

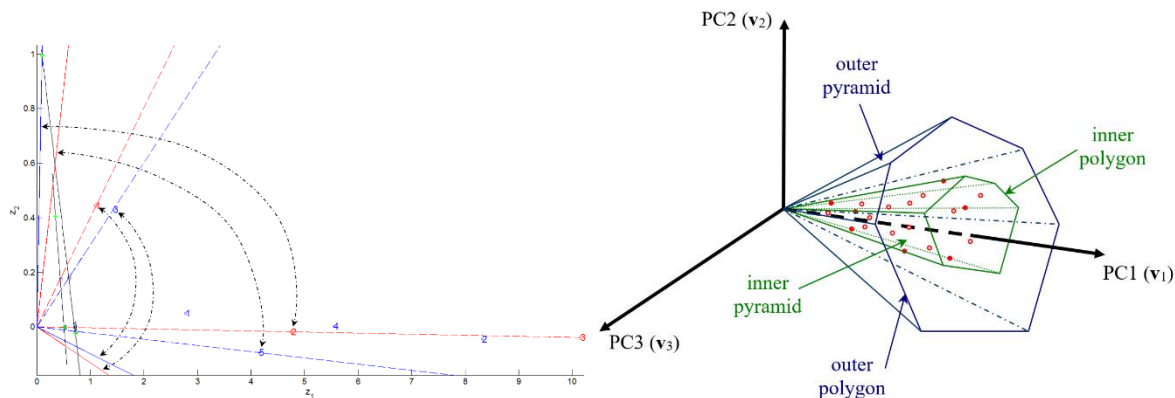
Ebben a tanulmányban vezettük be az Average Orthogonal Distance from the Linear Subspace (AODLS) elnevezésű mértéket, ami egy ponthalmaz átlagos ortogonális távolságát adja meg egy lineáris altértől:

$$\text{AODSL} = \frac{\left| \frac{\mathbf{X}^* \mathbf{V}^{*T} \mathbf{1} - \mathbf{1}}{\|\mathbf{V}^{*T} \mathbf{1}\|_2} \right|^T}{I} \quad (14)$$

Az AODSL segítségével kimutattuk, hogy a Lawton és Sylvestre által használt adathalmaz akár négy komponenst is tartalmazhat kettő helyett.



3. ábra A bal oldalon Borgen forgatott érintők algoritmusának illusztrálása, a jobb oldalon a magenta határoló görbe és a kék megengedett megoldás-tartományok



4. ábra Bal oldalon két komponens esetén a természetes dualitás, jobb oldalon három komponens esetén a belső és külső gúlának (pyramids) szemléltetése

Henry adatait ([38] 1. táblázata) felhasználó példán keresztül illusztráltam az elméletet (lásd 1. táblázatot is):

$$\mathbf{A}_{3 \times 5} = \begin{pmatrix} 0,3200 & 0,7680 & 0,5680 & 0,5440 & 0,3120 \\ 1,1700 & 3,5100 & 0,5850 & 2,3400 & 1,7550 \\ 2,5000 & 7,5000 & 1,2500 & 5,0000 & 3,7500 \end{pmatrix} =$$

$$\mathbf{X}_{3 \times 2} \mathbf{V}^T_{2 \times 5} = \begin{pmatrix} 1,1007 & 0,4434 \\ 4,7525 & -0,0185 \\ 10,1550 & -0,0394 \end{pmatrix} \begin{pmatrix} 0,2466 & 0,7382 & 0,1268 & 0,4924 & 0,3685 \\ 0,1095 & -0,1003 & 0,9661 & 0,0046 & -0,2110 \end{pmatrix} \quad (17)$$

$$\mathbf{A}^T_{5 \times 3} = \mathbf{Y}_{5 \times 2} \mathbf{U}^T_{2 \times 3} = \begin{pmatrix} 2,7783 & 0,0488 \\ 8,3161 & -0,0447 \\ 1,4290 & 0,4304 \\ 5,5472 & 0,0020 \\ 4,1510 & -0,0940 \end{pmatrix} \begin{pmatrix} 0,0977 & 0,4218 & 0,9014 \\ 0,9952 & -0,0414 & -0,0885 \end{pmatrix} \quad (18)$$

$$\mathbf{D}_{2 \times 2} = \begin{pmatrix} 11,2659 & 0,0000 \\ 0,0000 & 0,4455 \end{pmatrix} \quad (19)$$

1. táblázat Kölcsonösen egyértelmű ráképezések az egyenesek és pontok között

| j | $y_{j,1}$ | $y_{j,2}$ | $y_{j,1}/d_{1,1}$ | $y_{j,2}/d_{2,2}$ | $v_{j,1}$ | $v_{j,2}$ | $-\frac{y_{j,1} \cdot d_{2,2}}{y_{j,2} \cdot d_{1,1}}$ | $-v_{j,1}/v_{j,2}$ |
|-----|-----------|-----------|-------------------|-------------------|-----------|-----------|--|--------------------|
| 3 | 1,4290 | 0,4304 | 0,1268 | 0,9661 | 0,1268 | 0,9661 | -0,1313 | -0,1313 |
| 5 | 4,1510 | -0,0940 | 0,36846 | -0,2110 | 0,3685 | -0,2110 | 1,7463 | 1,7463 |
| i | $x_{i,1}$ | $x_{i,2}$ | $x_{i,1}/d_{1,1}$ | $x_{i,2}/d_{2,2}$ | $u_{i,1}$ | $u_{i,2}$ | $-\frac{x_{i,1} \cdot d_{2,2}}{x_{i,2} \cdot d_{1,1}}$ | $-u_{i,1}/u_{i,2}$ |
| 1 | 1,1007 | 0,4434 | 0,0977 | 0,9952 | 0,0977 | 0,9952 | -0,0982 | -0,0982 |
| 2 | 4,7525 | -0,0184 | 0,4218 | -0,0414 | 0,4218 | -0,0414 | 10,1860 | 10,1860 |

4.3 A megengedett megoldás-tartományok határainak számítása és a forgatási bizonytalanság mértékének meghatározása [P3]

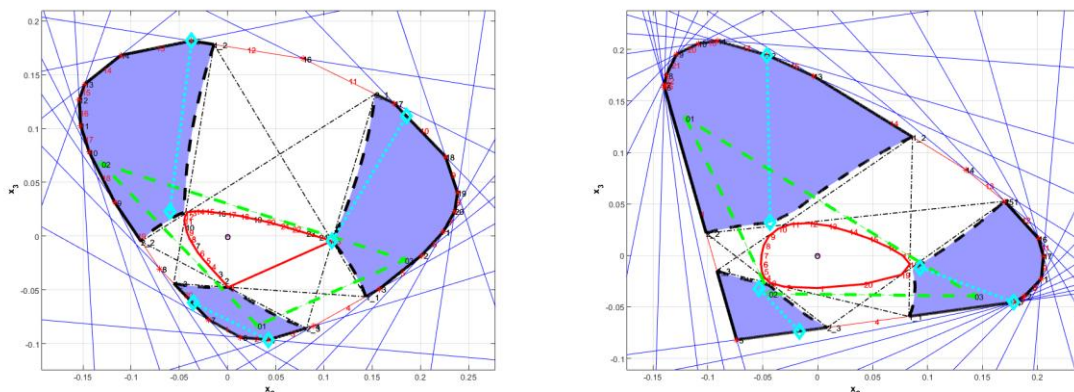
Ebben a tanulmányban a megengedett megoldás-tartományok határainak számításával kapcsolatos eredményeimet mutattam be. Gemperline [39] vezette be először a megengedett megoldás-tartományok terjedelmének (range of feasible solutions) meghatározását az egyes komponensek profiljainak görbe alatti területe (azaz az L_1 normájuk) maximalizálásával, ill. minimalizálásával:

$$\max\{\sum_{i=1}^I \sum_{j=1}^J (C_{n,i} S_{n,j})\} \text{ vagy } \min\{\sum_{i=1}^I \sum_{j=1}^J (C_{n,i} S_{n,j})\} \quad (20)$$

természetesen a nemnegativitás kényszerfeltétel és a profilok megfelelő kezelése mellett. Később Tauler [40] egy másik vektornormát, az L_2 euklideszi normát alkalmazta és bevezette a jel hozzájárulás függvényt (signal contribution function, SCF):

$$f_n(\mathbf{T}) = \frac{\|\mathbf{c}_n \mathbf{s}_n\|}{\left\| \begin{matrix} \mathbf{c} & \mathbf{s}^T \\ I \times N & N \times J \end{matrix} \right\|} \quad (21)$$

A \mathbf{T} rotációs (transzformációs) mátrix az n . komponenshez tartozó maximum és minimum normájú profilok meghatározásához kellett.

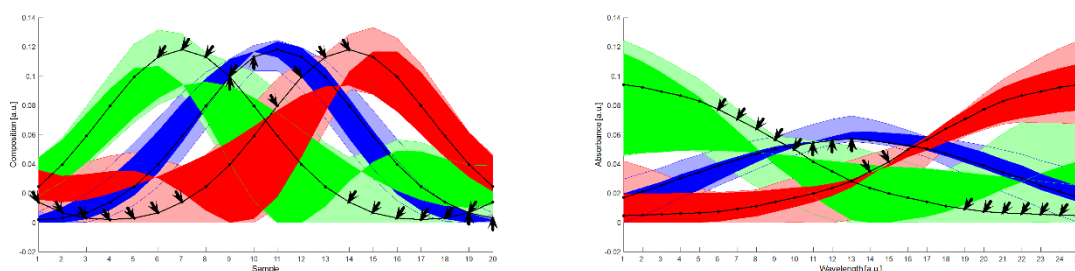


5. ábra Borgen grafikon illusztráció ami megmutatja, hogy két szélsőérték általában nem elegendő a teljes megengedett megoldás-tartomány leírására. Bal oldalon a koncentráció tér, a jobb oldalon a spektrális tér

Az elképzelést, hogy két szélsőértékkel elég lesz jellemezni a forgatási bizonytalanságot, nyilvánvalóan Lawton és Sylvestre kétkomponensű megoldása sugallta. Sikerkült azonban kimutatnom, hogy két szélsőérték általában nem elegendő. Ehhez az általam fejlesztetett és korábban bemutatott Borgen-plot rajzoló Matlab szkriptemet használtam fel. Az 5. ábrán a késsel jelölt megengedett megoldás-tartományok nem fedhetők le maradéktalanul a Gemperline-Tauler módszer által meghatározott min-max szélsőértékek definiálta türkiz szakaszokkal.

A 6. ábrán az 5. ábra koncentráció és spektrális terében kapott Borgen grafikonok alapján a megengedett megoldás-tartományok absztrakt pontjainak visszatranszformációjával kapott profilokat láthatjuk. A nyilak jelzik, hogy a Gemperline-Tauler módszerrel kapott (sötétebb színű) sávok hol nem tartalmazzák az adatgeneráláshoz használt profilok pontjait. A helyesen meghatározott világosabb színű sávok viszont tartalmazzák azokat és természetesen a Gemperline-Tauler módszerrel kapott hibás sávokat is.

A rotációs bizonytalanság számszerűsítésére bevezettem a komponensek koncentráció és spektrum szerinti Borgen grafikonbéli megengedett megoldás-tartományainak területe és a külső poligon területe között számolt arányt, valamint a Gemperline-Tauler módszerrel kapott maximum és minimum SCF értékek különbségét, melyeket az 2. táblázatban szerepeltetek.



6. ábra A fenti Borgen grafikonok megengedett megoldás-tartományainak pontjaiból transzformált profilok megjelenítése. Bal oldalon a koncentráció profilok, jobb oldalon a spektrumok

Az 2. táblázatból rögtön szembe tűnhet, hogy a Gemperline-Tauler módszer komponensenként, a koncentráció és spektrum hatás megkülönböztetése nélkül nyújt információt (utolsó sor), míg a terület-arány külön a koncentrációra és külön a spektrumokra is meghatározható (1. és 2. sorok).

A probléma megvilágításához képzeljük el, hogy az egyik komponens spektruma ismert. Így az ehhez tartozó rotációs bizonytalanság nulla lesz. Azonban a koncentrációhoz tartozó rotációs bizonytalanság nem feltétlenül lesz nulla. A Gemperline-Tauler módszerrel kapott maximum és minimum SCF értékek különbsége viszont ebben az esetben hibás eredményt ad: ha az érték nulla, akkor nem kezelte megfelelően a koncentrációban maradt rotációs bizonytalanságot; ha értéke nem nulla, akkor nem megfelelően kezelte az ismert spektrumot.

2. táblázat A rotációs bizonytalanság mérése különböző módszerekkel

| | Komp. 1 | Komp. 2 | Komp. 3 |
|--------------|---------|---------|---------|
| Koncentráció | 0,2912 | 0,5158 | 0,4722 |
| Spektrum | 0,2675 | 0,6309 | 0,3654 |
| Max-Min | 0,4710 | 0,4830 | 0,3000 |

4.4 Különböző Borgen-normák alkalmazhatóságának vizsgálata [P4]

Ebben a közleményben a Borgen-normák definícióját és részletes leírását adtam meg, amely téma először jelent meg a kémiai szakirodalomban. Néhány elméleti és gyakorlati vizsgálódás is szerepel az SMCR-módszerhez használt Borgen-normák adaptálhatóságáról.

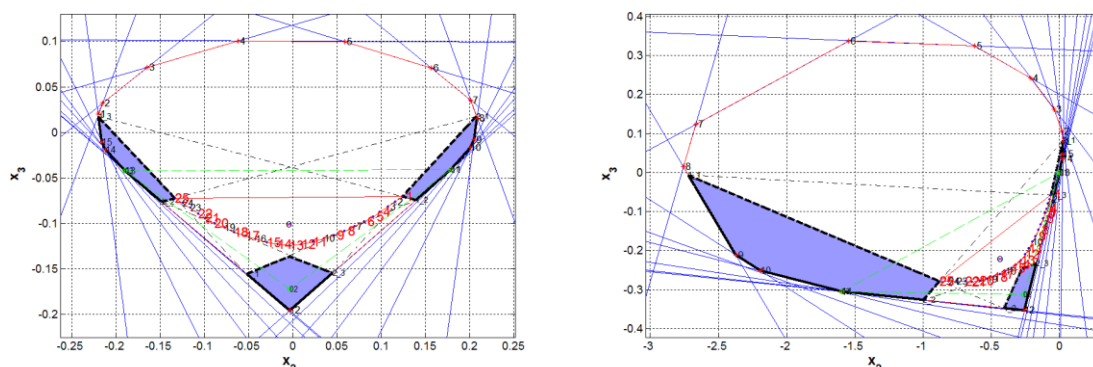
A Borgen-normák általános alakja:

$$\|\mathbf{x}\|_{B,z} = \mathbf{z}^T |\mathbf{x}| \quad \text{ha } \mathbf{z} > \mathbf{0} \quad (22)$$

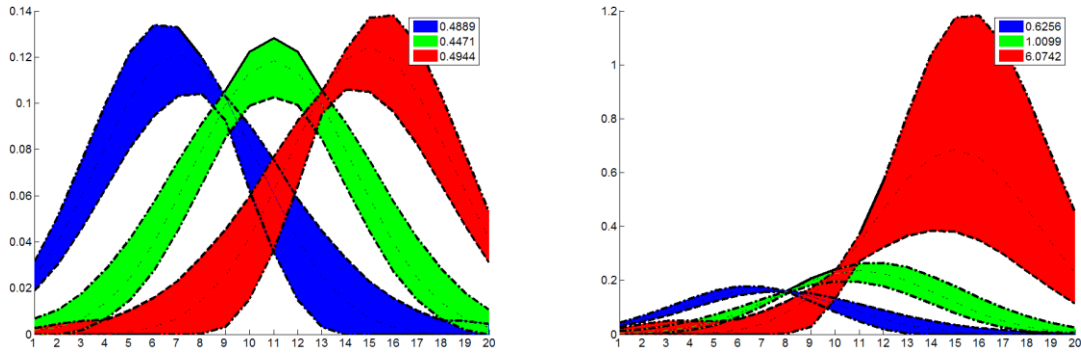
Tehát egy \mathbf{x} vektor Borgen normájának meghatározása úgy történik, hogy az \mathbf{x} vektor elemeinek abszolútértékeit vesszük, majd egy csak pozitív elemű vektorral skáláris szorzatot képzünk, azaz az azonos helyen lévő elemek szorzatainak összegét számoljuk.

Az így meghatározott normák teljesítik a normákra kötelező három feltételt:

1. pozitív homogenitás: $\|\alpha \mathbf{x}\|_{B,z} = \mathbf{z}^T |\alpha \mathbf{x}| = |\alpha| \mathbf{z}^T |\mathbf{x}| = |\alpha| \|\mathbf{x}\|_{B,z}$
2. háromszög-egyenlőtlenség: $\|\mathbf{x} + \mathbf{y}\|_{B,z} = \mathbf{z}^T |\mathbf{x} + \mathbf{y}| \leq \mathbf{z}^T (|\mathbf{x}| + |\mathbf{y}|) = \mathbf{z}^T |\mathbf{x}| + \mathbf{z}^T |\mathbf{y}| = \|\mathbf{x}\|_{B,z} + \|\mathbf{y}\|_{B,z}$
3. Nemnegativitás, illetve a nulla normaérték csak a nullvektorhoz rendelhetősége: mivel $\mathbf{z} > \mathbf{0}$, $\|\mathbf{x}\|_{B,z} = \mathbf{z}^T |\mathbf{x}| \geq 0$, illetve $\|\mathbf{0}\|_{B,z} = \mathbf{z}^T |\mathbf{0}| = 0$, és ha $\mathbf{x} \neq \mathbf{0}$, akkor $\|\mathbf{x}\|_{B,z} = \mathbf{z}^T |\mathbf{x}| \neq 0$



7. ábra Koncentráció tér Borgen grafikonjai, bal oldalon Borgen norma: L1; jobb oldalon Borgen norma: első komponens profilja alapján



8. ábra Koncentráció tér Borgen grafikonjaiból visszatranszformált profilsávok, bal oldalon Borgen norma: L_1 ; jobb oldalon Borgen norma: első komponens profilja alapján

Megállapíthatjuk, hogy az L_p -vel jelölt Hölder normák közül a $p=1$ eset, azaz az L_1 norma lesz Borgen norma: $\|\mathbf{x}\|_1 = \mathbf{1}^T |\mathbf{x}| = \|\mathbf{x}\|_{B,1}$, azaz $\mathbf{z} = \mathbf{1}$.

A közleményben megadtam a Brogen normák esetére az általános levezetéseket az SMCR-rel kapcsolatban [P4]. Először tekintsük az \mathbf{A} mátrix soronkénti normalizálását, azaz $\|\bar{\mathbf{a}}_{(i)}^T\|_{B,z} = \mathbf{z}^T \bar{\mathbf{a}}_{(i)}^T = 1$:

$$\bar{\mathbf{a}}_{(i)}^T = \mathbf{a}_{(i)}^T / (\mathbf{z}^T \mathbf{a}_{(i)}^T) = \sum_{n=1}^N c_{i,n} \mathbf{s}_n / \sum_{n=1}^N c_{i,n} \mathbf{z}^T \mathbf{s}_n =$$

$$\underbrace{\sum_{n=1}^N c_{i,n} \mathbf{z}^T \mathbf{s}_n}_{c_{i,n}^*} \frac{\mathbf{s}_n}{\mathbf{z}^T \mathbf{s}_n} / \underbrace{\sum_{n=1}^N c_{i,n} \mathbf{z}^T \mathbf{s}_n}_{c_{i,n}^*} \quad (23)$$

$$\bar{\mathbf{A}}_S = \begin{bmatrix} \bar{\mathbf{a}}_{(1)} \\ \vdots \\ \bar{\mathbf{a}}_{(i)} \\ \vdots \\ \bar{\mathbf{a}}_{(I)} \end{bmatrix}_{I \times J} = \underbrace{\begin{pmatrix} \begin{bmatrix} \mathbf{c}_1^{*T} \\ \vdots \\ \mathbf{c}_i^{*T} \\ \vdots \\ \mathbf{c}_I^{*T} \end{bmatrix}_{1 \times N} \\ \mathbf{1}_N \end{pmatrix}}_{\bar{\mathbf{W}}_C^*} \cdot \begin{pmatrix} \begin{bmatrix} \mathbf{c}_1^{*T} \\ \vdots \\ \mathbf{c}_i^{*T} \\ \vdots \\ \mathbf{c}_I^{*T} \end{bmatrix}_{1 \times N} \\ \mathbf{1}_N^T \end{pmatrix} \quad \bar{\mathbf{S}}^T = \bar{\mathbf{W}}_C^* \bar{\mathbf{S}}^T \quad (24)$$

$$\bar{\mathbf{S}} = \mathbf{S} \cdot \begin{pmatrix} \mathbf{1}_J \\ \mathbf{z}^T \mathbf{S} \end{pmatrix}_{J \times 1} \quad (25)$$

$$\mathbf{C}^* = \mathbf{C} \cdot \begin{pmatrix} \mathbf{1}_I \\ \mathbf{z}^T \mathbf{S} \end{pmatrix}_{I \times 1} \quad (26)$$

$$\overline{\mathbf{W}}_C^* = \mathbf{C}^* ./ \left(\mathbf{C}^* \mathbf{1}_N \mathbf{1}_N^T \right) \quad (27)$$

ahol ./ and .* az elemenkénti mátrixosztás és szorzás a Matlab koncepciójának megfelelően. Figyeljük meg, hogy a $\overline{\mathbf{W}}_C^*$ mátrix közvetlen kapcsolatban van a második úttal (spektrális profilok, \mathbf{S}) a 26. egyenleten keresztül.

Másodsorban tekintsük az \mathbf{A} mátrix oszloponkénti normalizálását, azaz $\|\bar{\mathbf{a}}_j\|_{B,z} = \mathbf{z}^T \bar{\mathbf{a}}_j = 1$:

$$\begin{aligned} \bar{\mathbf{a}}_j &= \mathbf{a}_j / (\mathbf{z}^T \mathbf{r}_j) = \sum_{n=1}^N \mathbf{c}_n s_{j,n} / \sum_{n=1}^N \mathbf{z}^T \mathbf{c}_n s_{j,n} = \\ &= \frac{\sum_{n=1}^N \frac{\mathbf{c}_n}{\mathbf{z}^T \mathbf{c}_n} \mathbf{z}^T \mathbf{c}_n s_{j,n}}{\sum_{n=1}^N \mathbf{z}^T \mathbf{c}_n s_{j,n}} \end{aligned} \quad (28)$$

$$\begin{aligned} \bar{\mathbf{A}}_C &= \begin{bmatrix} \bar{\mathbf{a}}_1 & \cdots & \bar{\mathbf{a}}_j & \cdots & \bar{\mathbf{a}}_J \end{bmatrix} = \\ &= \underbrace{\bar{\mathbf{C}} \left(\begin{bmatrix} \mathbf{s}_1^* & \cdots & \mathbf{s}_j^* & \cdots & \mathbf{s}_J^* \end{bmatrix} ./ \left(\mathbf{1}_N \mathbf{1}_N^T \begin{bmatrix} \mathbf{s}_1^* & \cdots & \mathbf{s}_j^* & \cdots & \mathbf{s}_J^* \end{bmatrix} \right) \right)}_{\overline{\mathbf{W}}_S^{*T}} = \\ &= \bar{\mathbf{C}} \overline{\mathbf{W}}_S^{*T} \end{aligned} \quad (29)$$

$$\bar{\mathbf{C}} = \mathbf{C} ./ \left(\mathbf{1}_I \mathbf{z}^T \mathbf{C} \right) \quad (30)$$

$$\mathbf{S}^* = \mathbf{S} .* \left(\mathbf{1}_J \mathbf{z}^T \mathbf{C} \right) \quad (31)$$

$$\overline{\mathbf{W}}_S^* = \mathbf{S}^* ./ \left(\mathbf{S}^* \mathbf{1}_N \mathbf{1}_N^T \right) \quad (32)$$

Ismét figyeljünk fel arra, hogy a $\overline{\mathbf{W}}_S^*$ közvetlen kapcsolatban lesz az első úttal (koncentráció profilok, \mathbf{C}) a 31. egyenleten keresztül.

Most már tisztázott, hogy a normalizált válaszjelek a normalizált tiszta komponens profilok konvex kombinációi lesznek (24. és 29. egyenletek), mivel a $\overline{\mathbf{W}}_C^*$ és $\overline{\mathbf{W}}_S^*$

mátrixok elemei mint kombinációs együtthatók (súlyok) nemnegatívak és a soronkénti összegek eggyel egyenlők a 27. és 32. egyenleteknek megfelelően (felhasználva $\mathbf{1}_N^T \mathbf{e}_1 = 1$ összefüggést, ahol \mathbf{e}_1 az egységmátrix első sora vagy oszlopa, azaz az első egységvektor):

$$\begin{aligned} \bar{\mathbf{W}}_C^* \mathbf{1}_N &= \left(\mathbf{c}^* ./ \left(\mathbf{c}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) \right) \mathbf{1}_N = \left(\mathbf{c}^* ./ \left(\mathbf{c}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) \right) \mathbf{1}_N \mathbf{1}_N^T \mathbf{e}_1 = \\ & \left(\left(\mathbf{c}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) ./ \left(\mathbf{c}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) \right) \mathbf{e}_1 = \mathbf{1}_I \mathbf{1}_N^T \mathbf{e}_1 = \mathbf{1}_I \end{aligned} \quad (33)$$

$$\begin{aligned} \bar{\mathbf{W}}_S^* \mathbf{1}_N &= \left(\mathbf{s}^* ./ \left(\mathbf{s}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) \right) \mathbf{1}_N = \left(\mathbf{s}^* ./ \left(\mathbf{s}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) \right) \mathbf{1}_N \mathbf{1}_N^T \mathbf{e}_1 = \\ & \left(\left(\mathbf{s}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) ./ \left(\mathbf{s}^* \quad \mathbf{1}_N \quad \mathbf{1}_N^T \right) \right) \mathbf{e}_1 = \mathbf{1}_J \mathbf{1}_N^T \mathbf{e}_1 = \mathbf{1}_J \end{aligned} \quad (34)$$

A fenti összefüggések igazolják, hogy a normalizált válaszjel-vektorokat az $N-1$ dimenziós szimplex fogja körbe, amely csúcspontjait az N db normalizált tiszta profilvektorok alkotják. A következő összefoglaló képleteket kapjuk az \mathbf{A} mátrix egyes sorai, ill. oszlopai Borgen norma szerinti normalizálására:

$$\bar{\mathbf{A}}_S = \mathbf{A} ./ \left(\mathbf{A} \quad \mathbf{z} \quad \mathbf{1}^T \right) \quad (35)$$

$$\bar{\mathbf{A}}_C = \mathbf{A} ./ \left(\mathbf{1} \quad \mathbf{z}^T \quad \mathbf{A} \right) \quad (36)$$

Mindenképpen megállapíthatjuk, hogy a Borgen-normák használata rendkívül gyümölcsöző lehet az SMCR alkalmazásokban. Az egyik legfontosabb és legmeglepőbb következtetésem, hogy a rotációs bizonytalanság mértéke függ az alkalmazott Borgen normától (7. és 8. ábrák) a minimális kényszerfeltételű (nemnegatív összetétel- és jel-profilok) SMCR esetén.

A Borgen grafikonok segítségével visszatranszformált profilsávokat utólag normalizálhatjuk, vagy változatlanul hagyhatjuk. A nem normalizált sávterületek alkalmasak a három komponens rotációs bizonytalanságának összehasonlítására egy

kiválasztott Borgen normára vonatkozóan; míg a normalizált sávterületek a különböző Borgen-normákból adódó rotációs bizonytalanság összehasonlítására használhatók egy kiválasztott komponensre.

Ezenkívül a különböző Borgen-normákkal elvégzett SMCR vizuális validációs eljárásként használható a valódi profilok unimodalitásának eldöntésére a profilsávok alapján, előzetes fizikai vagy kémiai ismeretek vagy az adatokra vonatkozó feltételezések nélkül.

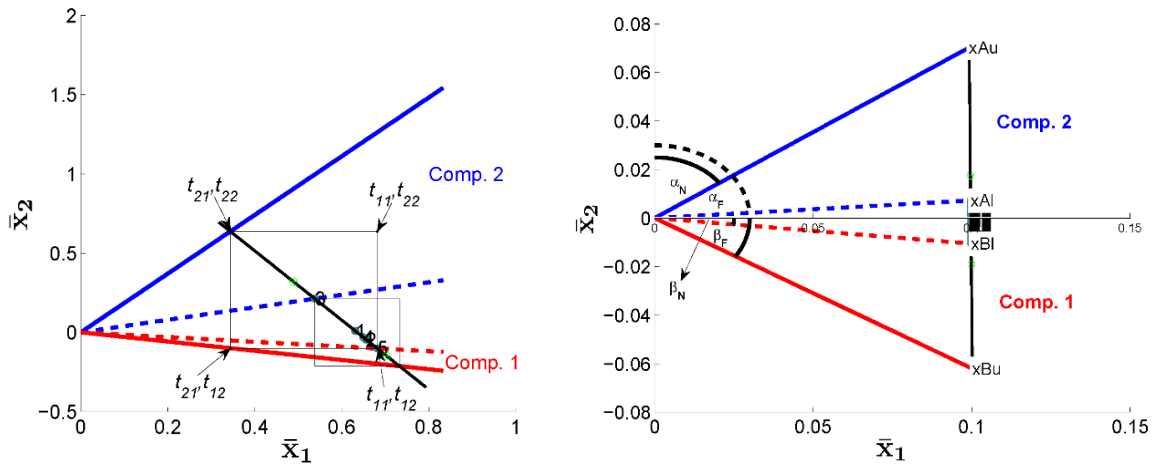
4.5 Elméleti fejtegetések kétkomponensű rendszerek megengedett megoldás-tartomány határainak értelmezéséhez és számításához [P5]

Korábban Abdollahi és mtsai [41] egy kétkomponensű rendszer SMCR-éhez a transzformációs mátrix elemeinek kiszámításához és sávhatárok meghatározásához a Tauler-féle *mcrbands* Matlab szkriptet [42] és a Maeder-féle rács módszert [43] használták. Egyik módszer sem analitikus, hanem iteratív meglehetősen nagy számításigénnyel. Régóta köztudott, hogy Lawton és Sylvestre megközelítése analitikusan, tehát nem iteratív módon is megadhatja a megengedett megoldás sávhatárait. Ebben a közleményemben, elsőként a szakirodalomban, egyértelmű kapcsolatot mutatok be Lawton és Sylvestre, Tauler, valamint Maeder megközelítései között. A 9. ábra bal oldalán egy hipotetikus, kétkomponensű L_1 norma alapon normalizált adatrendszer spektrális profiljainak LS diagramját láthatjuk. A középtájon átfutó folytonos fekete vonalat a normalizált és transzformált adatpontok határozzák meg. Az origóból kiinduló folytonos kék, ill. piros vonalak a külső határokat jelölik, míg a szaggatott vonalak a belső határokat: az 1. és 2. komponensekhez tartozó megengedett megoldás-tartományok az azonos színű folytonos és a szaggatott vonalak között terülnek el. A feltüntetett t -érték párok koordinátái (\bar{x}_1, \bar{x}_2) lesznek a \mathbf{T} transzformációs mátrix elemei.

A 9. ábra jobb oldalán a transzformációs mátrix főátlójában található elemekkel normalizált $(\bar{\mathbf{T}})$ esetet tüntettük fel, megfigyelhető, hogy ennek következtében az adatpontokat jelképező egyenes függőlegessé alakult.

Az xAu , xAl , xBl , xBu metszéspontok és az α_N , α_F , β_N , β_F szögek alapján teremthető meg a kapcsolat Tauler és az LS módszerek között. A 9. ábra bal oldalán bemutatott grafikonra vonatkozóan:

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} \quad (37)$$



9. ábra A levezetésekhez és értelmezéshez szükséges illusztrációk.

A 9. ábra jobb oldalán bemutatott grafikonra vonatkozóan:

$$\bar{\mathbf{T}} = \begin{pmatrix} 1 & 0 \\ t_{11} & 0 \\ 0 & 1 \\ & t_{22} \end{pmatrix} \mathbf{T} = \begin{pmatrix} 1 & 0 \\ t_{11} & 0 \\ 0 & 1 \\ & t_{22} \end{pmatrix} \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} = \begin{pmatrix} 1 & \bar{t}_{12} \\ \bar{t}_{21} & 1 \end{pmatrix} \quad (38)$$

A Tauler-Gemperline-féle SCF maximumának és minimumának kétkomponensű rendszerekre vonatkozó számításához szükséges $\bar{\mathbf{T}}$ transzformációs mátrix elemeket a következőképpen számolhatjuk az 1. komponensre:

$$\bar{t}_{12,MAX,1} = \frac{\bar{x}_{2,xBl}}{\bar{x}_{1,xBl}}, \bar{t}_{21,MAX,1} = \frac{\bar{x}_{1,xAu}}{\bar{x}_{2,xAu}}; \bar{t}_{12,MIN,1} = \frac{\bar{x}_{2,xBu}}{\bar{x}_{1,xBu}}, \bar{t}_{21,MIN,1} = \frac{\bar{x}_{1,xAl}}{\bar{x}_{2,xAl}} \quad (39)$$

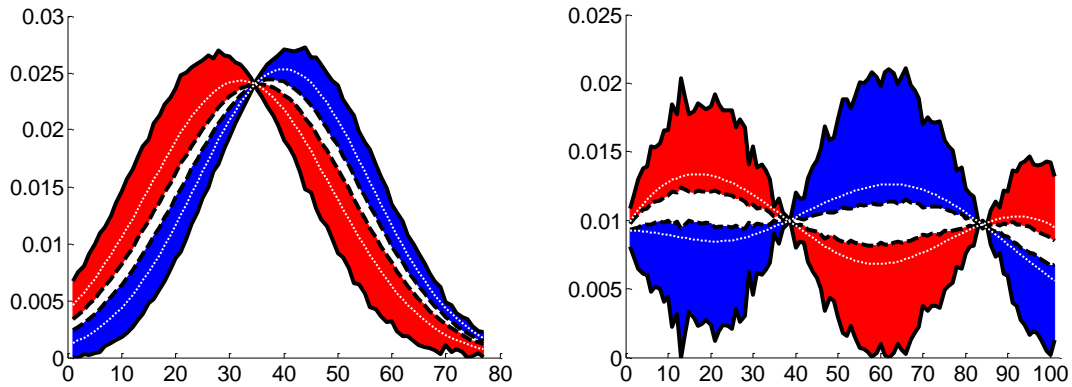
illetve a 2. komponensre:

$$\bar{t}_{12,MAX,2} = \frac{\bar{x}_{2,xBu}}{\bar{x}_{1,xBu}}, \bar{t}_{21,MAX,2} = \frac{\bar{x}_{1,xAl}}{\bar{x}_{2,xAl}}; \bar{t}_{12,MIN,2} = \frac{\bar{x}_{2,xBl}}{\bar{x}_{1,xBl}}, \bar{t}_{21,MIN,2} = \frac{\bar{x}_{1,xAu}}{\bar{x}_{2,xAu}} \quad (40)$$

Érdekességként megadjuk pl. az 1. komponensre vonatkozó 39. egyenletben szereplő \bar{t} értékeket a 9. ábra jobb oldali grafikonján szereplő szögek segítségével:

$$\begin{aligned} \bar{t}_{12,MAX,1} &= \tan(-\beta_N), \bar{t}_{21,MAX,1} = \tan(\alpha_N); \\ \bar{t}_{12,MIN,1} &= \tan(-\beta_F), \bar{t}_{21,MIN,1} = \tan(\alpha_F) \end{aligned} \quad (41)$$

A könnyebb ellenőrizhetőség végett tanulmányomhoz mellékeltem a részletes számításokat elvégző Matlab szkriptet, ill. néhány mesterségesen előállított hibaszint mellett a minimum és maximum SCF-ekhez tartozó sávhatárokat is, megfigyelve a hibaszint okozta jellegzetes torzulásokat (10. ábra).



10. ábra A koncentráció és spektrum profilok LS sávjai $SD=0,01$ additív Gauss zaj mellett.

Meg kell említeni, hogy amíg a Tauler-féle *mcrbands* és a Maeder-féle rácsmódszer az eredeti kezdő profilbecslésekhez ad határokat, addig az LS-alapú megközelítés csak normalizált határokat ad, azonban ez utóbbi eljárás analitikus és nem-iteratív.

A \mathbf{T} és $\bar{\mathbf{T}}$ mátrixok kiszámítását végző Matlab szkript:

```
[I,J] = size(R); % I is the number of rows,
                % J is the number of columns of R

comp = 2; % number of components
Run = R./ (sum(R')'*ones(1,J));
        % row-wise 1-norm normalization of R (R_unit_norm)

myoptions = pca('options');
myoptions.display = 'off';
myoptions.plots = 'none';

model = pca(Run,comp,myoptions); % pca is implemented in PLS_Toolbox
X = model.loads{1}; % the score matrix
V = model.loads{2}; % the loading matrix

% alternatively if PLS_Toolbox is not available:
% [U,Sv,V] = svds(Run); X = U(:,1:comp)*Sv(1:comp,1:comp); V = V(:,1:comp);
m_in_u = max(X(:,2)./X(:,1)); % inner area upper bound slope
m_in_l = min(X(:,2)./X(:,1)); % inner area lower bound slope
m_out_u = min(abs(V(V(:,2)<0,1)./V(V(:,2)<0,2)));
        % outer area upper bound slope
m_out_l = -min(abs(V(V(:,2)>0,1)./V(V(:,2)>0,2)));
        % outer area lower bound slope
m_unar = ones(1,J)*V; % coefficient vector a in eq. 10
```

```

xAu = ([m_unar; -m_out_u, 1]\[1; 0]); % intersection point of eq. 10 and
      % outer area upper bound line: upper (outer) bound point of component A
xA1 = ([m_unar; -m_in_u, 1]\[1; 0]); % intersection point of eq. 10 and
      % inner area upper bound line: lower (inner) bound point of component A
xBu = ([m_unar; -m_out_l, 1]\[1; 0]); % intersection point of eq. 10 and
      % outer area lower bound line: upper (outer) bound point of component B
xB1 = ([m_unar; -m_in_l, 1]\[1; 0]); % intersection point of eq. 10 and
      % inner area lower bound line: lower (inner) bound point of component B
tb_LS=[xB1; xAu; xBu; xA1; xBu; xA1; xB1; xAu];
      % transformation matrices given by mcrbands form
tb_LS_M = tb_LS./[tb_LS(1,[1 1]); tb_LS(2,[2 2]); tb_LS(3,[1 1]); ...
                 tb_LS(4,[2 2]); tb_LS(5,[1 1]); tb_LS(6,[2 2]); ...
                 tb_LS(7,[1 1]); tb_LS(8,[2 2])];
      % eq. 13 based simplification

Sini = V*[xB1; xA1]'; % initial estimations of spectra
% Sini = V*[xBu; xAu]'; % initial estimations in case of large noise
Cini = Run/Sini'; % initial estimations of concentration
[sb,cb,normb,tb,fb]=mcrbands(Run,Cini,Sini',eye(2));
      % mcrbands is Tauler's Matlab script (see below how to download it)
t_ = Sini'*V; % transformation matrix to get t_*V' = Sini'
for i=1:2:8, tb(i:i+1,:) = tb(i:i+1,)*t_; end; % modified tb for X and V
tb_tr = inv(diag(tb*sum(V)'))*tb;
      % transformation of tb being able to compare tb_tr with tb_LS
tb_tr_M = tb_tr./[tb_tr(1,[1 1]); tb_tr(2,[2 2]); tb_tr(3,[1 1]); ...
                 tb_tr(4,[2 2]); tb_tr(5,[1 1]); tb_tr(6,[2 2]); ...
                 tb_tr(7,[1 1]); tb_tr(8,[2 2])];
      % eq. 13 based simplification

% tb_M and tb_LS as well as tb_tr_M and tb_LS_M should be the same!!!

```

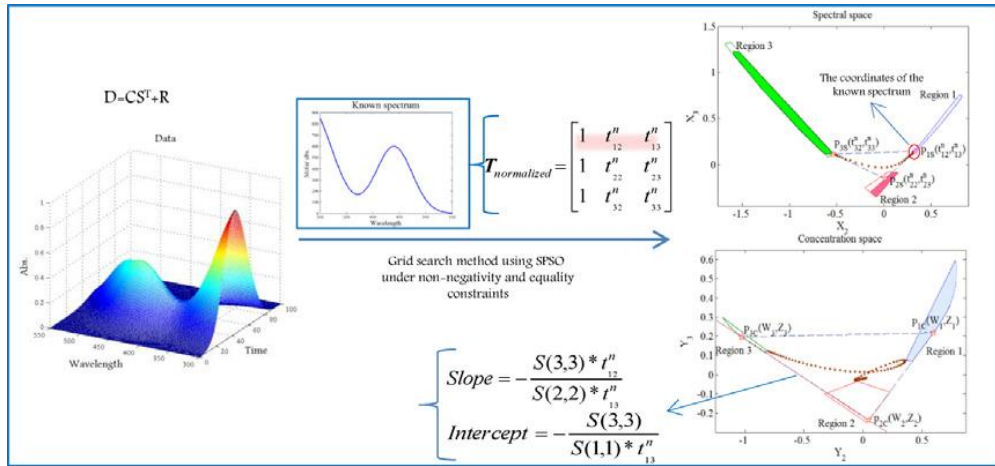
4.6 Ismert-profil kényszerfeltétel hatása a forgatási bizonytalanság csökkenésére [P6]

Ebben a tanulmányban háromkomponensű rendszerek esetén az összes összetartozó megengedett-megoldás halmazt egy speciális rácskeresés módszerrel határoztuk meg. Ez a módszer továbbfejlesztett formája egy korábbi rácskeresés módszernek [44], azaz a Matlab-ban elérhető *fminsearch* helyett a fajtaalapú részecske-raj-optimalizálás (Species-based Particle Swarm Optimization, SPSO) [45] algoritmuson alapul (11. ábra). Akkori legjobb tudásunk szerint ez volt az első olyan módszer, amely képes az összes összetartozó megoldást egyszerre megadni. A módszert zajmentes és zajos adatsor segítségével ellenőriztük.

Az eredmények azt mutatták, hogy a módszer alkalmas az adatok elemzésére, azonban a módszer meglehetősen időigényes. A kifejlesztett módszerrel az egyenlőség kényszerfeltételt (ismert koncentráció vagy spektrális profil figyelembe vétele: ismert-profil) is alkalmaztuk és várakozásunknak megfelelően a forgatási bizonytalanság és a megengedett megoldás-tartományok drasztikusan csökkentek (11. ábra).

Ebben a tanulmányban mutattuk meg részletekbe menően, hogy az ismert-profil kényszerfeltétel a komplementer komponensekre vonatkozóan egy lineáris egyenes megfelelő szakaszaira egyszerűsödő megengedett-tartományokat eredményez. Ehhez egy speciális Borgen norma normalizálást kell alkalmaznunk, amikor a \mathbf{z} vektor a csak pozitív elemeket tartalmazó első sajátvektor lesz. Ezt a normalizálást elérhetjük úgy is, hogy a \mathbf{T} transzformációs mátrix első oszlopának elemeivel elosztjuk a megfelelő sorban lévő további elemeket:

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix} \rightarrow \bar{\mathbf{T}} = \begin{pmatrix} \frac{1}{t_{11}} & 0 & 0 \\ 0 & \frac{1}{t_{21}} & 0 \\ 0 & 0 & \frac{1}{t_{31}} \end{pmatrix} \mathbf{T} = \begin{pmatrix} 1 & \bar{t}_{12} & \bar{t}_{13} \\ 1 & \bar{t}_{22} & \bar{t}_{23} \\ 1 & \bar{t}_{32} & \bar{t}_{33} \end{pmatrix} \quad (42)$$



11. ábra Az összes összetartozó megoldást egyszerre megadó, a fajtaalapú részecskekeraj-optimalizálást és ismert-profil kényszerfeltételt alkalmazó rácskeresés módszerrel

Legyen ismert az első komponens (mondjuk spektrális) profilja, azaz a 42. egyenlet utolsó mátrixában az első sor elemei ismertek. A normált $\bar{\mathbf{T}}$ transzformációs mátrix inverze:

$$\bar{\mathbf{T}}^{-1} = \frac{1}{\det(\bar{\mathbf{T}})} \begin{pmatrix} \bar{t}_{33}\bar{t}_{22} - \bar{t}_{32}\bar{t}_{23} & -(\bar{t}_{33}\bar{t}_{12} - \bar{t}_{32}\bar{t}_{13}) & \bar{t}_{23}\bar{t}_{12} - \bar{t}_{22}\bar{t}_{13} \\ -(\bar{t}_{33} - \bar{t}_{23}) & \bar{t}_{33} - \bar{t}_{13} & -(\bar{t}_{23} - \bar{t}_{13}) \\ \bar{t}_{32} - \bar{t}_{22} & -(\bar{t}_{32} - \bar{t}_{12}) & \bar{t}_{22} - \bar{t}_{12} \end{pmatrix} \quad (43)$$

Ezt meg kell szoroznunk a \mathbf{D} szinguláris érték mátrixszal, hogy a duális térben vizsgálódhassunk:

$$\mathbf{D}\bar{\mathbf{T}}^{-1} = \frac{1}{\det(\bar{\mathbf{T}})} \begin{pmatrix} d_{11}(\bar{t}_{33}\bar{t}_{22} - \bar{t}_{32}\bar{t}_{23}) & -d_{11}(\bar{t}_{33}\bar{t}_{12} - \bar{t}_{32}\bar{t}_{13}) & d_{11}(\bar{t}_{23}\bar{t}_{12} - \bar{t}_{22}\bar{t}_{13}) \\ -d_{22}(\bar{t}_{33} - \bar{t}_{23}) & d_{22}(\bar{t}_{33} - \bar{t}_{13}) & -d_{22}(\bar{t}_{23} - \bar{t}_{13}) \\ d_{33}(\bar{t}_{32} - \bar{t}_{22}) & -d_{33}(\bar{t}_{32} - \bar{t}_{12}) & d_{33}(\bar{t}_{22} - \bar{t}_{12}) \end{pmatrix}, \quad (44)$$

majd normalizálnunk kell (az inverz képzés miatt az oszlopok és sorok szerepe felcserélődött), azaz az első sor elemeivel a megfelelő oszlopok elemeit elosztjuk (a determináns értéke kiesik, mivel a számlálóban és nevezőben is megjelenik):

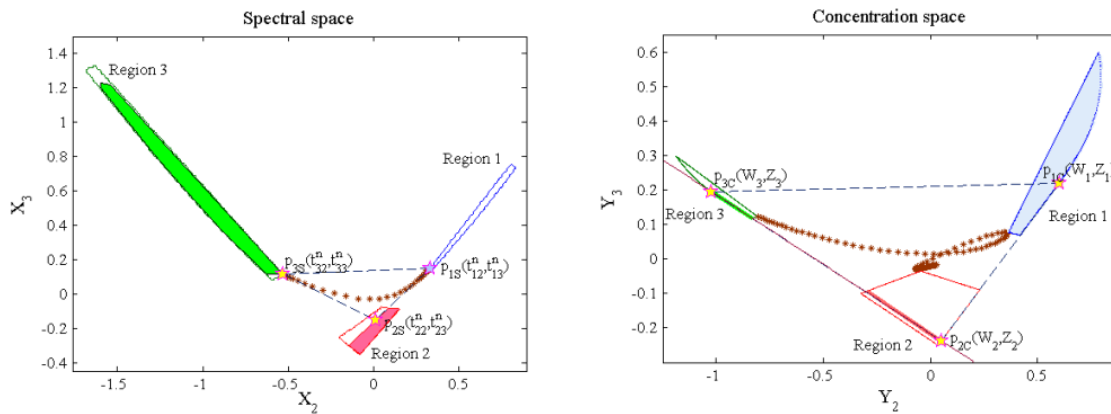
$$\overline{\overline{\mathbf{D}\bar{\mathbf{T}}^{-1}}} = \mathbf{M} = \begin{pmatrix} 1 & 1 & 1 \\ \frac{-d_{22}(\bar{t}_{33} - \bar{t}_{23})}{d_{11}(\bar{t}_{33}\bar{t}_{22} - \bar{t}_{32}\bar{t}_{23})} & \frac{d_{22}(\bar{t}_{33} - \bar{t}_{13})}{-d_{11}(\bar{t}_{33}\bar{t}_{12} - \bar{t}_{32}\bar{t}_{13})} & \frac{-d_{22}(\bar{t}_{23} - \bar{t}_{13})}{d_{11}(\bar{t}_{23}\bar{t}_{12} - \bar{t}_{22}\bar{t}_{13})} \\ \frac{d_{33}(\bar{t}_{32} - \bar{t}_{22})}{d_{11}(\bar{t}_{33}\bar{t}_{22} - \bar{t}_{32}\bar{t}_{23})} & \frac{-d_{33}(\bar{t}_{32} - \bar{t}_{12})}{-d_{11}(\bar{t}_{33}\bar{t}_{12} - \bar{t}_{32}\bar{t}_{13})} & \frac{d_{33}(\bar{t}_{22} - \bar{t}_{12})}{d_{11}(\bar{t}_{23}\bar{t}_{12} - \bar{t}_{22}\bar{t}_{13})} \end{pmatrix} \quad (45)$$

Ennek a mátrixnak az oszlopai határozzák meg a spektrális profilok pontjainak megfelelő egyeneseket a duális térben, az első komponenshez a 2. és 3. oszlopok segítségével számolhatunk meredekséget (SLOPE) és tengelymetszetet (ICEPT):

$$\text{SLOPE} = \frac{m_{32}-m_{33}}{m_{22}-m_{23}} = \frac{\frac{-d_{33}(\bar{t}_{32}-\bar{t}_{12})}{-d_{11}(\bar{t}_{33}\bar{t}_{12}-\bar{t}_{32}\bar{t}_{13})} \frac{d_{33}(\bar{t}_{22}-\bar{t}_{12})}{d_{11}(\bar{t}_{23}\bar{t}_{12}-\bar{t}_{22}\bar{t}_{13})}}{\frac{d_{22}(\bar{t}_{33}-\bar{t}_{13})}{-d_{11}(\bar{t}_{33}\bar{t}_{12}-\bar{t}_{32}\bar{t}_{13})} \frac{-d_{22}(\bar{t}_{23}-\bar{t}_{13})}{d_{11}(\bar{t}_{23}\bar{t}_{12}-\bar{t}_{22}\bar{t}_{13})}} = -\frac{d_{33} \bar{t}_{12}}{d_{22} \bar{t}_{13}} = -\frac{d_{33} t_{12}}{d_{22} t_{13}} \quad (46)$$

$$\text{ICEPT} = \frac{d_{33}}{d_{11}} \frac{1}{\bar{t}_{13}} = \frac{d_{33} t_{11}}{d_{11} t_{13}} \quad (47)$$

A 12. ábrán ismert spektrum információ okozta lineáris egyenesen lévő szakaszok mint megengedett megoldás-tartományok láthatók.



12. ábra Ismert spektrum (bal oldali grafikonon kék csillag, 1. komponens) mint ismert-profil kényszerfeltétel okozta megengedett megoldás-tartományok amik lineáris egyenesen lévő szakaszokra redukálódnak a duális térben (jobb oldali grafikonon a 2. és 3. komponensekhez tartozó piros és zöld szakaszok)

4.7 Újabb fejlesztések a görbeillesztés nélküli komponensprofil-kinyerésben az ismert-profil és unimodalitás kényszerfeltételek alkalmazásával [P7]

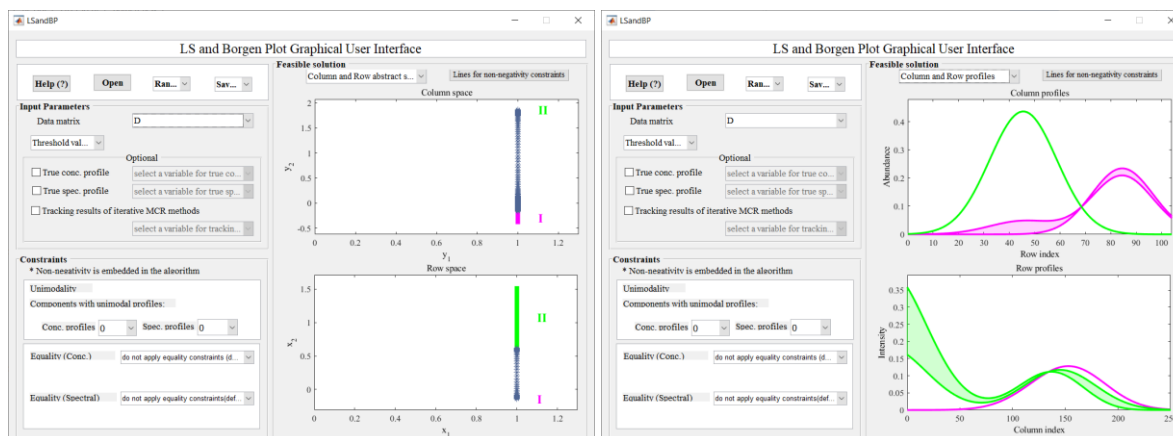
A görbeillesztés nélküli komponensprofil-kinyerés (SMCR) módszerei a mátrix alapú adatkészletekhez a teljes megengedett megoldás-tartományokat adják meg csak nemnegatív kényszerfeltétel alkalmazásával. A Lawton–Sylvestre módszer [20] volt az első, amit kétkomponensű rendszerek elemzésére vezettek be. Ezt általánosították háromkomponensű rendszerekre, pl. a Borgen grafikonok megrajzolásához is [28]. Úgy tűnik, hogy konvex geometriai szemlélet is szükséges a görbefelbontó módszerek megértéséhez és fejlesztéséhez, mivel a gyakran komplikált (lineáris) algebrai koncepciók kizárólagos alkalmazása megakasztotta a Borgen módszer alaposabb megértését 20 éven keresztül. Rajkó és István [P1] felülvizsgálták és tisztázták az SMCR módszerek elméletét, majd számítógépes geometriai módszereket vezettek be és alkalmaztak a Borgen grafikonok megrajzolásához háromkomponensű rendszerekhez. Ezek a fejlesztések, különösen a geometriai leírások, elméleti jellegűek és nehezen formalizálhatók zárt alakban megadható képletek segítségével.

Ebben a tanulmányban az analitikai SMCR módszereket vizsgáltuk felül és írtuk le a dualitás alapelveihez kapcsolódó egyszerű fogalmak segítségével. Ezenkívül, először a szakirodalomban, az egyenlőség (ismert-profil) és unimodalitás kényszerfeltételeket sikeresen implementáltuk a Lawton–Sylvestre módszerhez.

A Borgen grafikonok megrajzolásához is egy korszerűbb eljárást fejlesztettünk az ismert-profil kényszerfeltétel alkalmazásához.

Két- és háromkomponensű HPLC-DAD adatsorokat szimuláltunk és elemeztük őket az újabb kényszerfeltételek felhasználásával és azok nélkül.

A 13. ábrán kétkomponensű szimulált adatsor feldolgozása látható az általunk fejlesztett *LSandBP* Matlab szkript felhasználásával. A bal oldali képernyőképen az absztrakt térben a szakaszokkal jellemezhető megengedett megoldás-tartományokat láthatjuk, míg a jobb oldali képernyőképen az absztrakt térből visszatranszformált profilsávokat tanulmányozhatjuk.



13. ábra Kétkomponensű szimulált adatsor feldolgozása *LSandBP* Matlab szkripttel

A 14. ábrán háromkomponensű szimulált adatsor feldolgozása látható. A bal oldali képernyőképen az absztrakt térben a kék, lila és zöld színekkel jelölt megengedett megoldás-tartományokat nézhetjük meg. A pontozott szaggatott vonallal jelölt 2 db háromszög a Borgen szimplexek, amik a megengedett megoldás-tartományokat határolják be. A jobb oldali képernyőképen most is az absztrakt térből vissza-transzformált megengedett megoldásokat tartalmazó profilsávokat tanulmányozhatjuk.

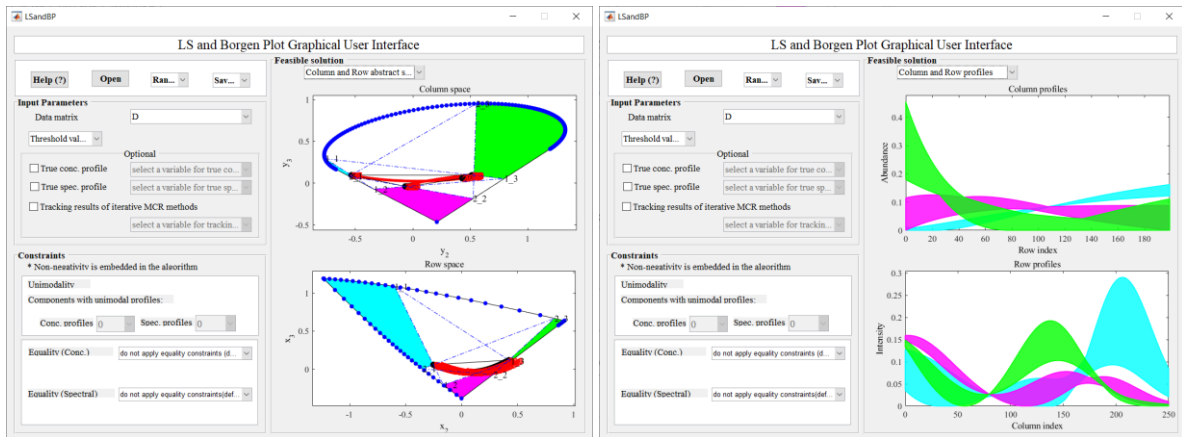
Az *LSandBP* Matlab szkriptünk a dualitás elvén alapuló egyszerűsítéseket tartalmazza így a 13. és 14. ábrákon bemutatott megjelenítés néhány másodperc alatt elkészült. A dualitás elvén alapuló egyszerűsítés pl., hogy az egyik absztrakt térben (mondjuk az oszlop térben) konvex burok algoritmussal meghatározott belső poligon a duális absztrakt térben (mondjuk a sor térben) a külső poligont határozza meg, ami igaz fordítva is, azaz a másik absztrakt térben (mondjuk a sor térben) meghatározott belső poligon az első absztrakt tér (mondjuk az oszlop tér) külső poligonját rögzíti. Így nem kell a számításigényes dupla-deszkripció (DD) számítógépes geometriai algoritmust igénybe vennünk. A másik egyszerűsítés, hogy elegendő csak az egyik absztrakt térben meghatározni a Borgen szimplexeket (Borgen háromszögek három komponens esetén), mert a dualitás elvének alkalmazásával a másik absztrakt térben iteráció nélkül megkapjuk azokat.

A továbbiakban a tanulmányunkban szereplő ábrákat bemutatva ismertetem az újabb kényszerfeltételek alkalmazását.

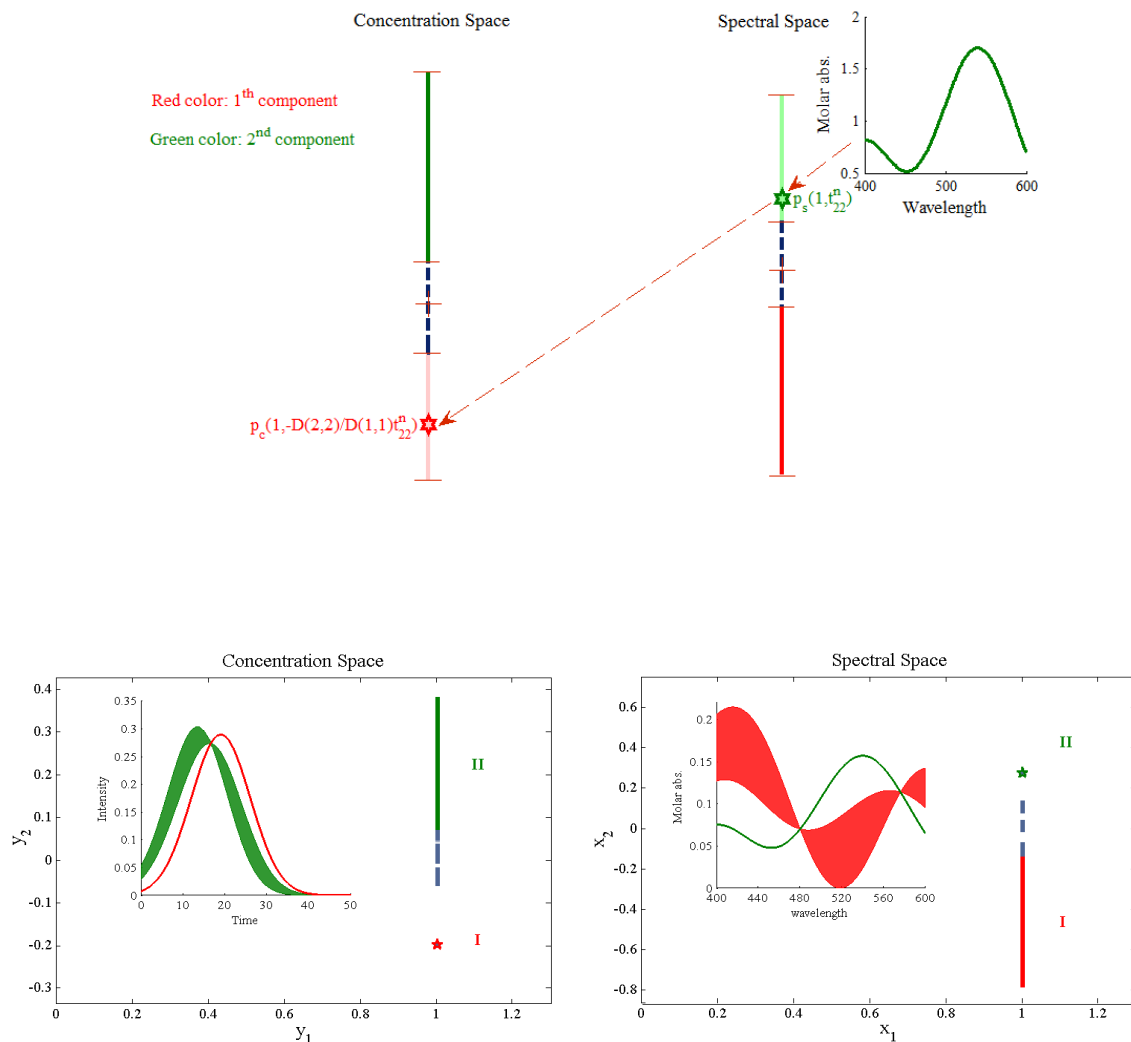
A 15. ábrán egy kétkomponensű rendszer LS grafikonjait láthatjuk a 2. komponens ismert spektruma felhasználásával. Vegyük észre, hogy az ismert spektrum a 2. komponens esetén az 1. komponens koncentráció profiljának egyértelmű meghatározását jelenti (a dualitás elve alapján), a megengedett megoldás-tartományok egyetlen pontra szűkülnek.

A 16. ábrán egy kétkomponensű rendszer LS grafikonjait láthatjuk az 1. komponens koncentráció profiljára az unimodális kényszerfeltétel alkalmazásával. Vegyük észre, hogy az unimodális kényszerfeltételnek csak egyetlen koncentráció profil felel meg, így a 2. komponens spektruma egyértelműen meghatározható (ismét a dualitás elve alapján), a megengedett megoldás-tartományok egyetlen pontra szűkülnek.

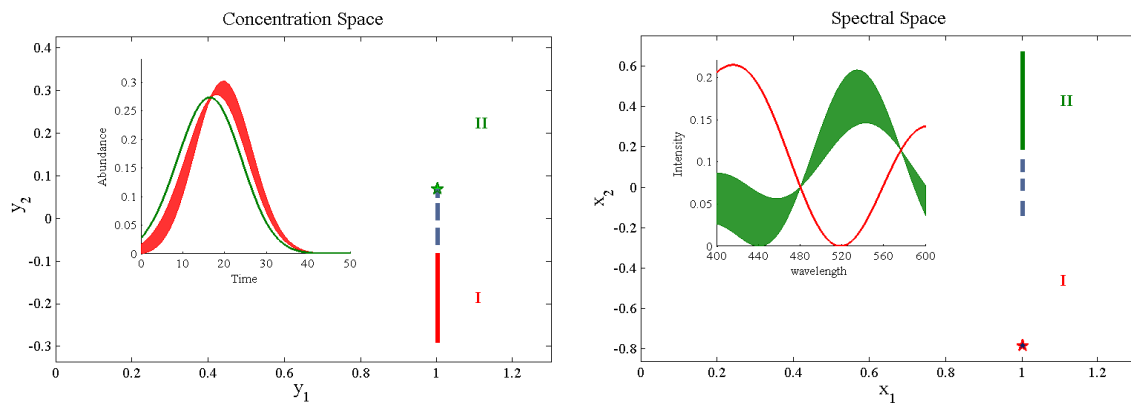
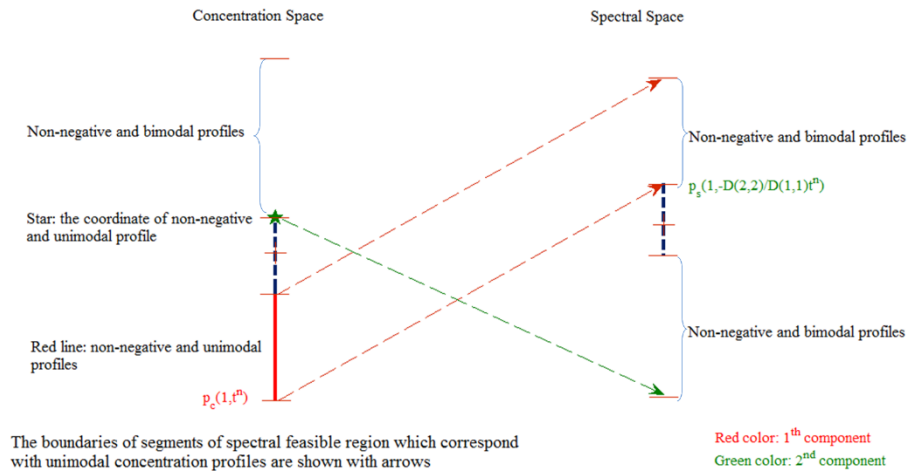
A 17. ábrán egy háromkomponensű rendszer Borgen grafikonjának megrajzolásának lépéseit láthatjuk, ahol a 2. komponens spektruma ismert. Első lépésként a belső és külső poligonok számítása és megjelenítése a feladat (17. ábra első sor). Második lépés a Borgen háromszögek meghatározása a dualitás elvének felhasználásával (17. ábra második sor). Harmadik lépés: az egyenlőség kényszerfeltétel alkalmazása, azaz a spektrális térben a külső poligon drasztikusan redukálódik (17. ábra harmadik sor). Végül a negyedik lépésben meghatározzuk a lecsökkenő megengedett megoldás-tartományokat és profilsávokat (17. ábra negyedik sor). Megfigyelhetjük, hogy a dualitás elvének megfelelően, a 2. komponens ismert spektruma a koncentráció térben az 1. és 3. komponens megengedett tartományait egyenesre korlátozza, az egyenes kék és piros szakaszaira (17. ábra negyedik sor jobb oldali grafikon).



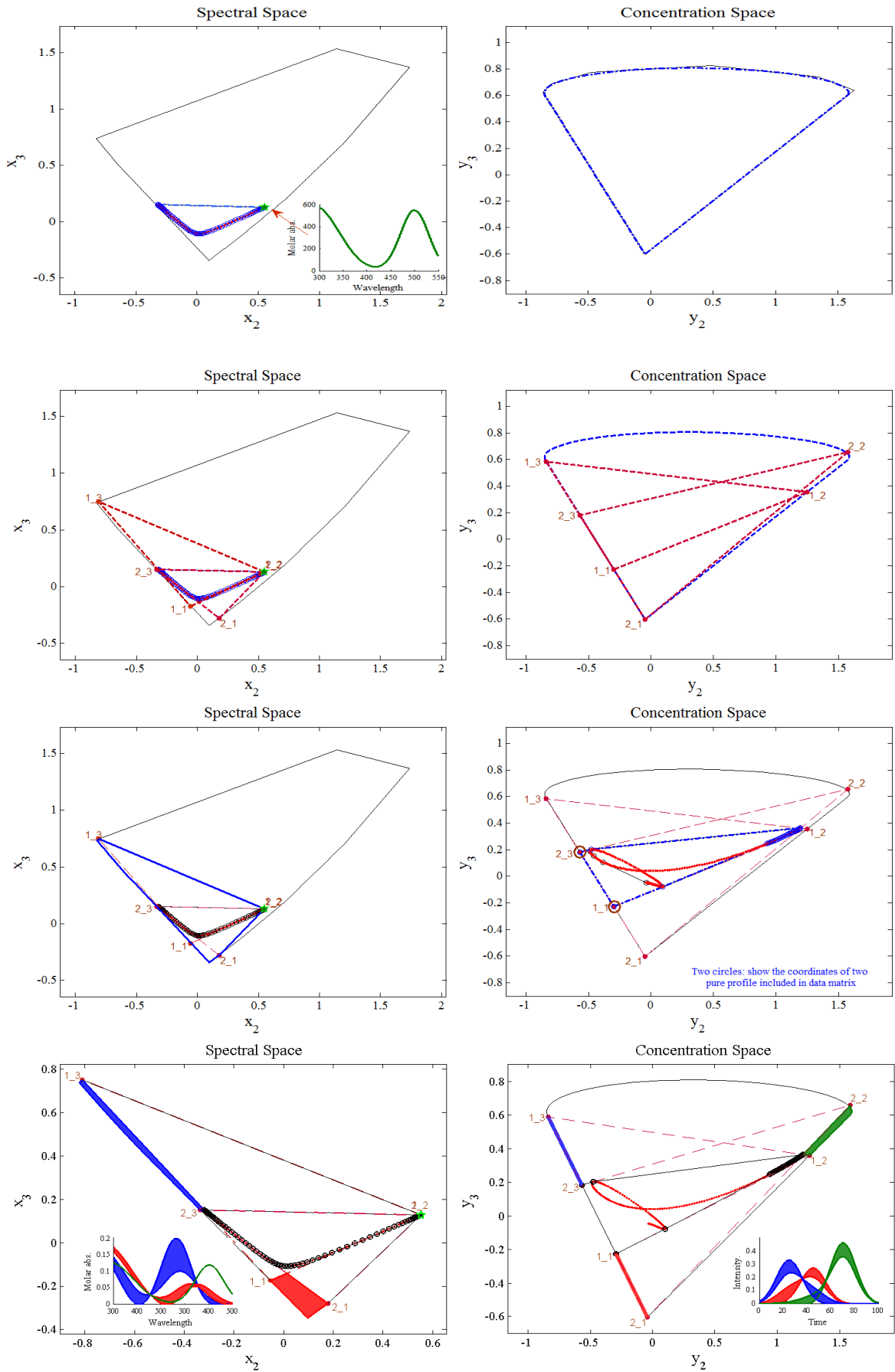
14. ábra Háromkomponensű szimulált adatsor feldolgozása LSandBP Matlab szkripttel



15. ábra Kétkomponensű rendszer esetén LS grafikonok megalkotása a 2. komponens ismert spektruma felhasználásával



16. ábra Kétkomponensű rendszer esetén LS grafikonok megalkotása az 1. komponens koncentráció profiljára az unimodális kényszerfeltétel alkalmazásával



17. ábra Háromkomponensű rendszer esetén Borgen grafikon megalkotása

4.8 Adat-alapú egyértelműen meghatározható profilbecslés definíciója és kimutatása [P8]

Ebben a tanulmányban a Manne tételek kritikáját és új elvek megfogalmazását végeztük el. Manne [46] a következő három tételt alkotta:

1. tétel – Ha minden zavaró kémiai komponens/vegyület, amely egy adott meghatározandó kémiai komponens/vegyület (analit) elúciós (koncentráció) jelenléti ablakán belül megjelenik és ezen az ablakon kívül is előfordul, akkor lehetséges az analit koncentráció profiljának meghatározása.
2. tétel – Ha minden zavaró kémiai komponens esetén az analit elúciós jelenléti ablakának vannak olyan részei, ahol az összes zavaró kémiai komponens hiányzik, akkor az analit spektrális profilját (spektrumát) lehet meghatározni.
3. tétel – Kromatográfiás vizsgálatokban csak rang információon alapuló mátrixfelbontás esetén az 1. és 2. tételek feltételei nemcsak elégségesek, hanem szükségesek is.

Manne a 3. tétel megfogalmazása előtt kijelenti, hogyha akár az 1., akár a 2. tétel feltételei közül valamelyik nem teljesül, akkor nem egyértelmű (non-unique) felbontást kapunk. Ezért gondolhatták azt Manne cikkét és tételeit felhasználók, hogyha viszont mindkét tétel feltételei teljesülnek, akkor a felbontás egyértelmű (unique), amit a 3. tétel megerősít, hiszen az 1. és 2. tételek nemcsak elégségesek, hanem szükségesek is.

Tanulmányunkban sikerült bemutatnunk egy ellenpéldát, amikor Manne 1. és 2. tétele teljesül, de a felbontás nem lesz egyértelmű. A 18. ábrán láthatjuk, hogy a Manne tételek feltételei teljesülnek, azaz az 1. tételre vonatkozóan, minden zavaró kémiai komponens megjelenik az analit jelenléti ablakában (lila ablak) és ezen kívül is (világos barna ablak). Ezek szerint a piros komponens koncentráció profilja egyértelműen meghatározható. A 2. tétel feltételei is teljesülnek, az analit jelenléti ablakának van olyan része, ahol a zavaró komponensek hiányoznak (sárga ablak). Tehát a Manne tételek alapján a piros komponens spektruma is meghatározható egyértelműen, azaz a komponens mindkét profilja.

Azonban a 19. ábrán láthatjuk, hogy sem a piros komponens, sem a többi komponens profilsávjainak egyike sem redukálódott egyetlen vonallá, ami az egyértelmű felbontást jelezné.

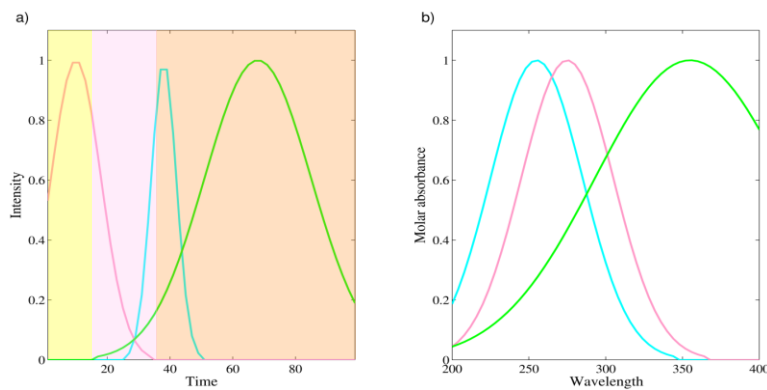
A megtalált ellenpélda ösztönzött bennünket, hogy áttekintsük újra az egyértelmű felbontás lehetőségeit, ill. a Manne tételek helyett megfelelő megfogalmazást találjunk. A Manne tételek a valamilyen iterációs módszerrel (pl. NMF vagy MCR-ALS) becsült profil becslésekre vonatkoznak. Ezzel ellentétben mi az LS és Borgen grafikonok által szolgáltatott sávmegoldásokat vettük alapul.

A Manne tételek így profil-alapú egyértelműségekre vonatkoznak, míg az általunk bevezetett eljárás az adat-alapú egyértelműségekre vonatkozik.

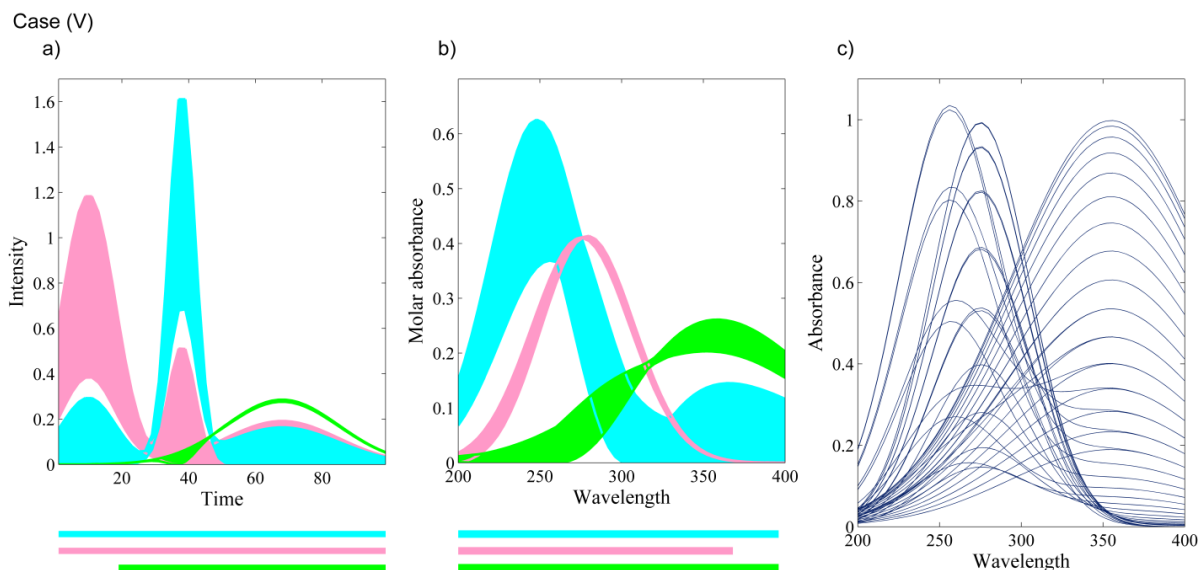
Theorem 1. *If all interfering compounds that appear inside the concentration window of a given analyte also appear outside this window, it is possible to calculate the concentration profile of the analyte.*

Theorem 2. *If for every interferent the concentration window of the analyte has a subwindow where the interferent is absent, then it is possible to calculate the spectrum of the analyte.*

Theorem 3. *For a resolution based only upon rank information in the chromatographic direction the conditions of Theorems 1 and 2 are not only sufficient but also necessary.*



18. ábra Ellenpélda a Manne tételekre: az 1. és 2. tételek feltételei teljesülnek



19. ábra Ellenpélda a Manne tételekre: a profilsávok alapján nincs egyértelmű megoldás

Két formában adhatjuk meg az adat-alapú egyértelműen megadható profilokra vonatkozó tételünket:

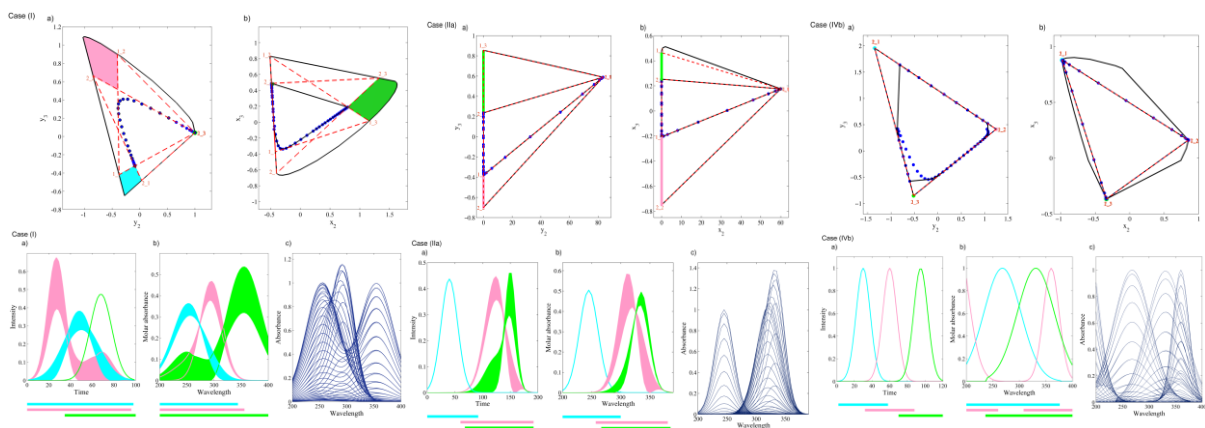
1. forma – Ha a Borgen grafikonon (vagy annak általánosításán háromnál több komponens esetén) találunk olyan pontot amely mind a belső és mind a külső poligonok valamely csúcspontjaihoz tartozik és ennek következtében a Borgen háromszögek (általánosan Borgen szimplexek) közös csúcspontja, ez a pont egyértelmű megoldás lesz az adott komponens részére.
2. forma – Ha egy komponensnek van szelektív ablaka az egyik irányban/módban/útban (pl. spektrum) és a másik irányt/módot/utat (pl. koncentráció) tekintve ennek a komponensnek van zéró ablaka, ahol az összes zavaró komponens megjelenik, akkor ennek a komponensnek a második irányban/módban/útban található (koncentráció) profilja egyértelműen megadható.

Elsőként a kemometriai szakirodalomban az alábbi egyértelműségi kategóriákat vezettük be:

- Nem-egyértelmű felbonthatóság (non-uniqueness), azaz egyetlen profil sem adható meg egyértelműen.
- Töredékes-egyértelműség (fractional uniqueness): egy komponens egyetlen profilja adható meg egyértelműen, a többi nem.

- Részleges-egyértelműség (partial uniqueness): egy komponens összes profilja egyértelműen megadható
- Teljes-egyértelműség (full uniqueness): az összes komponens összes profilja egyértelműen megadható.

A nem-egyértelműen felbontható adatmárixra a Manne tételekhez használt ellenpéldánkat hozhatjuk (19. ábra). A többi egyértelműség kategóriára a 20. ábrán találunk jellemző példákat.



20. ábra Bal oldali ábrahalmaz (Case I): töredékes-egyértelműség (fractional uniqueness), középső ábrahalmaz (Case IIa): részleges-egyértelműség (partial uniqueness), jobb oldali ábrahalmaz (Case IVb): teljes-egyértelműség (full uniqueness)

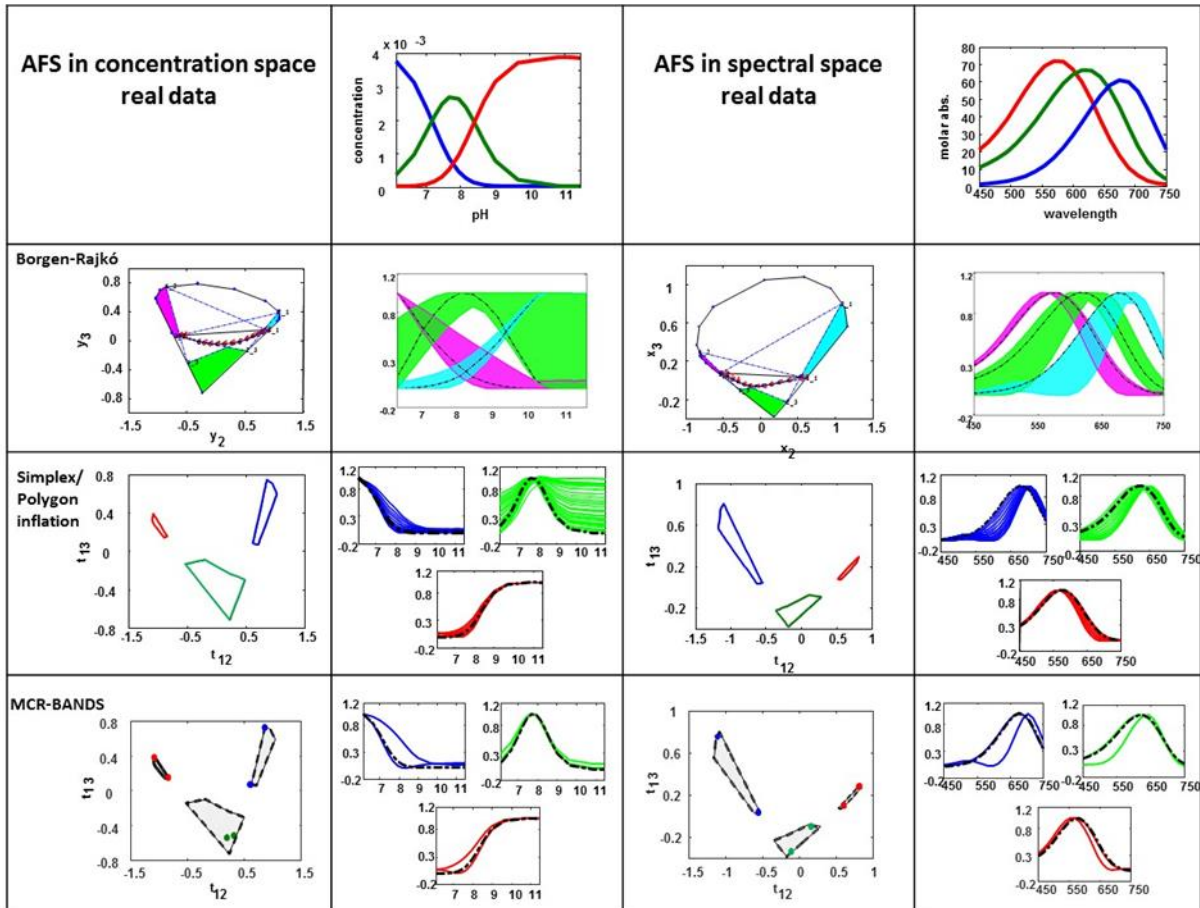
4.9 Újabb keletű görbeillesztés nélküli komponensprofil-kinyerő módszerek áttekintése [P9]

Ez az áttekintő mű az addig megjelent görbeillesztés nélküli komponensprofil-kinyerő módszerek áttekintését adta meg. Kiemelendő, hogy a társszerzők között Prof. Maeder, Prof. Tauler, Prof. Abdollahi és Prof. Neymeyr is ott voltak, az alkotás igazi nemzetközi kooperációban készült.

A Google Scholar alapján 2023. augusztus 31-én pl. Marcel Maeder 11483 hivatkozással és 47-es h-indexszel, Roma Tauler 29731 hivatkozással és 84-es h-indexszel rendelkezik.

A forgatási bizonytalanság feltárásával kapcsolatban szintetikus és valós adatokat elemezve megállapítottuk, hogy bár a végeredmény nagyon hasonló a bemutatott eljárások algoritmusai lényegi különbséget mutatnak. A Borgen-Rajkó grafikonok számítása explicit módon történik, így nagyon gyorsan megkapjuk az eredményt. Ez az elméleti megfontolásokon alapuló módszer érvényesítésre (validation) használható. Az egyetlen hátránya, hogy valós, zajjal terhelt adatokra nem mindig alkalmazható. A szimplex és poligon tágító (simplex and polygon inflation) módszerek eredménye nagyon hasonlít az előző elméleti alapú módszerrel kapottal. Az előnye ennek a közelítő módszernek, hogy valós zajjal terhelt adatokkal is működik. Az MCR-BANDS algoritmus teljesen különböző. Ez talán a legkönnyebben elérhető program és bármilyen komponens szám esetén használható. Az algoritmus könnyen implementálható, főképp Matlab-ban, mivel ott készen elérhető nemlineáris optimalizációs eljárások használhatók nemlineáris kényszerfeltételekkel. Itt a zajhatást újramintázó, pl. bootstrapping eljárásokkal térképezhetjük fel. A hátránya, hogy csak közelíteni tudja a valós sávmegoldásokat, mivel csak két extrém, norma-alapú megoldás alapján dolgozik.

A 21. ábrán összefoglalva láthatjuk a vizsgált eljárások eredményeit egy tipikus háromkomponensű modellrendszer esetén.



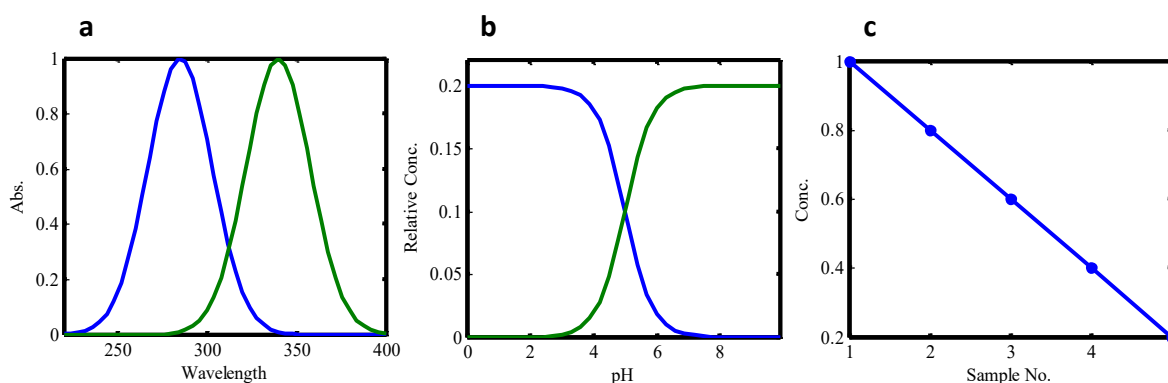
21. ábra Az áttekintő műben bemutatott módszerek összefoglaló ábrája

4.10 Két- és háromkomponensű három-utas adattömbök nem-negatív felbonthatóságának egyértelműsége [P10]

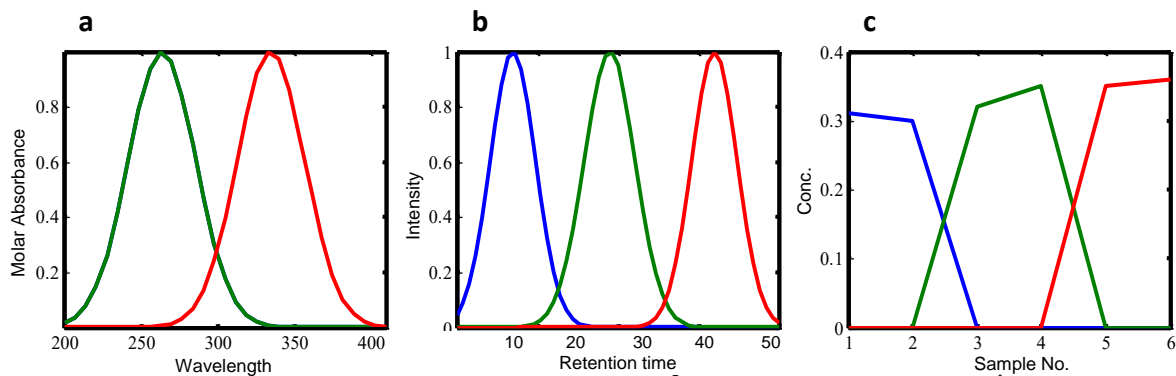
Az egyértelmű felbontás az egyik legvonzóbb tulajdonság, hogy trilineáris adattömböt hozunk létre kapcsolt technikákkal a méréseink során. Sajnos a triadikus dekompozíció identifikálása még nem teljesen feltárt, annak ellenére, hogy elméleti és gyakorlati fontossága ismert. A rang-vesztéssel (rank deficiency) sújtott loadingú adattömbökre egyelőre Kruskal által bevezetett egyenlőtlenség alkalmazható. Ten Berge és Sidiropoulos [47] megmutatta, hogy három-utas CP modellt feltételezve $F = 2$ és $F = 3$ esetén a $(k_E + k_S + k_C \geq 2F + 2)$ Kruskal egyenlőtlenség teljesülése szükséges és elégséges feltétele az egyértelmű felbontásnak, viszont $F > 3$ esetén már nem lesz szükséges. Itt k_E , k_S , és k_C az egyes **E**, **S**, **C** loading mátrixok ún. k -rangja, azaz egy mátrix minden lehetséges módon kiválasztott k oszlopa lineárisan független, de van legalább egy $k+1$ db oszlopból álló részmatrix amely oszlopai már nem lesznek lineárisan függetlenek.

A következő jelölést vezetjük be: $^k\{F_1, F_2, F_3\}$ jelentse, hogy $k_E=F_1$, $k_S=F_2$, és $k_C=F_3$.

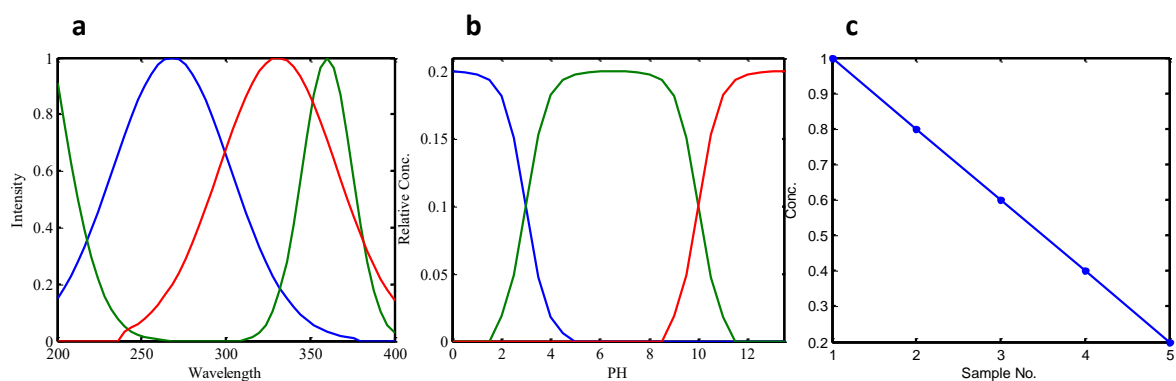
A tanulmányunkban két- és háromkomponensű három-utas adattömböket generáltunk speciális profil-tulajdonságokkal, de úgy, hogy a Kruskal egyenlőtlenség ne teljesüljön.



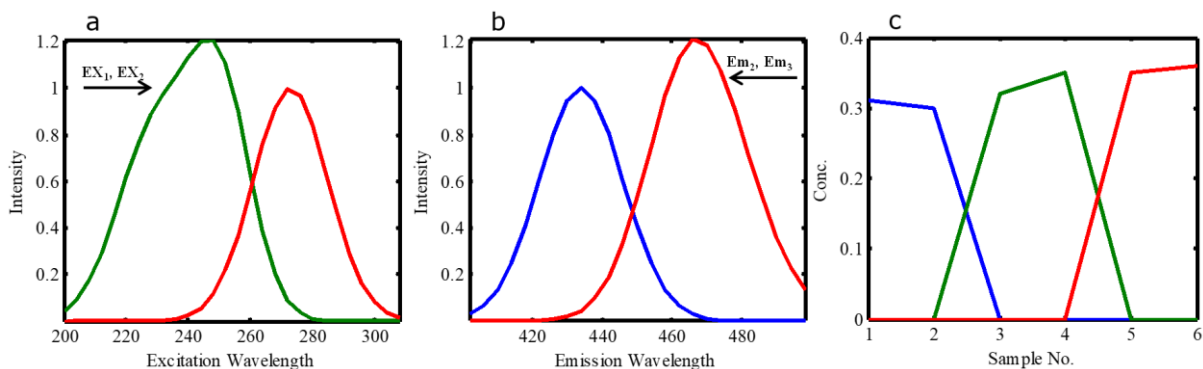
22. ábra Kétkomponensű szimulált másodrendű kalibráció spektro–pH-metriás titrálással. a) abszorpciós spektrumok, b) pH-függő összetétel profilok, c) relatív kezdeti koncentrációk a különböző mintákban



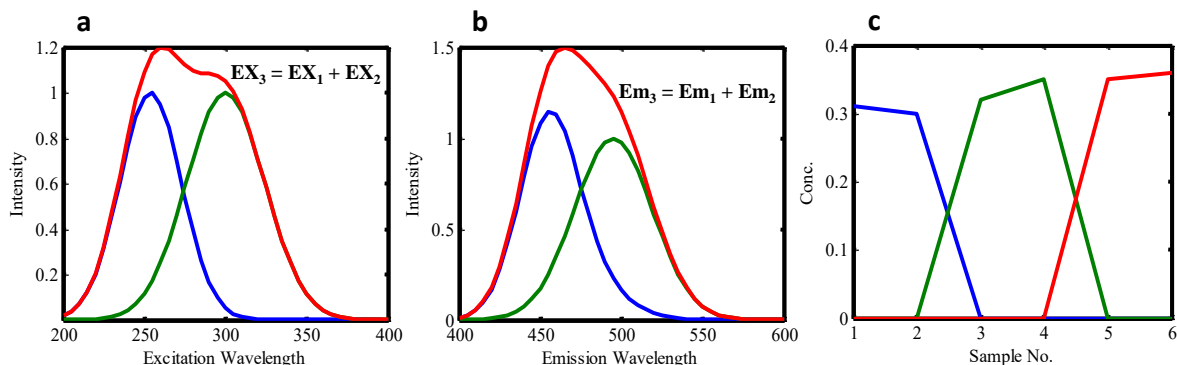
23. ábra Háromkomponensű szimulált HPLC-DAD adattömb. a) abszorpciós spektrumok, b) elúciós profilok, c) minták szerinti koncentráció profilok



24. ábra Háromkomponensű szimulált másodrendű kalibráció spektro-pH-metriás titrálással. a) abszorpciós spektrumok, b) pH-függő összetétel profilok, c) relatív kezdeti koncentrációk a különböző mintákban



25. ábra Háromkomponensű szimulált gerjesztett emissziós spektrum (EEM). a) gerjesztő spektrumok, b) emissziós spektrumok, c) minták szerinti koncentráció profilok



26. ábra Háromkomponensű szimulált gerjesztett emissziós spektrum (EEM). a) gerjesztő spektrumok, b) emissziós spektrumok, c) minták szerinti koncentráció profilok

A 22. ábrán kétkomponensű szimulált spektro-pH-metriás titrálással másodrendű kalibráció profiljai láthatók. A Kruskal egyenlőtlenség nem teljesül: $2 + 2 + 1 \not\geq 6$, mivel a 3. út szerint a két profil megegyezik. Azonban mégis egyértelmű megoldás/felbontás adható, mivel nemnegatív kényszerfeltételt használhatunk, ill. az 1. és 2. út szerint szelektív, ill. zéró ablakokat találunk.

Háromkomponensű rendszerek esetén, $F = 3$, az összes lehetőség, amikor a Kruskal feltételek nem teljesülnek a következők [47]:

1. Amikor valamelyik loading mátrix k -rangja kisebb mint 2.
2. Amikor valamelyik loading mátrix teljes rangú, a másik kettőben viszont lineárisan összefüggő profilok lesznek.
3. Amikor mindhárom loading mátrix k -rangja 2: $k\{2,2,2\}$.

A 23. ábrán háromkomponensű szimulált HPLC-DAD adattömb profiljait láthatjuk. Az 1. út szerinti loading mátrixban az 1. és 2. komponens abszorpciós spektruma megegyezik, azaz $k\{1,3,3\}$ és a Kruskal egyenlőtlenség nem teljesül: $1 + 3 + 3 \not\geq 8$. Itt is alkalmazhatjuk a nemnegatív kényszerfeltételt és megfigyelhetjük a szelektív, ill. zéró ablakokat mindhárom út szerint, ezért a felbontás egyértelműen elvégezhető.

A 24. ábrán háromkomponensű szimulált spektro-pH-metriás titrálással másodrendű kalibráció profiljai láthatók. A Kruskal egyenlőtlenség itt sem teljesül: $3 + 3 + 1 \not\geq 8$, mivel a 3. út szerint két profil megegyezik. Azonban itt is egyértelmű megoldás/felbontás adható, mivel nemnegatív kényszerfeltételt használhatunk, ill. az 1. és 2. út szerint szelektív, ill. zéró ablakokat találunk.

A 25. ábrán háromkomponensű szimulált gerjesztett emissziós spektrum (EEM) profiljait láthatjuk másodrendű kalibrációhoz. Az 1. út szerint az 1. és 2. komponens gerjesztett spektruma, ill. a 2. út szerint a 2. és 3. komponens emissziós spektruma megegyezik, azaz a Kruskal egyenlőtlenség nem teljesül: $1 + 1 + 3 \not\geq 8$. Mivel nemnegatív kényszerfeltételt használhatunk, ill. mindhárom út szerint szelektív, ill. zéró ablakokat találunk, a felbontás egyértelmű lesz.

A 26. ábrán szintén háromkomponensű szimulált gerjesztett emissziós spektrum (EEM) profiljait láthatjuk másodrendű kalibrációhoz, azonban most a 3. komponens gerjesztett és emissziós spektrumai a másik két komponens megfelelő spektrumainak összegéből áll elő, azaz a Kruskal egyenlőtlenség nem teljesül: $2 + 2 + 3 \not\geq 8$. Itt csak a 3. út szerint találunk szelektív, ill. zéró ablakokat, de a nemnegativitás, ill. trilinearitás kényszerfeltételek alkalmazása ismét biztosítja az egyértelmű felbontást.

A $k\{2,2,2\}$ esetben a Kruskal egyenlőtlenség nem teljesül: $2 + 2 + 2 \not\geq 8$. Mivel szelektív ablakok biztosan nem alakulhatnak ki mindhárom út szerinti profilok esetében lineáris függés miatt, egyértelmű felbontás nem lehetséges.

A fenti vizsgálódásokból levonhatjuk a következtetést, hogy Kruskal elégséges és szükséges feltétele két-, ill. háromkomponensű háromutas adattömbökre sérülhet és legfeljebb csak elégséges maradhat. Ten Berge és Sidiropoulos [47] igazolta, hogy $F > 3$ esetekben a nulla értékek speciális mintázata esetén lehetséges egyértelmű felbontás annak ellenére, hogy a Kruskal egyenlőtlenség nem teljesül, ezt most mi igazoltuk $F = 2$ és $F = 3$ esetekre is. Az adat-alapú egyértelműség elvének alkalmazásával a következő feltételeknek kell teljesülniük az egyértelmű felbontáshoz a Kruskal egyenlőtlenség nem teljesülése esetén: 1) trilinearitás, 2) szelektív ablakok jelenléte, ahogy Omidikia et al. [48] leírták, és végül 3) a speciálisan elhelyezkedő zérusok mintázata, ahogy ten Berge and Sidiropoulos [47], ill. Paatero et al. [49] jelezték.

A Kruskal egyenlőtlenség teljesülése tehát nem feltétlenül szükséges a *kényszerfeltételes trilineáris dekompozícióhoz* két-, ill. háromkomponensű háromutas adattömbök esetében.

4.11 Megengedett megoldás-tartományok analitikus megadása és értelmezése kétkomponensű három-utas adattömbök esetén [P11]

Ebben a tanulmányunkban a mátrix alapú adatok helyett a kétkomponensű három-utas (azaz 3D) adattömböket vizsgáltuk. Ily módon az LS módszer egy újabb általánosítását kaptuk.

Legyen \mathbf{A} egy $I \times J \times K$ méretű adattömb (adatkocka), egy tipikus eleme a_{ijk} . Akkor a tömb F -komponensű trilineáris felbontása:

$$a_{ijk} = \sum_{f=1}^F e_{if} s_{jf} c_{kf} \quad (48)$$

Egy trilineáris adatstruktúrát a \odot Khatri-Rao szorzással (oszloponkénti Kronecker szorzat) [20] is felírhatunk:

$$\mathbf{A}_{I,J,K}^1 = \mathbf{E}_{I,F} (\mathbf{C}_{K,F} \odot \mathbf{S}_{J,F})^T + \mathbf{R}_{I,J,K} \quad (49)$$

Ahol $\mathbf{A}_{I,J,K}^1$ mátrix az \mathbf{A} adattömb első irányában (útjában) vett kiterítése (vesd össze a 6. egyenlettel). A tanulmányban $F=2$, azaz kétkomponensű gerjesztett emissziós spektrumokat (Excitation Emission Matrix, EEM) szimuláltunk, így most az \mathbf{E} ($I \times 2$), \mathbf{S} ($J \times 2$), és \mathbf{C} ($K \times 2$) mátrixok a gerjesztő spektrum, emissziós spektrum és összetétel (koncentráció) mátrixokat jelentik.

Egy adattömb egyértelmű felbonthatóságát a k -rang (Kruskal-rang) bevezetésével fogalmazhatjuk meg. Egy mátrix k -rangú, ha minden lehetséges módon kiválasztott k oszlopa lineárisan független, de van legalább egy $k+1$ db oszlopból álló részmatrix amely oszlopai már nem lesznek lineárisan függetlenek [23,50]. A következő jelölést vezettük be: ${}^k\{F_1, F_2, F_3\}$ jelentse, hogy $k_E = F_1$, $k_S = F_2$, és $k_C = F_3$. Ebben a tanulmányban először ${}^k\{2, 2, 2\}$ adattömböt vizsgáltunk, ahol a $(k_E + k_S + k_C \geq 2F + 2)$ Kruskal egyenlőtlenség [23] teljesül, mivel $k_E + k_S + k_C = 6$ és $2F + 2 = 6$.

A megengedett megoldás-tartományok meghatározásához kétkomponensű három-utas adattömbök esetén először kiterítettük az adattömböt a választott irány-

ban (sorirány (I) vagy oszlopirány (J) vagy szálirány (K)), majd az LS módszert alkalmaztuk. Ha sorok irányában terítünk ki, akkor a következőképpen írhatjuk fel a kiterített mátrixot bilineáris formában:

$$\mathbf{A}_{I,J \cdot K}^1 = \mathbf{E}_{I,2} (\mathbf{S} \cdot \mathbf{C}^T)_{2,J \cdot K} = \mathbf{X}_{I,2} (\mathbf{V}^T)_{2,J \cdot K} \quad (50)$$

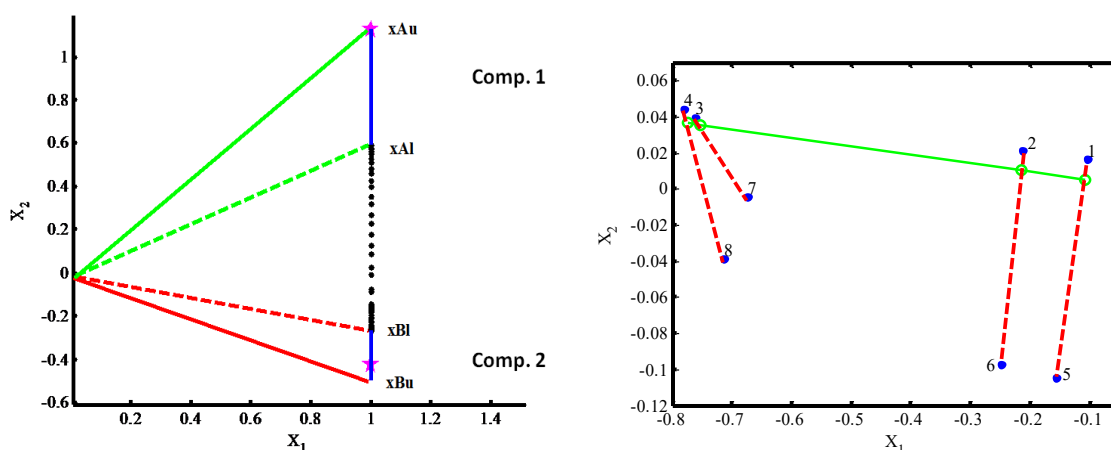
$\mathbf{S} \cdot \mathbf{C}$ ($J \cdot K \times 2$) az eredeti \mathbf{S} és \mathbf{C} mátrixok kiterítés okozta keverékmátrixa, amit Khatri-Rao szorzással kapunk (v.ö. 49. egyenlettel).

A trilineáris kényszerfeltétel alkalmazása a következő konvex lineáris kombinációt jelenti a 27. ábra bal oldalán lévő grafikonon bemutatott megengedett megoldástartomány mint szakasz végpontjainak jelölésével:

$$\alpha \cdot xAu + (1 - \alpha) xAl = xTR \quad (51)$$

α egy skalár 0 és 1 közötti értékkel, xAu és xAl az első komponens megoldás-tartományának külső és belső határoló pontjai, xTR pedig a trilineáris megoldás ehhez a komponenshez.

Mindkét komponens felső és alsó határoló pontjait visszatranszformáljuk, azaz megkapjuk az fAu , fAl , ill. fBu , fBl profilokat az első, ill. második komponensre vonatkozóan.



27. ábra Bal oldalon: az első irányban kiterített mátrix LS grafikonja. Jobb oldalon: az (1,2,3,4) kék pontok az \mathbf{fAuM} oszlopaihoz, az (5,6,7,8) kék pontok az \mathbf{fAIM} oszlopaihoz tartoznak, trilineáris megoldás szerint a zöld körökkel jelölt pontok α arányban osztják fel a szaggatott pirossal jelölt szakaszokat

Az $(1 \times J \cdot K)$ méretű profilokat $(J \times K)$ méretű mátrixszá alakítjuk, pl. az fAu $(1 \times J \cdot K)$ profilból $fAuM$ $(J \times K)$ mátrix lesz. A 51. egyenlethez hasonlóan tekintjük ezen mátrixok oszlopainak konvex lineáris kombinációit, pl.:

$$\alpha \cdot fAuM_k + (1 - \alpha) fAlM_k = fAMTR_k \quad (52)$$

Tehát az $fAMTR_k$ egy trilineáris megoldás lesz, ami az egyrangú $(J \times K)$ mátrix k . oszlopa. A $(J \times K)$ méretű **fAuM**, **fAlM**, **fBuM** és **fBIM** mátrixok oszlopait L_1 norma alapján normalizáljuk. Ha egy trilineáris kevert-profil mátrixszá alakítunk és oszlopok szerint normalizálunk, akkor minden oszlop meg fog egyezni a másikkal (ez a trilineáris tulajdonság). Így az **fAuM** és **fAlM** p . és q . oszlopainak normalizált lineáris kombinációi megegyeznek:

$$\frac{[\alpha \cdot fAuM_p + (1 - \alpha) fAlM_p]}{[\alpha \cdot \sum fAuM_p + (1 - \alpha) \sum fAlM_p]} = \frac{[\alpha \cdot fAuM_q + (1 - \alpha) fAlM_q]}{[\alpha \cdot \sum fAuM_q + (1 - \alpha) \sum fAlM_q]} \quad (53)$$

Átrendezve a fenti egyenlőséget az α -ra egy másodfokú egyenletet kapunk, aminek két megoldása közül azt a pozitív valós gyököt választjuk, ami 0 és 1 közé esik. A helyett, hogy a p és q összes kombinációjának megfelelő, azaz $\binom{K}{2}$ számolást végeznénk, a következő eljárást dolgoztuk ki. A legtávolabb lévő $fAuM_p - fAlM_p$ és $fAuM_q - fAlM_q$ szakaszok (a 27. ábra jobb oldalán lévő grafikonon az (1,5) és (4,8) pontpárok) alapján számoljuk az α -t, majd a teljes mátrixokra (az összes oszlop figyelembe vételével) meghatározzuk az **fAMTR** mátrixot:

$$\alpha \cdot \mathbf{fAuM} + (1 - \alpha) \mathbf{fAlM} = \mathbf{fAMTR} \quad (54)$$

Ha az α -t helyesen számoltuk ki, akkor az **fAMTR** mátrix oszlopai csak egy skála-paraméterben térnek el, azaz egymás skalárszorosai lesznek és a nem-normalizált térben egy nulla tengelymetszetű egyenesre illeszkednek (zöld körökkel meghatározott zöld vonal a 27. ábra jobb oldali grafikonján). Ha nincs egyértelmű megoldás, akkor azt vagy az illeszkedő egyenes hiánya, vagy az α értékének nem 0 és 1 közé esése jelzi. Egy másik kritérium lehet, hogy az **fAMTR** mátrix második sajátértéke nem nulla, azaz Matlab-ban nem kisebb mint 10^{-14} .

A tanulmányunkban elemzett adattömbök alapján a következő megállapításokat tettük. $k\{2,2,2\}$ adattömbök esetén a bemutatott algoritmusunk mindig gyorsan és

megfelelően működött, ugyanazt az egyértelmű megoldást szolgáltatva, mint pl. az iteratív eljárás alapuló PARAFAC. $k\{2,2,1\}$ adattömbök esetén a Kruskal egyenlőtlenség sérül: $2 + 2 + 1 \not\geq 6$, tehát az egyértelmű megoldás nem biztosított. A koncentráció mátrix k -rangja 1 (rangátfedés (rank overlap)), ami a profilok lineáris arányosságának esetleg megegyezésének következtében alakul ki. Ebben az esetben meglepő módon az összes megengedett megoldás-tartományba eső megoldás trilineáris lesz, azaz nem jelent előnyt a trilineáris kényszerfeltétel alkalmazása, nem kapunk egyértelmű megoldásokat. Ez a fajta három-utas adattömb bilineáris adatként kezelendő: minden egyes mátrixszelet vizsgálata ugyanarra az eredményre vezet. Adattömbként kezelés csak a szabotosságot és pontosságot csökkentheti a mérési zaj hatásától függően. Ezeket a megállapításokat általánosíthatjuk $k\{N,N,1\}$ típusú adattömbökre, amikor az utolsó út szerinti profilok megegyeznek, azaz $k_3=1$, N helyett.

A gyakorlatban az azonos profilok kezelésére elfogadott módszerként az U-PLS\RBL (unfolded partial least-squares and residual bilinearization) [51] jelent meg, azonban rangátfedés esetén ez nem alkalmazható. Olivieri et al. [52] bevezették az U-PLS/RBL-LD (U-PLS/RBL with linear dependencies) módosítást. Ugyanazokat a kényszerfeltételeket alkalmazták, mint az MCR-ALS-hez és ezért kapták ugyanazt az eredményt. A kényszerfeltételek (korreláció és megfelelés a speci-eszek között (correlation and corresponding among species), azaz kalibrációs hal-maz szerepeltetése ismert koncentrációval) alkalmazása miatt működtek az algoritmusok és nem a felépítésük miatt!

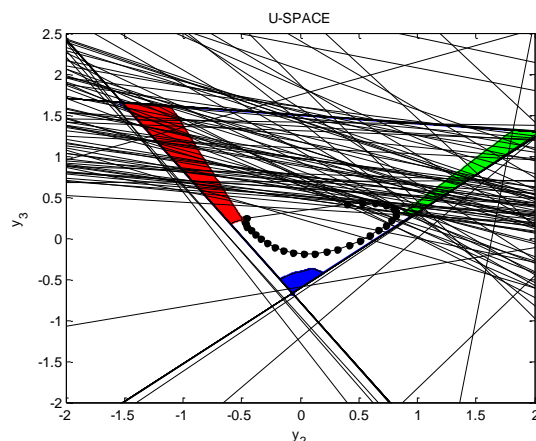
Hangsúlyozni kell tehát, hogy ugyanazon mérési körülmények és kényszerfeltételek alkalmazása mellett a kemometriai módszerek ugyanazt az eredményt fogják szolgáltatni, azonban eltérő mérési körülmények esetén illuzórikus az összehasonlítás.

4.12 Az L_0 , L_1 és L_2 normák vizsgálata ritkás profil-mátrixokra történő nemnegatív többváltozós görbefulbontás során [P12]

A nulla értékek jelenléte a mérési adatokban az analitikai kémia egyes területein természetes jelenség. A kromatográfiás elválasztás során (LC, GC) a jól elválasztott komponensek elúciós görbájén is megjelennek a nulla értékek. A spektrális irányban is kaphatunk nulla értékeket, pl. a tömeg-, ill. ionmobilitás spektrometriában. A napjainkban népszerű hiperspektrális képanalízis (pl. távérzékelés során kapott) adatprofiljai is nullákkal lehetnek tele, pl. mivel a kémiai vegyületek felszíni eloszlása nem egyenletes. A nullákat tartalmazó vektorokat, mátrixokat ritkásnak (sparse) nevezzük. Ebben a tanulmányunkban az SMCR módszer segítségével határoztunk meg ritkás profilokat és kiemeltük, hogy normális adatsor esetén ezek a külső poligonon helyezkednek el, ami az elválasztó határ a csak pozitív és a negatív értékeket is tartalmazó profilok absztrakt pontjai között. A legritkásabb megoldások a külső poligon csúcspontjai lesznek.

Másrészt, ha az adatmátrixunk csak nullákat tartalmazó sorokkal vagy oszlopokkal is rendelkezik (abnormális adatsor, 28. ábra), akkor a megengedett megoldástartomány belső pontjainak megfelelő profilok ugyanannyi nullát fognak tartalmazni (a nullák elhelyezkedése ezekben a profilokban függ a csak nullákat tartalmazó sorok (oszlopok) mátrixon belüli elhelyezkedésétől: egy teljesen nulla sor az adatmátrixban egy nulla elemet eredményez a koncentráció profilokban, és egy teljesen nulla oszlop az adatmátrixban egy nulla elemet eredményez a spektrális profilokban), de még ekkor is a külső poligon pontjainak megfelelő profilok több nullával fognak rendelkezni. A hasznos tartomány (region of interest, ROI [53]) módszerrel pl. nagyon könnyen kaphatunk abnormális adatsorokat.

A csak nullákat tartalmazó sorok (oszlopok) eltolást (offset) eredményeznek és ennek mértéke a csak nullákat tartalmazó sorok (oszlopok) számával lesz egyenlő. A külső poligon pontjaihoz kapcsolódó profiloknak nulla elemekkel kell rendelkezniük és ezek száma az egybevágó duális egyenesektől, a metszéspontjaik számától és az eltolástól függ.



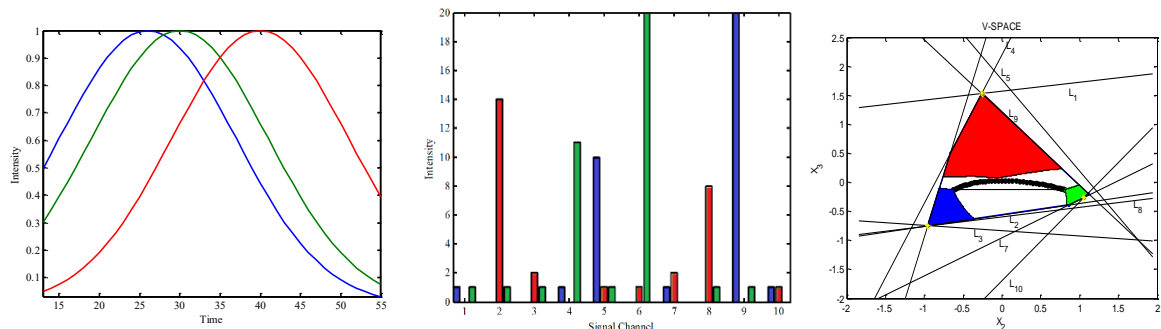
28. ábra Abnormális adatsor Borgen-Rajkó grafikonja. A fekete vonalakat a duális térben lévő adatpontokból határoztuk meg: csak nullákat tartalmazó sorok a belső poligon belsejében olyan absztrakt pontokat eredményeznek, amelyekből transzformált profilok azonos számú nullákat tartalmaznak.

Bár a tanulmányban csak a nemnegatív és ritkás kényszerfeltételek hatását vizsgáltuk, más kényszerfeltételeket is használhatunk mellettük, de a ritkás megoldásoknak mindig a külső poligonon kell elhelyezkedniük.

Az L_p ($p=0,1,2$) normák alkalmazását és hatásait a 29. ábrán bemutatott példa adaton szemléltetjük. A külső poligon oldalai nulla értéket is tartalmazó profilokat jelenítünk meg, míg a csúcspontokhoz tartozó profilok 4 (kék), 3 (piros) és 2 (zöld) nullát tartalmaznak, amelyek a szimulált tömegspektrumokban szereplő értékekkel megegyeznek (29. ábra középső és jobb oldali grafikon).

A megengedett megoldás-tartományok belsejében lévő pontokhoz kapcsolódó profilok nem tartalmaznak nullát, mivel két vektor (profil) konvex lineáris kombinációja csak akkor tartalmaz nullát, ha ugyanazon pozícióban mindkét vektor (profil) nullákat tartalmaz: $s_j^m = \alpha s_j^a + \beta s_j^b$ csak akkor 0, ha s_j^a és s_j^b mindkettő 0 ($0 < \alpha < 1$ és $0 < \beta < 1$).

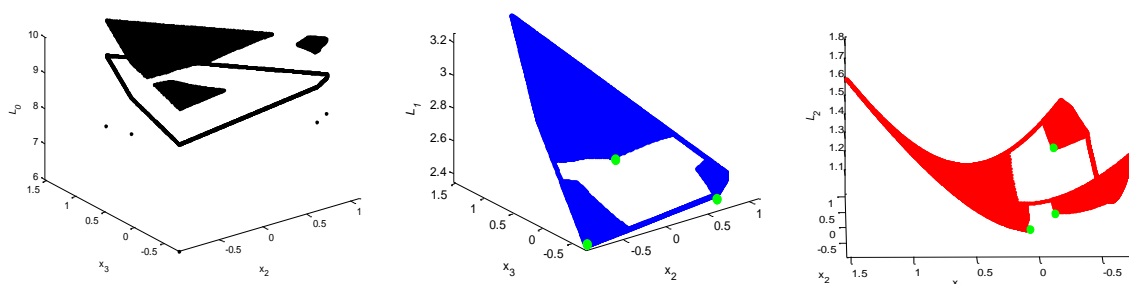
A 30. ábra bal oldali grafikonján láthatjuk, hogy az L_0 norma minimalizálásával helyes eredményt kapunk, azaz a legritkább megoldásokat. A 30. ábra középső grafikonján az L_1 norma minimalizálásával csak két komponens esetén kapunk helyes eredményt a harmadik komponens esetében az L_1 norma maximalizálása vezetne hibátlan eredményre.



29. ábra Háromkomponensű szimulált GC-MS adatsor. Bal oldalon: GC elúciós profilok; középen: MS profilok; jobb oldalon: a megengedett megoldás-tartományok a V-térben, a sárga csillagok a duális vonalak metszéspontjaiban megjelenő „valódi” ritkás MS profilok (középső diagram) absztrakt pontjai.

A 30. ábra jobb oldali grafikonja igazi meglepetést tartalmaz, az L_2 norma minimalizálásával a legrosszabb profil-bebecsléseket kapjuk, a legtávolabb lévőket a ritkás megoldástól. Az L_2 norma minimalizálása helyett maximalizálást kell alkalmazni, ekkor megkapnánk a kívánt ritkás megoldásokat.

A szakirodalomban a ritkás megoldás meghatározására az L_1 norma vagy az L_1 és L_2 normák kombinációjának minimalizálását javasolják (LASSO [54] és elasztikus háló (elastic net) [55]), azonban ez csak egyetlen komponens egyetlen profiljának ritkás meghatározásánál használható. Amennyiben 2, 3 vagy több komponens ritkás profiljait kell egyszerre meghatározni, akkor az L_0 norma minimalizálását vagy az L_2 norma maximalizálását és az L_1 norma alkalmazásának mellőzését javasoljuk.



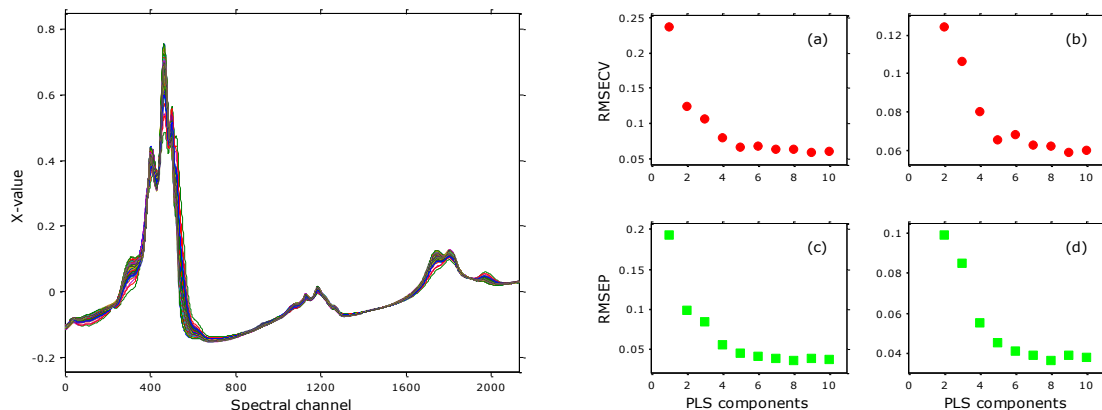
30. ábra Az L_p ($p=0,1,2$) normák alkalmazásának összefoglaló ábrái (megengedett megoldás-tartományok). Bal oldalon: L_0 norma minimalizálásával kapott értékek; középen: L_1 norma minimalizálásával kapott értékek; jobb oldalon: L_2 norma minimalizálásával kapott értékek.

5. Mérési adatok kiértékelése: gyakorlati fejlesztések és alkalmazások

5.1 Alternatív érvényesítési eljárás a túlillesztés elkerülésére [P13]

Ebben a tanulmányunkban kritikusan áttekintettük a többváltozós kalibrációnál jelentkező túlillesztés problémáját és a hagyományos érvényesítés (validation) alapú eljárást, ami a túlillesztés elkerülését célozza. Kifejlesztettünk egy alternatív eljárást, ami egy randomizációs-permutációs teszten alapul; ez lehetővé teszi, hogy minden egyes komponenshez amit a modellbe illeszteni kívánunk statisztikus szignifikancia szintet rendeljünk. Eljárásunkat összehasonlítottuk az elterjedt kereszt-érvényesítés és független tesztalapon alapuló eljárásokkal NIR kalibrációs adatokra PLS regressziót alkalmazva. Az eredmények azt mutatják, hogy az alternatív megközelítésünk objektívebb, mivel az érvényesítés alapú megközelítéssel ellentétben nem igényli a „puha” (soft) döntési szabályok alkalmazását. Az alternatív megközelítés tehát hasznos kiegészítésnek tűnik a kemométerek eszköztáraiban. A vizsgált adatsor F. Wahl (Institut Français du Pétrole) személyes adatközlése révén állt rendelkezésünkre. 239 gázolaj mintát vizsgáltak, amelyek NIR spektrumait (X) 4900–9000 cm^{-1} hullámszám-tartományban vették fel (31. ábra bal oldala). A függő változó (Y) a hidrogén tartalom volt. A referencia értékeket NMR méréssel határozták meg, a mérési hiba becslésére a tapasztalati standard eltérés (empirical standard deviation): $\sigma_{ref} = 0,025\text{g}/100\text{g}$.

A duplex algoritmus [56] segítségével a 239 mintát kalibrációs és érvényesítő adatkészletekre osztottuk: kalibrációhoz 84 minta, érvényesítéshez 155 minta jutott. A 31. ábra jobb oldalán bemutatott felső panelek alapján az RMSECV (root mean squared error of cross-validation) lokális minimuma 5 komponensű modellt, míg az alsó panelek alapján a külső érvényesítéssel kapott RMSEP (root mean squared error of prediction) lokális minimuma 8 komponensű modellt sugall.

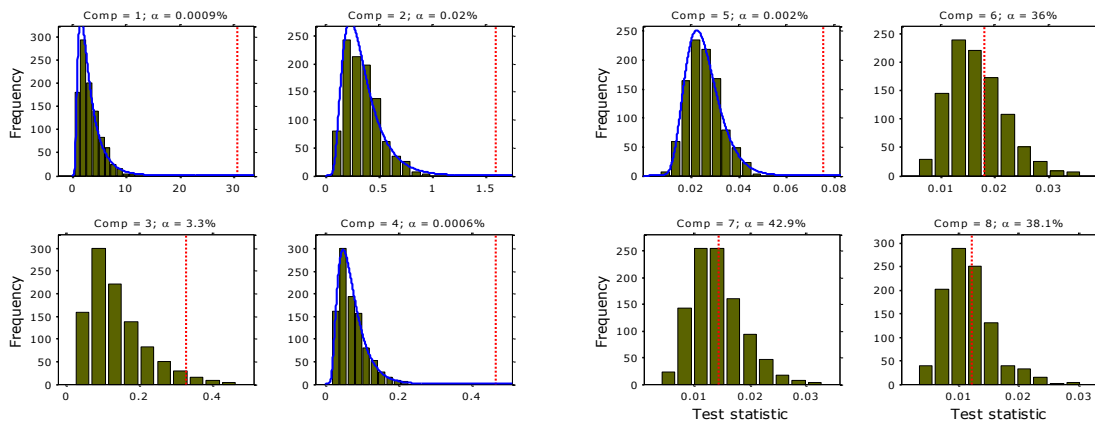


31. ábra Bal oldal: A vizsgált minták NIR spektrumai; jobb oldalon felső két diagram (piros körök): belső RMSECV a 84 kalibrációs mintán, jobb oldalon alsó két diagram (zöld négyzetek): külső RMSEP a 155 kalibrációtól független mintán (*b*, és *d*, panel: első adatpontok nélkül)

Alternatív módszerünkkel az 1-8 komponensekre generált hisztogramokat a 32. ábrán mutatjuk be, ahol a számított szignifikancia értékeket százalékban tüntettük fel. Megfigyelhetjük, hogy annak a valószínűsége, hogy a tesztstatisztika értékeket véletlen okozta rendkívül csekély az 1. (0,0009%), a 2. (0,02%), a 4. (0,0006%) és az 5. (0,002%) komponensek esetében. Érdekes azonban, hogy a 3. komponens szignifikanciája már 3,3%, ami annak tudható be, hogy ez a komponens érzékenyebb a spektrumokban lévő linearitástól eltérésre, mint a többi. A 6-8. komponensek magas szignifikancia értékei egyértelműen jelzik a túlillesztést az 5. komponens után. A következő megállapításokat tehetjük:

- alternatív eljárásunk enyhe feltételezések mellett alkalmazható,
- felhasználóbarát, csak a permutációk számát és a szignifikancia küszöbszintet kell megadni,
- az eredmények gyakran azonosak az érvényesítés alapúakkal (pl. Unscrambler vagy SIMCA), de most azokat teljesen objektív alapon kaptuk,
- helyettesítheti az érvényesítést a komponensek kiválasztásához, de kiegészítheti a szokásos diagramot (RMSEP becslések vs. komponensek) is,
- a randomizációs tesztet nem csak PLS regresszióra alkalmazhatjuk, hanem többutas kalibrációhoz is,

- tömörítésre lehet szükség nagy adathalmazok (big data) esetén, de arra figyelni kell, hogy a tömörítés ne okozzon függést a minták között,
- a bemutatott randomizációs teszt a kalibrációs adatkészleten működött, hogy a kereszt-érvényesítés objektivitását biztosítsuk, azonban semmi akadálya, hogy ezt a külső adatkészlettel érvényesítésre is alkalmazzuk.



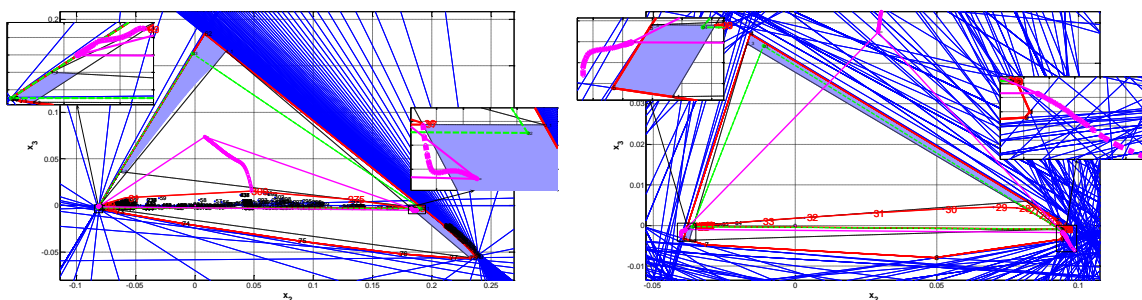
32. ábra Randomizációs eljárásunkkal kapott 1000 permutált adat alapján készített hisztogramsor. Az illesztéshez inverz Gauss függvényt használtunk (—), a teststatisztika értékek feltüntetése (---) mellett

5.2 Az MCR-ALS algoritmus néhány meglepő gyakorlati tulajdonságáról [P14]

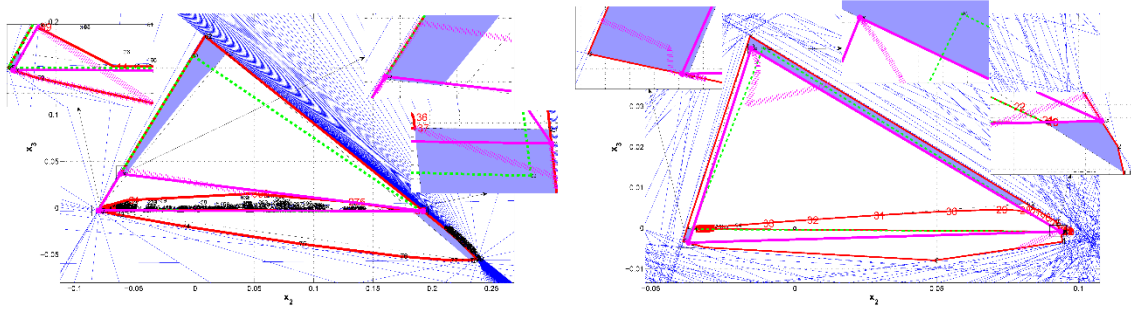
Ebben a tanulmányban az MCR-ALS algoritmus egy meglepő problémáját tártam fel: az iterációk során kapott rész megoldások és még a végső megoldás is az **A** adatmátrix nullterében lesznek. A Borgen grafikonon jól illusztrálható az iterációs nyomvonal és az, hogy a nulltérben lévő nemnegatív (rész)megoldás az **A** mátrix sor-, ill. oszlopterébe vetítve már negatív elemeket eredményez a profilokban (33. ábra).

A probléma orvoslásához egy új kényszerfeltételt vezettem be: becsült profilok zéró ortogonális résszel (zero orthogonal part for the estimated profiles). Az MCR-ALS algoritmus Matlab implementációjának fő függvényét (*als2004m.m*) [57,58] módosítottam úgy, hogy a profilok részbecslésénél csak az **A** mátrix sor-, ill. oszlopterébe vetített részét vesszük figyelembe. Sajnos ez a vetítés is tartalmazhat negatív elemeket, ha a külső poligonon kívülre kerül. Ezért egy ügyes algoritmus részlet módosította a negatív elemeket is tartalmazó részbecslést a vetítés és az origo közötti szakasz és a külső poligon metszéspontjára (34. ábra). Ráadásul az *fcnnls*-t [59] használtam a nemnegatív kényszerfeltétel alkalmazásához, ami jelentékeny konvergencia gyorsítást eredményezett.

A Borgen grafikon jelen állapotban is alkalmazható érvényesítő módszerként a görbefelbontó algoritmusok rejtett tulajdonságainak feltárására grafikus, vizuálisan könnyen értelmezhető produktuma miatt. Ezzel a tanulmánnyal is előrébb jutottunk a kemometria geometriájának (geometry of chemometry) feltárásában.



33. ábra Az eredeti MCR-ALS algoritmus a nulltérben konvergál, így a vetítések olyan profilokat eredményeznek, amelyek zérus értékeket tartalmaznak



34. ábra Az új kényszerfeltétellel és a külső poligonon belülré kényszerítés korrekcióval az MCR-ALS algoritmus már nem a nulltérben konvergál

5.3 Kritikai megjegyzések NIR hiperspektrális adatok kiértékelése során alkalmazott minimális térfogat kritériumra vonatkozóan [P15]

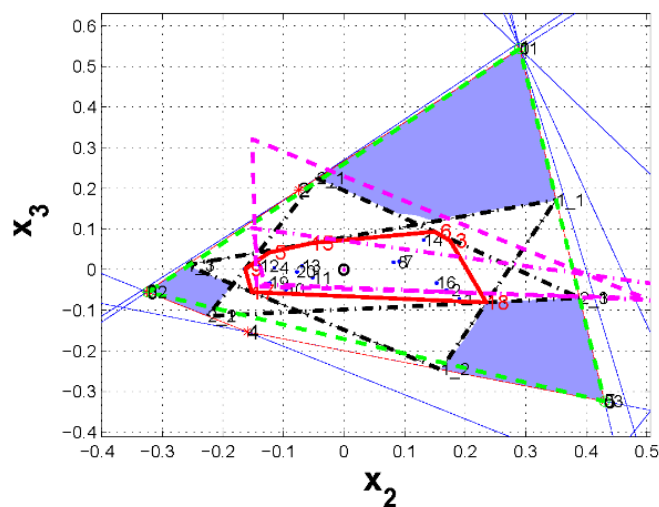
Egy megjegyzés közleményt jelentettem meg Lopes et al. [60] által az *Analytical Chemistry*-ben közölt tanulmánnyal kapcsolatban. Hamisított gyógyszer-tabletták kemometriai jellemzéséről számoltak be NIR hiperspektrális adatok SISAL (simplex identification via split augmented Lagrangian), MVSA (minimum volume simplex analysis), és MVES (minimum-volume enclosing simplex) módszerekkel történt felbontása alapján. Elsőként azt jegyeztem meg, hogy a tiszta pixel nélküli hiperspektrumok esetére Craig [61] földtudományok területén megjelent munkáját hivatkozták, mivel ő az absztrakt térben az adatpontokat tartalmazó minimális térfogatú szimplex csúcspontjait javasolta becslésként. Azonban Perczel et al. [62,63] már Craig előtt javasolták ugyanazon elven alapuló CCA (convex constraint analysis) módszerüket kémiai szaklapban megjelent cikkeikben. Valamint Henry [64] is foglalkozott a többváltozós receptor modellek kapcsán az n -dimenziós éldetektálással, amit szintén nem vettek figyelembe. Másodsorban azon kijelentésüket kritizáltam, amiben azt állították, hogy „Az MCR-ALS döntő hátránya az ún. forgatási bizonytalanság, azaz különböző orientációjú szimplexek, amelyek tartalmazzák az adatpontokat a legkisebb négyzetek kritériumát egyaránt minimalizálják.” Megjegyeztem, hogy a forgatási bizonytalanság nem az MCR-ALS vagy más algoritmus tulajdonsága, hanem a bilineáris adatmátrix belső sajátja!

Lopes et al. [60] a nemnegativitást alkalmazták kényszerfeltételként, valamint egy szükséges feltételt: „a szimplex minden $(p-1)$ dimenziós oldala, azaz lapja (facet) tartalmazzon $p-1$ spektrális adatvektort”. Ez kemometriailag azt jelenti, hogy minden komponensre vonatkozóan léteznie kell legalább $p-1$ spektrális sávnak (hullámhossz, hullámszám stb.), ahol csak ez egyik komponensnek lesz nulla az elnyelése. Ez olyan erős megkötés, ami csak ritkán fordul elő a gyakorlatban. Másrésről az MCR-ALS annyira rugalmasan alakítható, hogy könnyen implementálhatjuk a minimális térfogatú szimplex kritériumot és a szigorú feltételt.

Továbbiakban illusztráltam, hogy a koncepció amit használtak nem minden esetben ad elfogadható eredményt. Henry és Kim [32] adatait felhasználva bemutattam, hogy nemnegatív kényszerfeltétel alkalmazása mellett a minimális térfogatú szimplex (három komponensű rendszerek esetén minimális területű háromszög) a duális térben negatív értékeket is tartalmazó profilkok absztrakt pontjait eredményezheti, mert ezek kívül esnek a külső poligonon. A 35. ábrán a pontozott és a sima szaggatott vonalakkal jelzett háromszögek csúcspontjai a SISAL és MVSA módszerek megoldásai. A csúcspontok közül egy vagy kettő a külső poligonon kívülre esik, így a nemnegativitás feltétele sérül. Ráadásul egyetlen csúcspont sem esik egyetlen megengedett megoldás-tartományba sem!

Megállapíthatjuk tehát, hogy nem létezik olyan általános algoritmus, ami a legjobb (optimális) az összes adatkészletre.

Azután az „egyértelmű” és az „egyetlen” megoldás fogalmait jártam körül. A programozók és matematikusok olyan algoritmusokat fejlesztenek, amik egyetlen megoldást eredményeznek, de ez még nem biztos, hogy egyértelmű is lesz, hiszen létezhet hasonló optimális tulajdonsággal bíró másik megoldás is. A megengedett megoldás-tartományok ezeket gyűjtik össze. A PCA „egyértelmű” megoldása az alkalmazott matematikai kényszerfeltételeknek köszönhető, többek között az ortogonalitásnak.



35. ábra Henry és Kim adatának Borgen grafikonja a SISAL és MVSA megoldásokkal

Azonban a nemnegatív profilok szinte sohasem ortogonálisak (a zérus értékek speciális helyzete szükséges ehhez: ahol az egyik profil nem nulla értékű, ott a másik profil nullát kell tartalmazzon), így a PCA absztrakt megoldása csak egy matematikai műtermékként kezelendő a legtöbb esetben. Kémiaileg/fizikailag értelmezhető profilmegoldások csak további kémia/fizikai jelentéssel bíró kényszerfeltétel (pl. nemnegativitás, unimodalitás stb.) alkalmazása mellett kapható.

Egy további kritikus probléma a normalizálás. A természetben az adatok általában nem normalizáltak. A normalizálásra az adatok egyszerűbb értelmezése és/vagy az alkalmazott algoritmus működése miatt van szükség. Ezért a megjegyzésem tárgyakként vett cikkben az 1. és 2. egyenletet a következőképpen kellett volna értelmezni: az \mathbf{M} mátrix a lineáris Napier-féle abszorpciós koefficiens mátrix [65], így az \mathbf{A} mátrix tartalmazza a behatolási mélységet, ami a koncentrációval arányos. Azonban ez a mátrix nem természetes módon normalizált a Ref. [60] 2. egyenlete szerint, a behatolási mélység nem kapcsolható közvetlenül a 0 és 1 közé eső összetétel arányhoz. Ezért az 1-norma feltételezése önkényes ebben az esetben. Az \mathbf{Y} adatmátrixot értelemszerűen normalizálhatjuk: $\bar{\mathbf{Y}}_C = \mathbf{Y} ./ \left(\mathbf{1} \mathbf{1}^T \mathbf{Y} \right) = \bar{\mathbf{M}} \bar{\mathbf{W}}$, ahol

$$\bar{\mathbf{M}} = \mathbf{M} ./ \left(\mathbf{1} \mathbf{1}^T \mathbf{M} \right), \quad \bar{\mathbf{W}} = \mathbf{A}^* ./ \left(\mathbf{1} \mathbf{1}^T \mathbf{A}^* \right), \quad \text{és} \quad \mathbf{A}^* = \mathbf{A} .* \left(\mathbf{M}^T \mathbf{1} \mathbf{1}^T \right).$$

A $./$ and $.*$ a Matlab-ban szokásos mátrixokra vonatkozó elemenkénti osztás, ill. szorzás művelete. Természetesen $\bar{\mathbf{W}}^T \mathbf{1} = \mathbf{1}$ lesz, de az így kapott normalizált összetétel mátrix az \mathbf{M} spektrális mátrixtól is függeni fog.

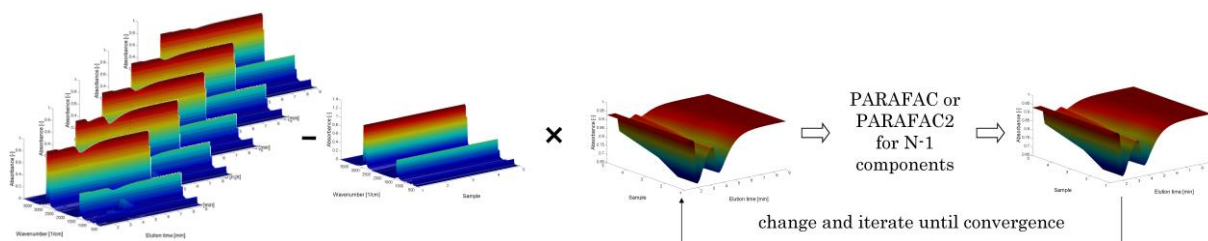
5.4 HPLC/FT-IR mérések során az eluens-hatás elimináció kemometriai megvalósítása [P16]

A kutatásaink célja jelen esetben az volt, hogy HPLC/FT-IR mérések során az átfolyós folyadékcellával megvalósított IR detektálás előnyeit (gyorsaság, aránylag egyszerű kivitelezés, sokféle eluens alkalmazhatósága) kihasználva a fizikai oldószereliminációt és annak hátrányos tulajdonságait (csak illékony eluens alkalmazható, az elpárologtatás során meghatározandó komponens is távozhat, bonyolult kivitelezés) [66,67] kemometriai módszerre cseréljük és egy megfelelő algoritmus segítségével vonjuk ki az eluens hatását.

Ehhez megvizsgáltuk a PARAFAC (PARAllel FACtor analysis, párhuzamos faktor-elemzés), PARAFAC2 és az SMCR kemometriai módszereket az ALDRICH elektronikus adatbázisában található n-hexán, 1-klórhexán, 1,3-dioxolán és 2-metoxi-1,3-dioxolán spektrumok (referenciaspektrumok) felhasználásával.

A PARAFAC módszer nem alkalmazható, mivel nem teljesül az alkalmazásának egyik feltétele, nevezetesen, hogy a profiloknak mintáról mintára meg kell egyezniük, mivel az eluens elúciós profilja természetesen más és más lesz a különböző összetételű minták esetén. A PARAFAC2 módszerrel a spektrumok kinyerése lehetséges volt legalább 95%-os átfedéssel, de a koncentrációk és az elúciós profilok meghatározása nem volt kielégítő.

Kifejlesztettünk egy új iterációs eljárást, az OSSS-IU-PARAFAC-ot (Objective Subtraction of Solvent Spectrum with Iterative Use of PARAFAC), amellyel mind a spektrumok, mind az elúciós profilok és a koncentrációk is 100%-os egyezéssel kinyerhetők voltak. Sajnos kiderült, hogy ez a jó egyezés, ill. magának az algoritmusnak a működése a használt adatmodell speciális voltának köszönhető, nevezetesen a legegyszerűbb elkülönülő háromszög alakú elúciós profiloknak. Ha már zajt kevertünk hozzá, vagy más a valósághoz közelebbi elúciós profil alakot választottunk, akkor az eljárás már nem konvergált. Zajmentes, de a valóságot jobban közelítő elúciós profilok alkalmazása esetén a PARAFAC2 módszert kellett iteratívan alkalmazni a kielégítő kvalitatív és kvantitatív eredmények eléréséhez. Az OSSS-IU-PARAFAC algoritmusunk képi megjelenítése:



A mért adatoknak nemcsak a realiztikus kromatográfiás csúcsai, hanem az eluens-eluátum kölcsönhatásából származó spektrális módosulásai is hozzájárultak ahhoz, hogy a szimulált adatsorokra megalkotott algoritmusunk változtatás nélküli alkalmazása eredménytelen lett. A valós adatsoron több elő és utó adatfeldolgozó eljárásokat kellett alkalmaznunk. Összehasonlítva a szimulált adathalmazban lévő eluens spektrumok hányadát ($\approx 20\%$) a mért adathalmazban lévővel ($\approx 70\%$), valószínűsíthető, hogy a kemometriai felbontásra szánt kiindulási adattömb által reprezentált eluens spektrumhányad is befolyásolja a számítás végeredményét. Számításaink további szakaszában ezért sort kerítettünk a mért adattömb értelem-szerű megcsonkítására, vagyis kihagytuk azokat az időszakaszokat a hozzájuk tartozó spektrummal együtt, amelyekhez csak eluens spektrum rendelhető (konkrétan: az első kromatográfiás csúcs előtti és az utolsó kromatográfiás csúcs utáni spektrumsorozatokat).

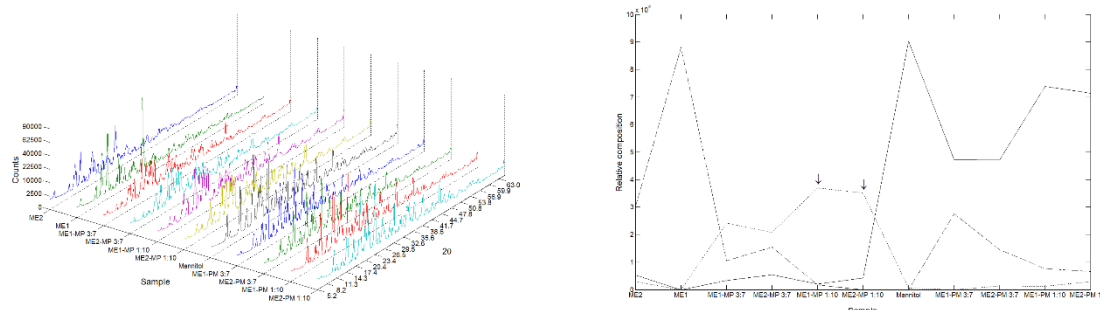
A megmaradt adathalmaz zajának csökkentése érdekében - a spektrumok hullámszámfüggő zajának elemzése után - csak a legkevésbé zajos és egyben az azonosítási szempontból legértékesebb tartományt, az $1800\text{-}650\text{ cm}^{-1}$ közötti részeket használtuk fel; illetve e tartományon belül a spektrumok 1-nél nagyobb abszorban-ciájú, Lambert-Beer törvényt nem követő részeit (vagyis az $1490\text{-}1425\text{ cm}^{-1}$, $1392\text{-}1365\text{ cm}^{-1}$ és $739\text{-}714\text{ cm}^{-1}$ tartományokat, ahol egyébként az eluens erős elnyelési sávjai találhatóak) kihagytuk. A redukált adattömbbel végrehajtott kétlépéses OSSS-IU-PARAFAC2 eljárás az elúciós és koncentrációprofilokat már jól visszaadta, de a kívánt spektrumokat csak egy újabb szubjektív elimináció végrehajtása után kaptuk meg. Bizonyos, hogy az általunk kifejlesztett eljárás nagymértékben függ a kezdőértékként megadott eluens elúciós profiljától, így az SMCR módszerek vizsgálatába kezdtünk.

5.5 A meloxikám-mannit binér rendszer fiziko-kémiai jellemzése [P17]

Vizsgáltuk a meloxikám (ME) és a mannit (MA) mint hordozóanyag fizikai és olvadék keverékének (PM - physical mixture, ill. MP - melted product) kioldódási viselkedését. Kétféle részecskeméretű meloxikám mintát használtunk: az ME1-gyel jelölt kb. 40-szer nagyobb méretű volt mint az ME2-vel jelölt minta amit őrléssel mikronizáltak:

| Minták | d (10%) (µm) | d (90%) (µm) | Fajlagos felület (m ² /g) |
|--------|--------------|--------------|--------------------------------------|
| ME1 | 50,50 | 206,2 | 0,071 |
| ME2 | 0,72 | 5,97 | 2,514 |
| Mannit | 17,61 | 239,45 | 0,226 |

A kemometriai feladat ennél a problémánál az volt, hogy megválasszunk azt a kérdést miszerint a röntgendiffrakciós mérések alapján értelmezni tudjuk-e az olvadékforma legelőnyösebb kioldódási viselkedését. MCR-ALS módszert alkalmaztunk és az eredményekből arra következtettünk, hogy az olvadékban egy új forma alakult ki (blend), ami 1:10 arány mellett nagyobb mértékben van jelen, mint 3:7 arány esetén. Megfigyeltük, hogy a fizikai keverékben elhanyagolható, míg a cseppentő módszerrel kapott mintában a legnagyobb mértékben található az új alakulat (36. ábra). A legjobb kioldódást a cseppentő módszerrel kapott termék mutatta.

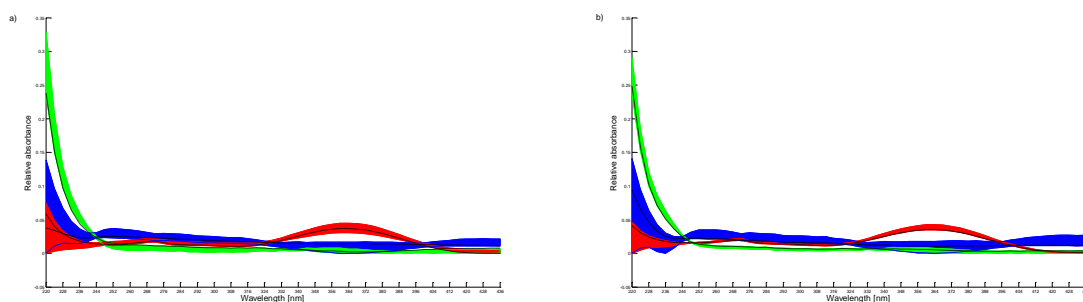


36. ábra Bal oldalon: ME-MA binér rendszerek röntgendiffrakciós diagramjai; jobb oldalon: az MCR-ALS által kapott három kristályalakulat összetétel profilja, folytonos vonal – MA, szaggatott vonal - ME, pontozott vonal - 'blend'

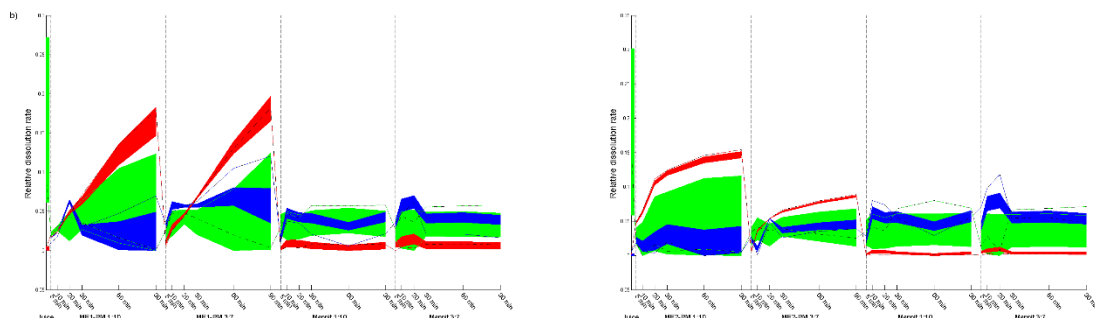
5.6 Görbeillesztés nélküli komponensprofil-kinyerő módszer alkalmazása meloxicám-mannit binér rendszerek kioldódási vizsgálata során [P18]

Kioldódási vizsgálatok többváltozós kemometriai módszerekkel történő kiértékelése jelent meg az Analytical Chemistry-ben 2006-ban. Úgy találtuk, hogy Wiberg és Hultin [68] által használt kemometriai eljárások nem alkalmazhatók az általunk vizsgált meloxicám-mannit fizikai keverék rendszer (ME-MA PM) kioldódási eredményeinek értékelésére (a minták és így a részecskeméret adatok is megegyeznek az előző szakaszban bemutatott táblázatban lévőkkel). Többek között azért nem, mert az eltérő részecskeméretű meloxicámnak eltérő kioldódási görbét mértünk (itt arra kell figyelni, hogy az eltérő méretű meloxicám UV-Vis spektruma azonos ugyan, de a kioldódási görbék eltérőek), ezért pl. a PARAFAC nem alkalmazható (a profil nem azonos az összes mintára). Sajnos a PARAFAC2 alkalmazása sem adott kielégítő eredményeket. Ezért az adatsort kettébontottuk a kétféle részecskeméretnek megfelelően és az általunk kifejlesztett (és korábban részletezett) SMCR módszert megvalósító Matlab programot használtuk. Megállapítottuk, hogy az ME2 jelzéssel ellátott részecskeméretű meloxicám kioldódása kétszer jobb, mint a többi mintáé.

A spektrális sávmegoldások az ME1 és az ME2 adatkészletekre nagyon hasonlóak lettek, igazolva eljárásunk következetességét (37. ábra).

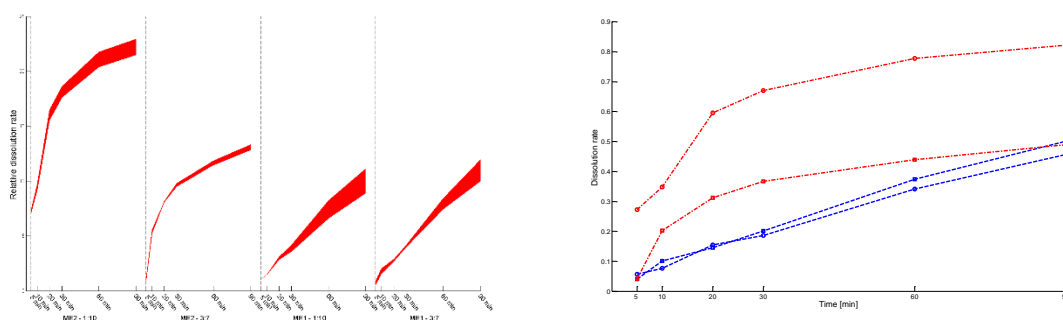


37. ábra Spektrális sávmegoldások. Bal oldalon: ME1, jobb oldalon: ME2



38. ábra Kioldódási-görbe sávmegoldások. Bal oldalon: ME1; jobb oldalon: ME2

A 38. ábra az SMCR alapján kapott kioldódási-görbe sávmegoldásokat mutatja be. Ezekből kigyűjtöttük a meloxikámra vonatkozó sávokat a 39. ábra bal oldalán. Az 1:10 arányú ME2-PM kioldódása sokkal jobb, mint a 3:7 arányú ME2-PM-é és az ME1-PM mindkét arányú változatánál, ahogy azt korábban megállapítottuk [P17]. Amennyiben az akkori gyakorlatnak megfelelően egyetlen karakterisztikus hullámhosszon (364 nm) vizsgálódunk, akkor a 39. ábra jobb oldalán látható görbéket kapjuk, ami egybevág az eddigi megállapításainkkal. A bemutatott módszerünk azonban nemcsak a kioldódási-görbe sávokat, de a spektrális sávokat is szolgáltatja, valamint látható, hogy a 39. ábra jobb oldalán a két ME1-hez tartozó kék görbe egyike az ME2-höz tartozó alsó piros görbe felé fut be, de ez a sávmegoldásokról nem következik. Tehát a mannit+kapszula elnyelése 364 nm-en nem hanyagolható el (vö. 37. ábrával), amit a gyakorlatban tévesen mégis megtesznek.



39. ábra ME1 és ME2 kioldódási görbék. Bal oldalon: kigyűjtött sávmegoldások; jobb oldalon: 364 nm-es karakterisztikus hullámhosszon mért adatok alapján

Legjobb tudomásunk szerint ez az első beszámoló az SMCR módszerek alkalmazásáról háromkomponensű kioldódási adatokra. A sávmegoldások természetesen csak kvalitatív képet tudnak adni, de a legtöbb esetben a sávok elég szűkek ahhoz, hogy a mérési hibán belül kvantitatív következtetéseket is levonhassunk.

5.7 A lokális rang-vesztés értelmezése és vizsgálata

[P19]

A lokális rang információ kulcsszerepet játszik a többkomponensű kémiai rendszerek kapcsán felmerülő görbefelbontásnál. Lokális rang kényszerfeltételt alkalmazva a forgatási bizonytalanság mértéke csökkenthető és bizonyos esetekben meg is szüntethető minek következtében egyértelmű felbontást kapunk [19]. A szakirodalomban szokásosan azt feltételezik, hogy a lokális rang vagy egyenlő, vagy nagyobb (zaj hatása miatt) mint a kémiai komponensek száma a vizsgált (soronkénti vagy oszloponkénti) mátrixrészletekben. Például megszokott az, hogy csak egyetlen kémiai komponenst tételezünk fel, ha a lokális rang is 1. Tegyük fel, hogy feltételezésünk igaz: a lokális rangfeltáró módszerek, pl. EFA [70] vagy más EFA-alapú módszerek [71-73] által feltárt lokális rang-térkép alapján olyan lényegi tulajdonságokat tudunk megadni, mint pl. az egyes komponensekhez tartozó szelektív tartomány(ok) vagy zérus tartomány(ok). Végző soron azt mondhatnánk, hogy az egyes komponensek egyedi jelenléti ablaka (pl. kromatográfiás adatok esetén az elúciós ablakok) a lokális rang-információ alapján megadható. Mi történik a lokális rang-információn alapuló MCR-módszerek eredményeivel, ha ezek a feltételezések helytelenek? Az analitikai kémiában MCR-módszereket alkalmazó kutatók számára érdekes lenne tudni, hogy ezek a feltételezések mely esetekben korrektek és mely esetekben kerülendőek. Ebben a közleményünkben részletesen tanulmányoztuk a lokális rang kényszerfeltételt, és bilineáris kémiai adatkészletek vizsgálatával kerestük a lokális rang információ és a komponensek jelenléti ablakai közötti megfelelést.

Látni fogjuk, hogy a lokális rang olyan matematikai fogalom, amely esetleg nincs teljes összhangban a kémiai komponensek számával. Így a forgatási bizonytalanság csökkentése érdekében a lokális rang kényszerfeltétel alkalmazása az MCR módszerekben helytelen megoldásokhoz vezethet! Ez a probléma az általunk bevezetett szakkifejezés szerint a „lokális rang-vesztés” (local rank deficiency) következménye.

Az előre és visszairányuló (forward and backward) EFA alapján a példának választott adatmátrixunkra a következő jelenléti ablakmátrixot kaptuk:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | sorszám |
|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|--------------|------------|
| 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | lokális rang | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | I komp. |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | II komp. |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | III. komp. |

Az első 8 retenciós időből álló ablakban a spektrumok lokális rangja 2 (az első két és az utolsó két lokális rang nem feltétlenül helyes: a dőlt és aláhúzott 0 értékeket a korábban NaN (Not a Number) határozatlan reprezentációról helyettesítette az EFA algoritmus). A következő 14 retenciós időből álló ablakban a lokális rang 3, a 23. és 28. közötti retenciós idő-ablakban ismét 2 a lokális rang, végül az utolsó 6 retenciós időből álló ablakban a lokális rang 1.

A 40. ábra bal oldali Borgen grafikonján csillagokkal jelölt megoldáspontoknak megfelelő duális elúciós megoldásgörbék a 41. ábra bal oldalán láthatók. A kinagyított grafikonrészlet jól mutatja, hogy a 29. ponttól csak a piros komponens van jelen, tehát a lokális rang egyenlő a komponensek számával, a fentebb részletezeteknek megfelelően. A 40. ábra középső Borgen grafikonján újabb megengedett megoldáspontokat mutatunk be. Azonban az ezeknek megfelelő duális elúciós megoldásgörbéken (41. ábra középső grafikonja) a 29. ponttól két komponenst (a pirosat és a kéket) láthatunk. Végül a 40. ábra jobb oldali Borgen grafikonján további megengedett megoldáspontokat tüntettünk fel. Ezeknek megfelelő duális elúciós megoldásgörbék (41. ábra jobb oldali grafikonja) alapján a 29. ponttól mindhárom komponens jelenléte megerősítést nyer.

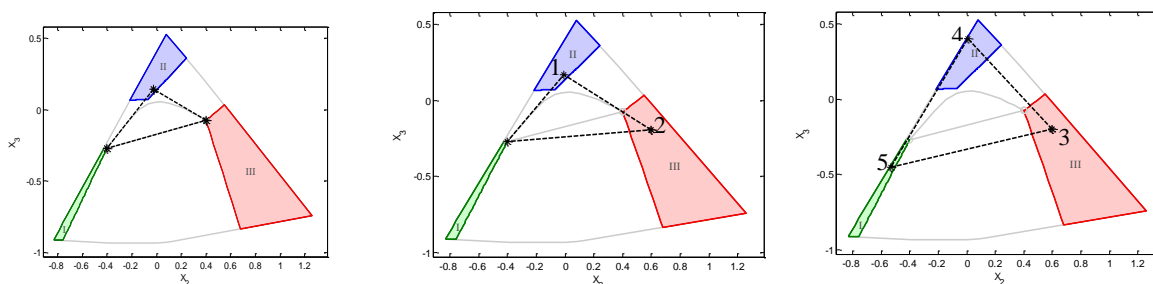
A bemutatott példát is felhasználva, három komponensű rendszerekre a lehetséges lokális rang előfordulásokat a következőkben foglalhatjuk össze:

1. ha az egy lokális rangú ablak a duális térben a belső és külső polinom közös csúcspontján lévő pontot határoz meg, akkor az a primer térben egy szelektív ablak lesz,

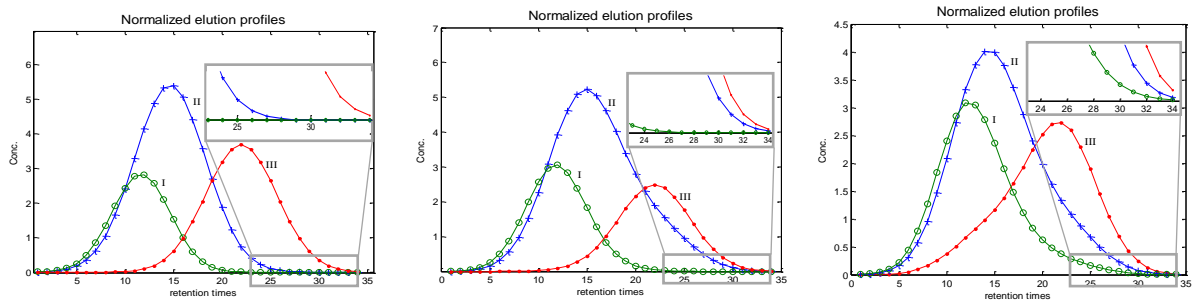
2. ha az egy lokális rangú ablak a duális térben a belső és külső polinomok oldalainak metszetén átmenő egyenesen található, akkor a primer térben két komponens lineáris konvex kombinációja lesz,
3. ha az egy lokális rangú ablak a duális térben a belső poligon belső pontja lesz, akkor három komponens lineáris konvex kombinációja lesz,
4. egyéb esetekben az egy lokális rangú ablak a fenti lehetőségek kombinációjaként adódik,
5. ha a két lokális rangú ablak a duális térben a belső és külső polinom közös csúcspontján lévő pontot határoz meg, akkor a primer térben tényleg két komponens fog megjelenni,
6. ha a két lokális rangú ablak a belső poligon belső pontja lesz, akkor három komponens lineáris konvex kombinációja lesz,
7. egyéb esetekben a két lokális rangú ablak kettő vagy három komponens is tartalmazhat.

A fenti esetek egyike sem eredményezheti azonban a forgatási bizonytalanság csökkenését, hiszen a megengedett megoldás-tartományok nem változnak!

Ha a rang-vesztés csak a választott ablakban fordul elő és nem a teljes mátrixban, akkor pusztán a lokális rang alapján csökkenteni a megengedett megoldás-tartományt kockázatosnak tűnik, hacsak nincs megbízható kémiai információk az adott ablakra vonatkozóan.



40. ábra Borgen grafikonok három különböző megoldással



41. ábra A 40. ábrán mutatott különböző megoldásokhoz tartozó elúciós görbe profilok

Tehát ebben a munkánkban igazoltuk, hogy az adatmátrixok mikrostruktúrája, amit a Borgen grafikon tár fel, alapvető a lokális rang értelmezése terén. Természetes folytatása ez korábbi eredményünknek, miszerint Manne görbefelbontási tételai bár szükségesek, de általában nem elégségesek, így az adat-alapú egyértelműséget (data-based uniqueness) kell használnunk [P8].

5.8 Dualitás alapú általánosított rang-eltüntető algoritmus koncentráció meghatározáshoz, realisztikus hiba-becsléssel [P20]

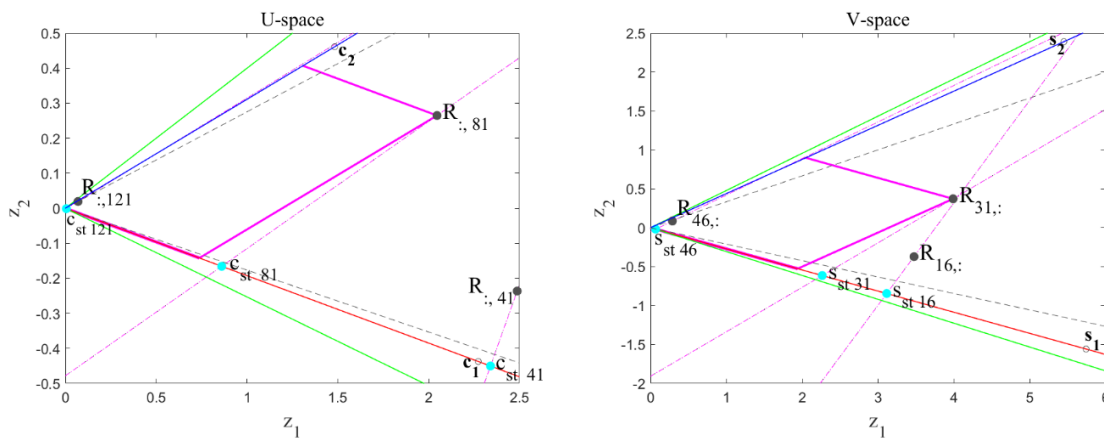
Az analitikai kémia egyik leggyakrabban felmerülő problémája a komplex kémiai minták elemzése úgy, hogy a (zavaró) komponensek többségét nem veszik figyelembe a kalibrációs modellalkotás során. Ezekben az esetekben kívánatos, hogy mennyiségi információt tudjunk szereznünk egy kiemelt komponensről (analit) anélkül, hogy a minta többi összetevőjével foglalkoznánk. Azt a tulajdonságot, hogy ismeretlen összetevők jelenlétében is meg tudjuk határozni az analit mennyiségét, másodrendű előnynek (second-order advantage) [22] nevezzük. A másodrendű előny eléréséhez olyan mérés technikára van szükség, amely másodrendű bilineáris adatmátrixot képes előállítani, például nagy teljesítményű folyadékkromatográfia (HPLC) összekapcsolása UV-látható diódasorral (UV-Vis DAD), azaz kapcsolt technikák (hyphenated techniques) alkalmazására. A másodrendű kalibrációs adatok meghatározó tulajdonsága a trilinearitás, így a kalibrációs mintákhoz rendelt adatmátrixok csoportja alapvetően háromutas adattömbbe rendezhető. A háromutas adatok elemzésére többféle módszer létezik, amelyek többsége a másodrendű előnyök felhasználásával oldja meg a kalibrációs feladatot [74-76]. Ezen módszerek közül a rang-eltüntető faktorelemzés (Rank Annihilation Factor Analysis, RAFA) két bilineáris adatmátrixszal dolgozik, az első a kalibrációs mintához, \mathbf{R}_s , a második az ismeretlen mintához, \mathbf{R} , tartozik. Ennél a módszernél az adott komponens (analit) hozzájárulását kivonjuk a mért spektrumok mátrixából, így a maradékmátrix \mathbf{E} rangja eggyel csökken:

$$\mathbf{E} = \mathbf{R} - \lambda \mathbf{R}_s \quad (55)$$

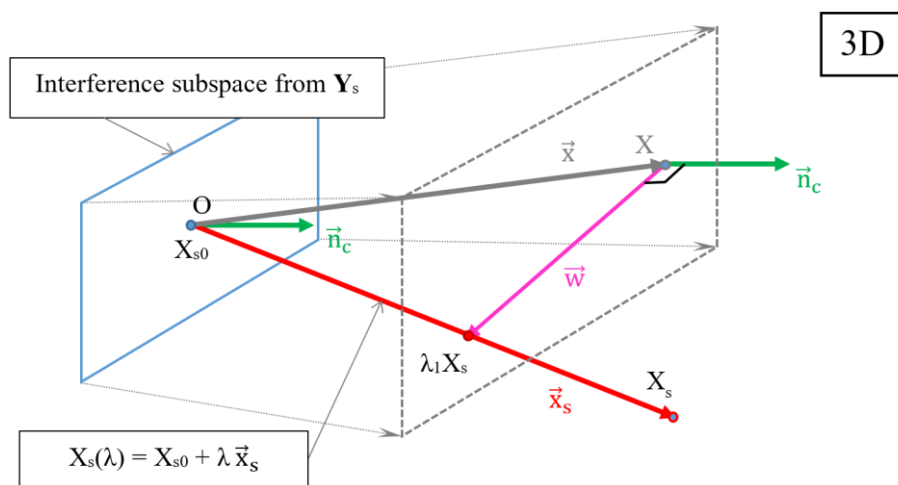
ahol λ az analit koncentráció aránya az ismeretlen és a kalibrációs minták között. A RAFA-t először Ho et al. [77] vezette be iteratív eljárás-ként az adott komponens (analit) meghatározására. 1986-ban Sanchez és Kowalski bemutatta az általánosított rang-eltüntető módszert (Generalized Rank Annihilation Method, GRAM) a rang-eltüntető probléma nem iteratív módon történő megoldására [78].

Egy második tanulmány is megjelent a GRAM tenzoriális perspektívából való leírására, de nem szolgáltatott semmilyen képi illusztrációt a módszer működéséről [79]. Azonban számos újabb tanulmány kimutatta, hogy az algebrai magyarázatok mellett a kemometriában a geometriai megközelítésnek is fontos szerepe van az általános fogalmak megértéséhez [P6-P8,80,P19,81,82].

Jelen munkánkban egy újszerű algoritmust mutatunk be a másodrendű előny elérésére, amely a dualitáson alapul. A javasolt módszerhez a matematikai képletek mellett informatív geometriai szemléltetést is adtunk.



42. ábra Bal oldalon az oszlop-, jobb oldalon a sor-térben mutatjuk be a standard és a zavaró komponens lineáris kombinációjaként adódó adatpontokat



43. ábra Az \vec{n}_c vektor ortogonális a zavaró komponensek alterére

A szimulációs és kísérleti adatok felhasználásával a vizsgálat eredményei azt mutatják, hogy a dualitás elve alapján megbízható λ számítható, amelyből könnyen és gyorsan meghatározható a tényleges koncentráció. A javasolt módszer további előnye, hogy adatvezérelt hibaintervallumokat lehet számolni. Ezzel szemben a meglévő GRAM algoritmusok önmagukban nem tudnak ilyen jellegű információt szolgáltatni, hibaszámítás csak számítógép-intenzív módszerek (pl. Jackknife, Bootstrap stb.) segítségével nyerhető.

A 42. ábrán illusztráljuk kétkomponensű (egy analit és egy zavaró komponens) rendszeren a λ kiszámítására alkalmazott módszerünket. Az egyszerűség kedvéért csak néhány adatpontot (fekete körök) és standard pontot (kékeszöld körök) tüntettünk fel az oszlop- és sortérben, a 42. ábra bal, ill. jobb oldali grafikonján. A zöld, piros és kék vonalak reprezentálják a megengedett megoldás-tartomány külső határait, a standard és a zavaró komponensek altereit. A tiszta komponensek koncentráció ($\mathbf{c}_1, \mathbf{c}_2$) és spektrális ($\mathbf{s}_1, \mathbf{s}_2$) profilpontjait is feltüntettük üres fekete körökkel jelölve. Példaként az \mathbf{R} adatmárix 81. oszlopát a magenta paralelogrammával jelölve határozhatjuk meg a tiszta és a zavaró komponensek lineáris kombinációjaként. A Bouguer-Lambert-Beer-törvény teljesülése esetén pl. a standard mátrix 81. oszlopának ($\mathbf{c}_{st,81}$) távolsága a tőle balra lévő paralelogrammától ($\mathbf{R}_{:,81}$ ponton átmenő magenta paralelogramma oldala és a piros vonal metszéspontjától) meghatározza λ értékét. Másszóval, ha $\mathbf{c}_{st,81}$ -et λ -val megszorozzuk, akkor az a magenta paralelogramma jobb alsó sarkára kerül. Ugyanezt megfigyelhetjük a sortérben is, az \mathbf{R} adatmárix 31. sorát határozzuk meg a magenta paralelogrammával és az $\mathbf{s}_{st,31}$ pont kerül a jobb alsó sarkára a λ -val való szorzás után. A 43. ábra háromkomponensű rendszerre mutatja be a szituációt. Az \vec{n}_c vektor ortogonális a zavaró komponensek alterére amit a dualitás elve alapján határozhatunk meg. Az \vec{n}_c és a \vec{w} vektorok is ortogonálisak, a \vec{w} vektor a meghatározott λX_s analit és az X mintapont közötti különbség vektor lesz: $\vec{w} = \lambda X_s - X$ vagy $\vec{w} = \lambda \vec{x}_s - \vec{x}$. Az \vec{n}_c és \vec{w} vektorok közötti ortogonálitás miatt: $\vec{n}_c(\lambda \vec{x}_s - \vec{x}) = \lambda \vec{n}_c \vec{x}_s - \vec{n}_c \vec{x} = 0$.

Ebből λ kifejezhető:

$$\lambda = \frac{\vec{n}_c \vec{x}}{\vec{n}_c \vec{x}_s} = \frac{\vec{n}_c X}{\vec{n}_c X_s} \quad (56)$$

Ennek általánosítása bármennyi zavaró komponens esetén:

$$\mathbf{\Lambda}_x = (\mathbf{X} \mathbf{X}_{ort}^T) ./ (\mathbf{X}_s \mathbf{X}_{ort}^T) \quad (57)$$

$$\mathbf{\Lambda}_y = (\mathbf{Y} \mathbf{Y}_{ort}^T) ./ (\mathbf{Y}_s \mathbf{Y}_{ort}^T) \quad (58)$$

A $\mathbf{\Lambda}_x$ és $\mathbf{\Lambda}_y$ mátrixok elemeiből átlagot és standard deviációt (SD) számolhatunk. Kiderült azonban, hogy az egyes mátrixelemek nagy zaj esetén szélsőséges értékeket vehetnek fel, ezért a mediánt és a medián abszolút eltérést (median absolute deviation, MAD) mint robusztus becsléseket alkalmaztuk.

HPLC UV-Vis DAD szimulált adatokon mutatjuk be a módszerünket. Két GRAM algoritmust használtunk kontroll módszereként: egyet a PLS Toolboxban [83] implementálva és Faber által házilag készítettet [84]. Különböző zajszinteken a $\lambda = 0.8571$ valós értékkel a következő eredményeket kaptuk:

| Example 2 | $\lambda_{Mean} (\pm SD)$ | $\lambda_{Med} (\pm MAD)$ | $\lambda_{PLS_Toolbox}$ | λ_{Faber} |
|-----------------------|---------------------------|---------------------------|--------------------------|-------------------|
| Noise free data | 0.8571 \pm (0) | 0.8571 \pm (0) | 0.8571 | 0.8571 |
| Noise level of 0.05 % | 0.8576 \pm (0.0006) | 0.8576 \pm (0.0001) | 0.8561 | 0.8583 |
| Noise level of 0.1 % | 0.861 \pm (0.001) | 0.8610 \pm (0.0002) | 0.8577 | 0.8593 |
| Noise level of 0.3 % | 0.853 \pm (0.003) | 0.8539 \pm (0.0009) | 0.8500 | 0.8570 |

Zajmentes adatra mindhárom módszer azonos eredményt adott. A zajszint emelésével mind az SD mind a MAD értékek növekedtek. Az $\text{átlag} \pm k \cdot \text{SD}$ intervallumok tartalmazzák a valódi értéket $k=4$ esetben.

Bár akkurátusan levezett formulák ismertek és publikáltak, pl. [85], hogy megkapjuk a GRAM-becslések szórását, de azokat csak úgy tudjuk felhasználni, ha például a mérési mátrixokhoz kapcsolódó szórásokat előre ismerjük, ami nem mindig áll rendelkezésre, például a nagyon korlátozott ismétlési lehetőségek miatt; így ezek a képletek elsősorban elméleti vizsgálatokra vagy kísérletek tervezésére alkalmasak miután feltártuk a leginkább befolyásoló faktorokat. A mi korlátozott vizsgálódásunk kimutatta, hogy valós helyzetekben a GRAM algoritmus által szolgáltatott becsléseknek nincs szabályos hibaeloszlása, például véges várható értékkel (ami igen erős feltétel a hibaterjedés használatához), ezért néhány rendkívül kis és nagy érték is előfordulhat, így az elméleti levezetések sikertelenek lehetnek, és a gyakorlatban robusztus becsléseket kell használnunk, ahogy ezt mi is tettük a fentiekben.

6. Az új tudományos eredmények tézisszerűen rendezett összefoglalása

1. Az általam Borgen grafikonnak (Borgen plot) elnevezett ábra elkészítéséhez először az adatmátrixunk normalizálását végeztem el az egyes sorok, ill. oszlopok összegeit felhasználva. A normalizálás okozta dimenzió-csökkenés miatt az absztrakt 2-dimenziós síkon lévő adatpontok konvex burkának meghatározását a *convhull* Matlab függvénnyel végeztem (első számítógépes geometriai módszer). Megkaptuk a belső poligont (inner polygon), amit egy megengedett megoldás szimplexe - vagyis három komponens esetén háromszög -, egyetlen oldala sem metszhet, legfeljebb érintheti.

A Fourier-Motzkin eliminációs módszer (FMEM) egy lineáris egyenlőtlenség-rendszerből a redundáns egyenlőtlenségeket szűri ki. A redundáns egyenlőtlenségek észlelése a csúcsok és szélső irányok listájának kezelésével a dupla-deszkripció (double description (DD)) módszerrel történt, amit a Matlabban a *cddmex* harmadik fél által fejlesztett C/C++ alacsony szintű programnyelven készített és előre lefordított külső függvénnyel valósítottam meg (második számítógépes geometriai módszer). A nem redundáns félsíkok a külső poligont (outer polygon) határozzák meg. A megengedett megoldás-tartományok a belső és a külső poligonok között keresendők, tehát a nemnegatív profilok absztrakt pontjai a külső poligonon belül helyezkednek el, és egy háromszög csúcsaként a belső poligont körbeveszik, azaz lineáris konvex kompozícióként az adatpontok előállíthatók.

Ebben a tanulmányban vezettük be az Average Orthogonal Distance from the Linear Subspace (AODLS) elnevezésű mértéket, ami egy ponthalmaz átlagos ortogonális távolságát adja meg egy lineáris altértől. Az AODSL segítségével kimutattuk, hogy a Lawton és Sylvestre által használt adathalmaz akár négy komponenset is tartalmazhat kettő helyett.

2. A kapcsolódó közleményben sikerült általánosan igazolnom, hogy ha a koncentráció és spektrális profilokra csak a nemnegativitás feltételezéssel élünk, akkor a V -térben lévő pontok számával fog megegyezni az U -térben lévő hipersíkok

száma és fordítva. Sőt a konvex csúcsok definícióját értelmezve azt a szigorúbb megállapítást tehetjük, hogy a V-térben lévő külső konvex csúcsok által meghatározott lapok (facets) megfelelnek az U-térben lévő megfelelő belső konvex csúcsoknak és fordítva.

Mivel a belső konvex csúcsokhoz a megfelelő térbe transzformált profilok mint pontok rendelhetők, így az SMCR feladat redukálható a belső politópok (három komponens esetén 2-dimenzióban poligonok) meghatározására és így az esetlegesen időigényes és bonyolult számítógépes geometriai módszerek mellőzhetők és/vagy egyszerűbbre cserélhetők.

A természetes dualitást a következő összefüggés-hálózattal illusztráltam:

$$\begin{array}{l}
 \text{V-tér:} \quad \mathbf{X}_{I \times N} \text{ pontok} \quad \leftarrow \quad \mathbf{Y}_{J \times N} \mathbf{D}^{-1} \mathbf{z} = \mathbf{V}_{J \times N} \mathbf{z} \geq \mathbf{0}_{J \times 1} \text{ hipersíkok} \\
 \text{U-tér:} \quad \mathbf{Y}_{J \times N} \text{ pontok} \quad \leftarrow \quad \mathbf{X}_{I \times N} \mathbf{D}^{-1} \mathbf{z} = \mathbf{U}_{I \times N} \mathbf{z} \geq \mathbf{0}_{I \times 1} \text{ hipersíkok}
 \end{array}$$

3. Az elképzelést, hogy két szélsőértékkel elég lesz jellemezni a forgatási bizonytalanságot, nyilvánvalóan Lawton és Sylvestre kétkomponensű megoldása sugallta. Azonban kimutattam, hogy két szélsőérték általában nem elegendő. Ehhez az általam fejlesztetett és korábban bemutatott Borgen-plot rajzoló Matlab szkriptemet használtam fel.

A rotációs bizonytalanság számszerűsítésére bevezettem a komponensek koncentráció és spektrum szerinti Borgen grafikonbéli megengedett megoldás-tartományainak területe és a külső poligon területe között számolt arányt, valamint a Gemperline-Tauler módszerrel kapott maximum és minimum SCF értékek különbségét. A Gemperline-Tauler módszer komponensenként, a koncentráció és spektrum hatás megkülönböztetése nélkül nyújt információt, míg a terület-arány külön a koncentrációra és külön a spektrumokra is meghatározható. A probléma megvilágításához képzeljük el, hogy az egyik komponens spektruma ismert. Így az ehhez tartozó rotációs bizonytalanság nulla lesz. Azonban a koncentrációhoz tartozó rotációs bizonytalanság nem feltétlenül lesz nulla. A Gemperline-Tauler módszerrel kapott maximum és minimum SCF értékek különbsége viszont ebben az esetben hi-

bás eredményt ad: ha az érték nulla, akkor nem kezelte megfelelően a koncentrációban maradt rotációs bizonytalanságot; ha értéke nem nulla, akkor nem megfelelően kezelte az ismert spektrumot.

4. A Borgen-normák definícióját és részletes leírását adtam meg, amely téma először jelent meg a kémiai szakirodalomban. A Borgen-normák általános alakját a következőképpen határoztam meg: $\|\mathbf{x}\|_{B,z} = \mathbf{z}^T |\mathbf{x}|$, ha $\mathbf{z} > \mathbf{0}$.

Bemutattam még a Borgen normák esetére az általános levezetéseket az SMCR-rel kapcsolatban. Az egyik legfontosabb és legmeglepőbb következtetésem, hogy a rotációs bizonytalanság mértéke függ az alkalmazott Borgen normától a minimális kényszerfeltételű (nemnegatív összetétel- és jel-profilok vélelmezése) SMCR esetén.

Tisztáztam, hogy a Borgen grafikonok segítségével visszatranszformált profilsávokat utólag normalizálhatjuk, vagy változatlanul hagyhatjuk. A nem normalizált sávterületek alkalmasak a komponensek rotációs bizonytalanságának összehasonlítására egy kiválasztott Borgen normára vonatkozóan; míg a normalizált sávterületek a különböző Borgen-normákból adódó rotációs bizonytalanság összehasonlítására használhatók egy kiválasztott komponensre.

Ezenkívül megállapítottam, hogy a különböző Borgen-normákkal elvégzett SMCR vizuális validációs eljárásaként használható a valódi profilok unimodalitásának eldöntésére a profilsávok alapján, előzetes fizikai vagy kémiai ismeretek vagy az adatokra vonatkozó feltételezések nélkül.

5. Elsőként a szakirodalomban, egyértelmű kapcsolatot mutattam be Lawton és Sylvestre, Tauler, valamint Maeder megközelítései között.

A Tauler-Gemperline-féle SCF maximumának és minimumának kétkomponensű rendszerekre vonatkozó számításához szükséges $\bar{\mathbf{T}}$ transzformációs mátrix elemeit a következőképpen adtam meg az 1. komponensre:

$$\bar{t}_{12,MAX,1} = \frac{\bar{x}_{2,xBl}}{\bar{x}_{1,xBl}}, \bar{t}_{21,MAX,1} = \frac{\bar{x}_{1,xAu}}{\bar{x}_{2,xAu}}; \bar{t}_{12,MIN,1} = \frac{\bar{x}_{2,xBu}}{\bar{x}_{1,xBu}}, \bar{t}_{21,MIN,1} = \frac{\bar{x}_{1,xAl}}{\bar{x}_{2,xAl}}$$

illetve a 2. komponensre:

$$\bar{t}_{12,MAX,2} = \frac{\bar{x}_{2,xBu}}{\bar{x}_{1,xBu}}, \bar{t}_{21,MAX,2} = \frac{\bar{x}_{1,xAl}}{\bar{x}_{2,xAl}}; \bar{t}_{12,MIN,2} = \frac{\bar{x}_{2,xBl}}{\bar{x}_{1,xBl}}, \bar{t}_{21,MIN,2} = \frac{\bar{x}_{1,xAu}}{\bar{x}_{2,xAu}}$$

A könnyebb ellenőrizhetőség végett tanulmányomhoz mellékeltem a részletes számításokat elvégző Matlab szkriptet, ill. néhány mesterségesen előállított hibaszint mellett a minimum és maximum SCF-ekhez tartozó sávhatárokat is, és megfigyeltem a hibaszint okozta jellegzetes torzulásokat.

Megemlítem, hogy amíg a Tauler-féle *mcrbands* és a Maeder-féle rácsmódszer az eredeti kezdő profilbecslésekhez ad határokat, addig az LS-alapú számításom csak normalizált határokat ad, azonban ez utóbbi eljárás analitikus és nem-iteratív.

6. Háromkomponensű rendszerek esetén az összes összetartozó megengedett-megoldás halmazt egy speciális rácskeresés módszerrel határoztuk meg, ami a fajtaalapú részecske-raj-optimalizálás (Species-based Particle Swarm Optimization, SPSO) algoritmuson alapul. A módszert zajmentes és zajos adatsor segítségével ellenőriztük. Az eredmények azt mutatták, hogy a módszer alkalmas az adatok elemzésére, azonban meglehetősen időigényes.

A kifejlesztett módszerrel az egyenlőség kényszerfeltételt (ismert koncentráció vagy spektrális profil figyelembevétele: ismert-profil) is alkalmaztuk és várakozásunknak megfelelően a forgatási bizonytalanság és a megengedett megoldás-tartományok drasztikusan csökkentek.

Részletekbe menően megmutattuk még, hogy az ismert-profil kényszerfeltétel a komplementer komponensekre vonatkozóan egyenesek megfelelő szakaszaira egyszerűsödő megengedett-tartományokat eredményez.

7. Úgy tűnik, hogy konvex geometriai szemlélet is szükséges a görbefelbontó módszerek megértéséhez és fejlesztéséhez, mivel a gyakran komplikált (lineáris) algebrai koncepciók kizárólagos alkalmazása megakasztotta a Borgen módszer alaposabb megértését 20 éven keresztül. Az analitikai SMCR módszereket vizsgáltuk felül és írtuk le a dualitás alapelveihez kapcsolódó egyszerű fogalmak segítségével. Ezenkívül, először a szakirodalomban, az egyenlőség (ismert-profil) és unimodalitás kényszerfeltételeket sikeresen implementáltuk a Lawton–Sylvestre módszerhez.

A Borgen grafikonok megrajzolásához is egy korszerűbb eljárást fejlesztettünk az ismert-profil kényszerfeltétel alkalmazásához. Két- és háromkomponensű HPLC-DAD adatsorokat szimuláltunk és elemeztük őket az újabb kényszerfeltételek felhasználásával és azok nélkül.

Az *LSandBP* Matlab szkriptünk a dualitás elvén alapuló egyszerűsítéseket tartalmazza így a Borgen grafikonok néhány másodperc alatt elkészültek. A dualitás elvén alapuló egyszerűsítés pl., hogy az egyik absztrakt térben (mondjuk az oszlop térben) konvex burok algoritmussal meghatározott belső poligon a duális absztrakt térben (mondjuk a sor térben) a külső poligont határozza meg, ami igaz fordítva is, azaz a másik absztrakt térben (mondjuk a sor térben) meghatározott belső poligon az első absztrakt tér (mondjuk az oszlop tér) külső poligonját rögzíti. Így nem kell a számításigényes dupla-deszkripció (DD) (*cddmex*) számítógépes geometriai algoritmust igénybe vennünk.

A másik egyszerűsítés, hogy elegendő csak az egyik absztrakt térben meghatározni a Borgen szimplexeket (Borgen háromszögek három komponens esetén), mert a dualitás elvének alkalmazásával a másik absztrakt térben iteráció nélkül megkapjuk azokat.

8. A Manne tételek kritikáját és új elvek megfogalmazását végeztük el. Sikerült bemutatnunk egy ellenpéldát, amikor Manne 1. és 2. tétele teljesül, de a felbontás nem lesz egyértelmű. A megtalált ellenpélda ösztönzött bennünket, hogy áttekintsük újra az egyértelmű felbontás lehetőségeit, ill. a Manne tételek helyett megfelelő megfogalmazást találjunk.

A Manne tételek a profil-alapú egyértelműségekre vonatkoznak, míg az általunk bevezetett eljárás az adat-alapú egyértelműségekre vonatkozik.

Két formában fogalmazhatjuk meg az adat-alapú egyértelműen megadható profilokra vonatkozó tételünket:

3. forma – Ha a Borgen grafikonon (vagy annak általánosításán háromnál több komponens esetén) találunk olyan pontot, amely mind a belső és mind a külső poligonok valamely csúcspontjaihoz tartozik és ennek következtében

a Borgen háromszögek (általánosan Borgen szimplexek) közös csúcspontja, ez a pont egyértelmű megoldás lesz az adott komponens részére.

4. forma – Ha egy komponensnek van szelektív ablaka az egyik irányban/módban/útban (pl. spektrum) és a másik irányt/módot/utat (pl. koncentráció) tekintve ennek a komponensnek van zéró ablaka, ahol az összes zavaró komponens megjelenik, akkor ennek a komponensnek a második irányban/módban/útban található (koncentráció) profilja egyértelműen megadható.

Elsőként a kemometriai szakirodalomban az alábbi egyértelműségi kategóriákat vezettük be:

- Nem-egyértelmű felbonthatóság (non-uniqueness), azaz egyetlen profil sem adható meg egyértelműen.
- Töredékes-egyértelműség (fractional uniqueness): egy komponens egyetlen profilja adható meg egyértelműen, a többi nem.
- Részleges-egyértelműség (partial uniqueness): egy komponens összes profilja egyértelműen megadható
- Teljes-egyértelműség (full uniqueness): az összes komponens összes profilja egyértelműen megadható.

9. A forgatási bizonytalanság feltárásával kapcsolatban szintetikus és valós adatokat elemezve megállapítottuk, hogy bár a végeredmény nagyon hasonló a bemutatott eljárások algoritmusai lényegi különbséget mutatnak. A Borgen-Rajkó grafikonok számítása explicit módon történik, így nagyon gyorsan megkapjuk az eredményt. Ez az elméleti megfontolásokon alapuló módszer érvényesítésre (validation) használható. Az egyetlen hátránya, hogy valós, zajjal terhelt adatokra nem mindig alkalmazható. A szimplex és poligon tágítás (simplex and polygon inflation) módszerek eredménye nagyon hasonlít az előző elméleti alapú módszerrel kapottra. Az előnye ennek a közelítő módszernek, hogy valós zajjal terhelt adatokkal is működik. Az MCR-BANDS algoritmus teljesen különböző. Ez talán a legkönnyebben elérhető program és bármilyen komponens szám esetén használható. Az algoritmus könnyen implementálható, főképp Matlab-ban, mivel ott készen elérhető

nemlineáris optimalizációs eljárások használhatók nemlineáris kényszerfeltételekkel. Itt a zajhatást újramintázó, pl. bootstrapping eljárásokkal térképezhetjük fel. A hátránya, hogy csak közelíteni tudja a valós sávmegoldásokat, mivel csak két extrém, norma-alapú megoldás alapján dolgozik.

10. Vizsgálódásainkból levontuk a következtetést, hogy Kruskal elégséges és szükséges feltétele két-, ill. háromkomponensű háromutas adattömbökre sérülhet és legfeljebb csak elégséges maradhat. Ten Berge és Sidiropoulos igazolta, hogy $F > 3$ esetekben a nulla értékek speciális mintázata esetén lehetséges egyértelmű felbontás annak ellenére, hogy a Kruskal egyenlőtlenség nem teljesül, ezt most mi igazoltuk $F = 2$ és $F = 3$ esetekre is.

Az adat-alapú egyértelműség elvének alkalmazásával a következő feltételeknek kell teljesülniük az egyértelmű felbontáshoz a Kruskal egyenlőtlenség nem teljesülése esetén: 1) trilinearitás, 2) szelektív ablakok jelenléte, és végül 3) a speciálisan elhelyezkedő zérusok mintázata.

Fontos konklúziónk, hogy a Kruskal egyenlőtlenség teljesülése nem feltétlenül szükséges a kényszerfeltételes trilineáris dekompozícióhoz két-, ill. háromkomponensű háromutas adattömbök esetében.

11. Az elemzett adattömbök alapján a következő megállapításokat tettük. $k\{2,2,2\}$ adattömbök esetén a bemutatott algoritmusunk mindig gyorsan és megfelelően működött, ugyanazt az egyértelmű megoldást szolgáltatva, mint pl. az iteratív eljárás alapuló PARAFAC.

$k\{2,2,1\}$ adattömbök esetén a Kruskal egyenlőtlenség sérül: $2 + 2 + 1 \not\geq 6$, tehát az egyértelmű megoldás nem biztosított. A koncentráció mátrix k -rangja 1 (rangátfedés (rank overlap)), ami a profilok lineáris arányosságának, esetleg megegyezésének következtében alakul ki. Ebben az esetben meglepő módon az összes megengedett megoldás-tartományba eső megoldás trilineáris lesz, azaz nem jelent előnyt a trilineáris kényszerfeltétel alkalmazása, nem kapunk egyértelmű megoldásokat. Fontos következtetésünk, hogy ez a fajta három-utas adattömb bilineáris adatként kezelendő: minden egyes mátrixszelet vizsgálata ugyanarra az eredményre vezet.

Adattömbként kezelés csak a szabatoságot és pontosságot csökkentheti a mérési zaj hatásától függően.

Kimondtuk, hogy ezeket a megállapításokat általánosíthatjuk $k\{N,N,1\}$ típusú adattömbökre, amikor az utolsó út szerinti profilok megegyeznek, azaz $k_3=1$ (N helyett).

12. Az SMCR módszer segítségével határoztunk meg ritkás (sparse) profilokat és kiemeltük, hogy normális adatsor esetén ezek a külső poligonon helyezkednek el, ami az elválasztó határ a csak pozitív és a negatív értékeket is tartalmazó profilok absztrakt pontjai között. A legritkásabb megoldások a külső poligon csúcspontjai lesznek.

Másrésről, ha az adatmátrixunk csak nullákat tartalmazó sorokkal vagy oszlopokkal is rendelkezik (abnormális adatsor), akkor a megengedett megoldás-tartomány belső pontjainak megfelelő profilok ugyanannyi nullát fognak tartalmazni (a nullák elhelyezkedése ezekben a profilokban függ a csak nullákat tartalmazó sorok (oszlopok) mátrixon belüli elhelyezkedésétől: egy teljesen nulla sor az adatmátrixban egy nulla elemet eredményez a koncentráció profilokban, és egy teljesen nulla oszlop az adatmátrixban egy nulla elemet eredményez a spektrális profilokban), de még ekkor is a külső poligon pontjainak megfelelő profilok több nullával fognak rendelkezni. A hasznos tartomány (region of interest, ROI) módszerrel pl. nagyon könnyen kaphatunk abnormális adatsorokat.

A szakirodalomban a ritkás megoldás meghatározására az L_1 norma vagy az L_1 és L_2 normák kombinációjának minimalizálását javasolják (LASSO és elasztikus háló (elastic net)), azonban megállapítottuk, hogy ez csak egyetlen komponens egyetlen profiljának ritkás meghatározásánál használható. Amennyiben 2, 3 vagy több komponens ritkás profiljait kell egyszerre meghatározni, akkor az L_0 norma minimalizálását vagy az L_2 norma maximalizálását és az L_1 norma alkalmazásának mellőzését javasoljuk.

13. Ebben a tanulmányunkban kritikusan áttekintettük a többváltozós kalibrációnál jelentkező túlillesztés problémáját és a hagyományos érvényesítés (validation) alapú eljárást, ami a túlillesztés elkerülését célozza.

Kifejlesztettünk egy alternatív eljárást, ami egy randomizációs-permutációs teszten alapul; ez lehetővé teszi, hogy minden egyes komponenshez, amit a modellbe illesztünk statisztikus szignifikancia szintet rendeljünk.

A következő megállapításokat tettük: (1) alternatív eljárásunk enyhe feltételezések mellett alkalmazható, (2) felhasználóbarát, csak a permutációk számát és a szignifikancia küszöbszintet kell megadni, (3) az eredmények gyakran azonosak az érvényesítés alapúakkal (pl. Unscrambler vagy SIMCA), de most azokat teljesen objektív alapon kaptuk, (4) helyettesítheti az érvényesítést a komponensek kiválasztásához, de kiegészítheti a szokásos diagramot (RMSEP becslések vs. komponensek) is, (5) a randomizációs tesztet nem csak PLS regresszióra alkalmazhatjuk, hanem többutas kalibrációhoz is, (6) tömörítésre lehet szükség nagy adathalmazok (big data) esetén, de arra figyelni kell, hogy a tömörítés ne okozzon függést a minták között, (7) a bemutatott randomizációs teszt a kalibrációs adatkészleten működött, hogy a kereszt-érvényesítés objektivitását biztosítsuk, azonban semmi akadálya, hogy ezt külső adatkészlettel érvényesítésre is alkalmazzuk.

14. Az MCR-ALS algoritmus egy meglepő problémáját tártam fel: az iterációk során kapott rész megoldások és még a végső megoldás is az \mathbf{R} adatmátrix nullterében lesznek.

Megmutattam, hogy a Borgen grafikonon jól illusztrálható az iterációs nyomvonal és az, hogy a nulltérben lévő nemnegatív (rész)megoldás az \mathbf{R} mátrix sor-, ill. oszlopterébe vetítve már negatív elemeket eredményez a profilokban.

A probléma orvoslásához egy új kényszerfeltételt vezettem be: becsült profilok zéró ortogonális résszel (zero orthogonal part for the estimated profiles).

15. Elsőként azt jegyeztem meg, hogy a tiszta pixel nélküli hiperspektrumok esete a földtudományok területén megjelent munkára hivatkoztak, pedig a kémiai szakirodalomban a CCA (convex constraint analysis) algoritmust már korábban leközölték.

Másodsorban arra hívtam fel a figyelmet, hogy a forgatási bizonytalanság nem az MCR-ALS vagy más algoritmus tulajdonsága, hanem a bilineáris adatmátrix belső sajátja!

Lopes et al. a nemnegativitást alkalmazták kényszerfeltételként, valamint egy szükséges feltételt: „a szimplex minden $(p-1)$ dimenziós oldala, azaz lapja (facet) tartalmaz $p-1$ spektrális adatvektort”. Ez kemometriailag azt jelenti, hogy minden komponensre vonatkozóan léteznie kell legalább $p-1$ spektrális sávnak (hullámhossz, hullámszám stb.), ahol csak ez egyik komponensnek lesz nulla az elnyelése. Rávilágítottam, hogy ez olyan erős megkötés, ami csak ritkán fordul elő a gyakorlatban.

Továbbiakban illusztráltam, hogy a koncepció, amit használtak nem minden esetben ad elfogadható eredményt: nemnegatív kényszerfeltétel alkalmazása mellett a minimális térfogatú szimplex (három komponensű rendszerek esetén minimális területű háromszög) a duális térben negatív értékeket is tartalmazó profilok absztrakt pontjait eredményezheti.

Egy további kritikus problémaként az alkalmazott normalizálást jártam körül.

16. A HPLC/FT-IR mérések során az átfolyós folyadékcellával megvalósított IR detektálás előnyeit (gyorsaság, aránylag egyszerű kivitelezés, sokféle eluens alkalmazhatósága) kihasználva a fizikai oldószereliminációt és annak hátrányos tulajdonságait (csak illékony eluens alkalmazható, az elpárologtatás során meghatározandó komponens is távozhat, bonyolult kivitelezés) kemometriai módszerre cseréltük és egy megfelelő algoritmus segítségével vontuk ki az eluens hatását. Kifejlesztettünk egy új iterációs eljárást, az OSSS-IU-PARAFAC-ot (Objective Subtraction of Solvent Spectrum with Iterative Use of PARAFAC).

Zajmentes, de a valóságot jobban közelítő elúciós profilok alkalmazása esetén a PARAFAC2 módszert kellett iteratíván alkalmazni a kielégítő kvalitatív és kvantitatív eredmények eléréséhez.

Sort kerítettünk a mért adattömb értelemszerű megcsonkítására, vagyis kihagytuk azokat az időszakaszokat a hozzájuk tartozó spektrummal együtt, amelyekhez csak eluens spektrum rendelhető (konkrétan: az első kromatográfiás csúcs előtti és az utolsó kromatográfiás csúcs utáni spektrumsorozatokat). A redukált adattömbbel

végrehajtott kétlépéses OSSS-IU-PARAFAC2 eljárás az elúciós és koncentrációprofilokat már jól visszaadta, de a kívánt spektrumokat csak egy újabb szubjektív elimináció végrehajtása után kaptuk meg.

17. A kemometriai feladat ennél a problémánál az volt, hogy megválaszoljam azt a kérdést miszerint a röntgendiffrakciós mérések alapján értelmezni tudjuk-e az olvadékforma legelőnyösebb kioldódási viselkedését. MCR-ALS módszert alkalmaztam és az eredményekből arra következtettem, hogy az olvadékban egy új forma alakult ki (blend), ami 1:10 arány mellett nagyobb mértékben van jelen, mint 3:7 arány esetén.

Megfigyeltük, hogy a fizikai keverékben elhanyagolható, míg a cseppentő módszerrel kapott mintában a legnagyobb mértékben található az új alakulat. A legjobb kioldódást a cseppentő módszerrel kapott termék mutatta.

18. Úgy találtuk, hogy Wiberg és Hultin által használt kemometriai eljárások nem alkalmazhatók az általunk vizsgált meloxikám-mannit fizikai keverék rendszer kioldódási eredményeinek értékelésére. Többek között azért nem, mert az eltérő részecskeméretű meloxikámnak eltérő kioldódási görbét mértünk (itt arra kell figyelni, hogy az eltérő méretű meloxikám UV-Vis spektruma azonos ugyan, de a kioldódási görbék eltérőek), ezért pl. a PARAFAC nem alkalmazható (a profil nem azonos az összes mintára). Sajnos a PARAFAC2 alkalmazása sem adott kielégítő eredményeket.

Ezért az adatsort kettébontottam a kétféle részecskeméretnek megfelelően és az általam kifejlesztett (és korábban részletezett) SMCR módszert megvalósító Matlab programot használtam.

Megállapítottuk, hogy az ME2 jelzéssel ellátott részecskeméretű meloxikám kioldódása kétszer jobb, mint a többi mintáé. A spektrális sávmegoldások az ME1 és az ME2 adatkészletekre nagyon hasonlóak lettek, igazolva eljárásunk következetességét. Legjobb tudomásunk szerint ez az első beszámoló az SMCR módszerek alkalmazásáról háromkomponensű kioldódási adatokra. A sávmegoldások természetesen csak kvalitatív képet tudnak adni, de a legtöbb esetben a sávok elég szűkek ahhoz, hogy a mérési hibán belül kvantitatív következtetéseket is levonhassunk.

19. Megállapítottuk, hogy a lokális rang olyan matematikai fogalom, amely esetleg nincs teljes összhangban a kémiai komponensek számával. Így a forgatási bizonytalanság csökkentése érdekében a lokális rang kényszerfeltétel alkalmazása az MCR módszerekben helytelen megoldásokhoz vezethet!

Ez a probléma az általunk bevezetett szakkifejezés szerint a „lokális rang-vesztés” (local rank deficiency) következménye.

Felhívtuk a figyelmet arra, hogy ha a rang-vesztés csak a választott ablakban fordul elő és nem a teljes mátrixban, akkor pusztán a lokális rang alapján csökkenteni a megengedett megoldás-tartományt kockázatosnak tűnik, hacsak nincs megbízható kémiai információnk az adott ablakra vonatkozóan.

Tehát ebben a munkánkban igazoltuk, hogy az adatmátrixok mikrostruktúrája, amit a Borgen grafikon tár fel, alapvető a lokális rang értelmezése terén. Természetes folytatása ez korábbi eredményünknek, miszerint Manne görbefeletési tételei bár szükségesek, de általában nem elégségesek, így az általunk előzőleg bevezetett és részletesen bemutatott adat-alapú egyértelműséget (data-based uniqueness) kell használnunk.

20. Egy újszerű algoritmust mutattunk be a másodrendű előny elérésére, amely a dualitáson alapul. A javasolt módszerhez a matematikai képletek mellett informatív geometriai szemléltetést is adtunk.

A szimulációs és kísérleti adatok felhasználásával a vizsgálat eredményei azt mutatták, hogy a dualitás elve alapján megbízható λ számítható, amelyből könnyen és gyorsan meghatározható a tényleges koncentráció.

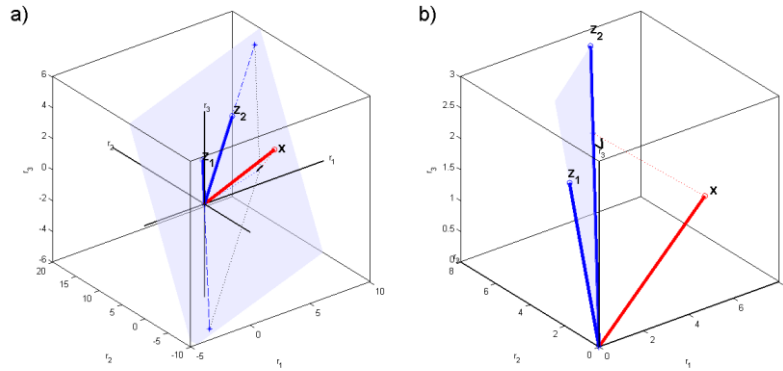
A javasolt módszer további előnye, hogy adatvezérelt hibaintervallumokat tudunk számolni. Ezzel szemben a meglévő GRAM algoritmusok önmagukban nem tudnak ilyen jellegű információt szolgáltatni, hibaszámítás csak számítógép-intenzív módszerek (pl. Jackknife, Bootstrap stb.) segítségével végezhető.

7. Az eredmények más hasznosításának lehetőségei, újabb kutatási irányok, tervek

A 2000-ben alapított *Kemometria és Molekulamodellzés Munkabizottság* (KeMoMo mb., MTA SZAB Kémiai Szakbizottság, Szeged) társelnökeként a mai napig összegyűjtjük a kemometriával és elméleti kémiával foglalkozó nemzetközi rangú magyar szakembereket az évente legalább egyszer megtartott *KeMoMo-QSAR Szimpóziumainkon*, legutóbb 2023. május 18-án és 19-én.

2012-ben a nemzetközi *CAC2012 Budapest* konferencia szatellit rendezvényeként a forgatási bizonytalansággal foglalkozó nemzetközi kutatókat Szegedre invitáltam egy kötetlen találkozóra, amelyről beszámoló is készült [86]. Itt elhatároztuk egy szűk szakmai összejövetelsorozat megrendezését: First Intercontinental Meeting on Rotational Ambiguity, *FIMORA* (Szeged, 2012); Second Intercontinental Meeting on Rotational Ambiguity, *SIMORA*; stb. Sajnos évekig nem történt semmi ezzel kapcsolatban, azonban 2017-ben Marcel Maeder Newcastle-ban (Ausztrália) megszervezte a *TIC2017* (Topics in Chemometrics) konferenciát. Ezen alkalommal én megalapítottam, az inkább szimbolikus, de eszmeileg annál jelentősebb díjat (a díj egy a rendezők által elkészített tic(k), pipa (megjelölés) szobrocska): Comprehensive Cooperation in Chemometrics, C^3 . A díjat első alkalommal Prof. Hamid Andollahi kapta. 2019-ben bár Zanjan (Irán) lett volna a rendező város, az ismert politikai és gazdasági okok miatt Szegedre helyeztük át, így én lettem a házigazdája a *TIC2019 Szeged(Zanjan)* konferenciának, amit a Science Hotelben tartottunk meg bentlakásos formában. A C^3 díjat Prof. Marcel Maeder kapta. Hosszabb kényszerpihenő után 2023 szeptemberében lesz a következő *TIC2023 Rostock* (Németország) konferencia.

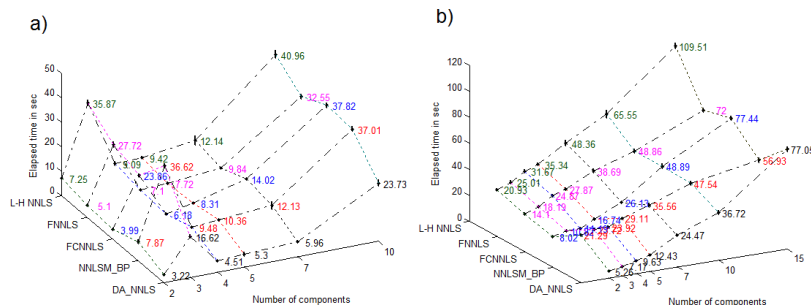
A nemnegatív legkisebb négyzetek (non-negative least squares, NNLS) módszere alapvető fontosságú a minimális kényszerfeltételekkel megvalósított MCR és SMCR módszerek esetében, hiszen éppen a koncentrációk és a spektrumok nemnegativitását írjuk elő.



44. ábra $d_1z_1 + d_2z_2 = x$ a) OLS megoldás, b) NNLS megoldás

2014-ben sikerült egy addig még nem publikált algoritmust kifejleszteni konvex geometriai alapon [87]. Felfedeztük, hogy az NNLS feladat egyenértékű azzal, amikor egy pont és egy konvex gúla közötti távolságot kell meghatároznunk (44. ábra). Így egy duális feladat megoldásaként a távolság algoritmus (distance algorithm, DA) alkalmazásával kifejlesztettük a DA_NNLS algoritmusunkat.

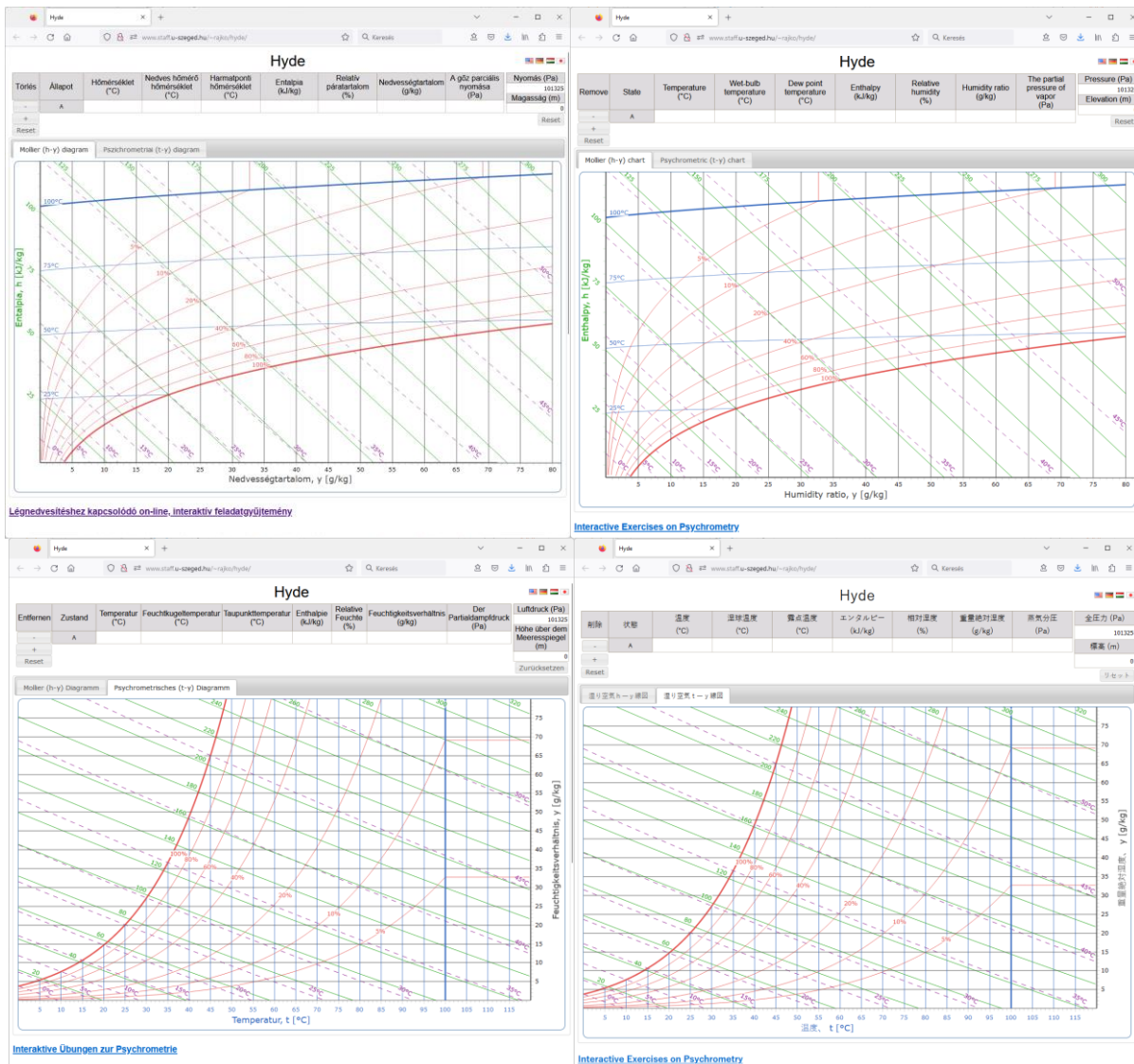
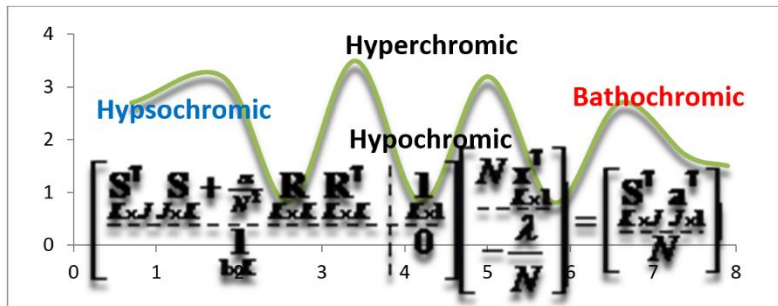
A DA_NNLS a leggyorsabbnak bizonyult kis és közepes méretű lineáris regressziós feladatoknál (45. ábra). Később sikerült egy olyan új NNLS algoritmust kifejlesztünk, ami a teszteléseink során a leggyorsabbnak bizonyult akár kis, közepes vagy nagyméretű lineáris regressziós feladatok esetén amellet, hogy mind az elméleti megalapozottsága mind a programozhatósága a konkurens eljárások közül a leg-egyszerűbb. Szabadalmi bejelentést tettünk, de az sajnos az adminisztráció ingoványában megrekedt. Most az új NNLS algoritmusunk finomítása és a publikálás előkészítése zajlik.



45. ábra MCR-ALS-be ágyazott NNLS algoritmusok vizsgálata gyakorlati adatokon.

a) közepes méretű (310×407) NIR-, b) nagyméretű (120×3403) Raman spektrumok

Az NNLS-sel kapcsolatban merült fel egy érdekes együttműködési lehetőség fizikai-kémikus kollégákkal [88]. Az elméleti kémiában használt DISS (Direct Inversion in the Spectral Subspace) módszerre adtunk meg zárt formulát, amivel iteráció mentesen lehet számolni. A DISS módszer regularizációs változatával (ez már iterációt igényel) a spektrumok eltolódását tudtuk kezelni:



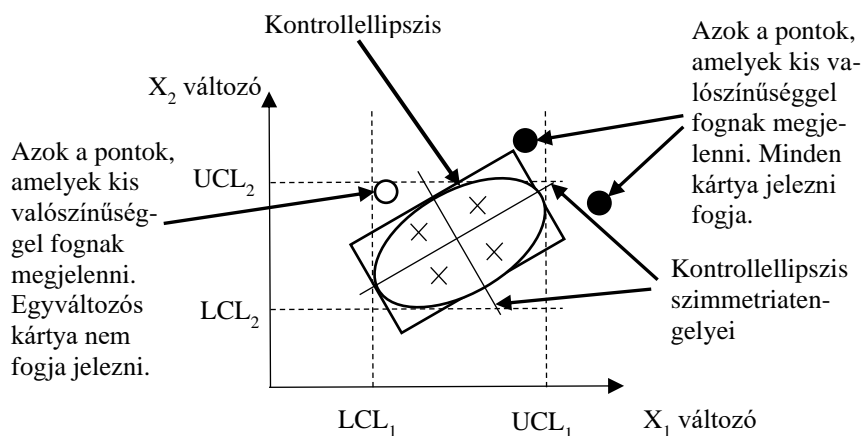
46. ábra A Hyde négy nyelven érhető el

Az élelmiszeripari műveletteni ismereteinket kamatoztattuk amikor kifejlesztettük az interaktív pszichometriai diagramokat légtechnikai számítások elvégzésére [89]. Tovább lépve, egy az oktatásban, iparban és a tudományos kutatásokban is használható webes felületet hoztunk létre, ami négy nyelven érhető el (46. ábra):

<http://sol.cc.u-szeged.hu/~rajko/hyde>

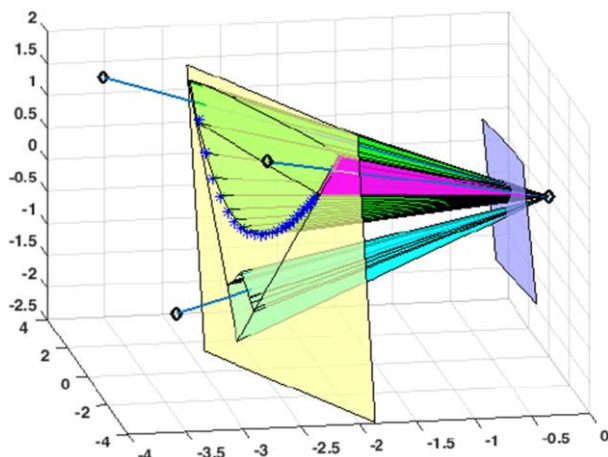
Erdélyi Péter hallgatóm MÉTE OTDK (2014) 2., OTDK (2015) 1. helyezést ért el, majd 2016-ban „A leginnovatívabb TDK munkáért” kategóriában SZTE Innovációs Díjban részesült.

Más élelmiszeripari műveletteni ismereteink kapcsán megvizsgáltuk a többváltozós folyamatszabályozás alkalmazásának előnyeit és hátrányait az élelmiszeriparban [90,91]:



47. ábra Egy- és többváltozós módszerek összehasonlítása

2018-ban részt vettem egy nagyobb lélegzetű tisztázó projektben, ami a closure (záródás, affin kombináció, anyagmérleg) kényszerfeltételt vette górcső alá [92]. A sok értékes megállapítás közül itt most egyet emelek ki, amely a többszörös normalizálás feleslegességére, ill. veszélyére hívja fel a figyelmet és a fenti 48. ábra illusztrálja: az ábrán a closure kényszerfeltétellel kapott sík sárga, míg az L_1 normával normalizált sík lila. A rombusz alakú jelölések három koncentráció profil becslés absztrakt vektorait jelzik, amik megengedett megoldások. Viszont láthatjuk, hogy a closure kényszerfeltétel nem teljesül. Nem fog teljesülni akkor sem, ha a lila síkra vetítjük ezeket a pontokat, és csak akkor teljesül, ha a sárga síkra vetített pontokat használjuk.



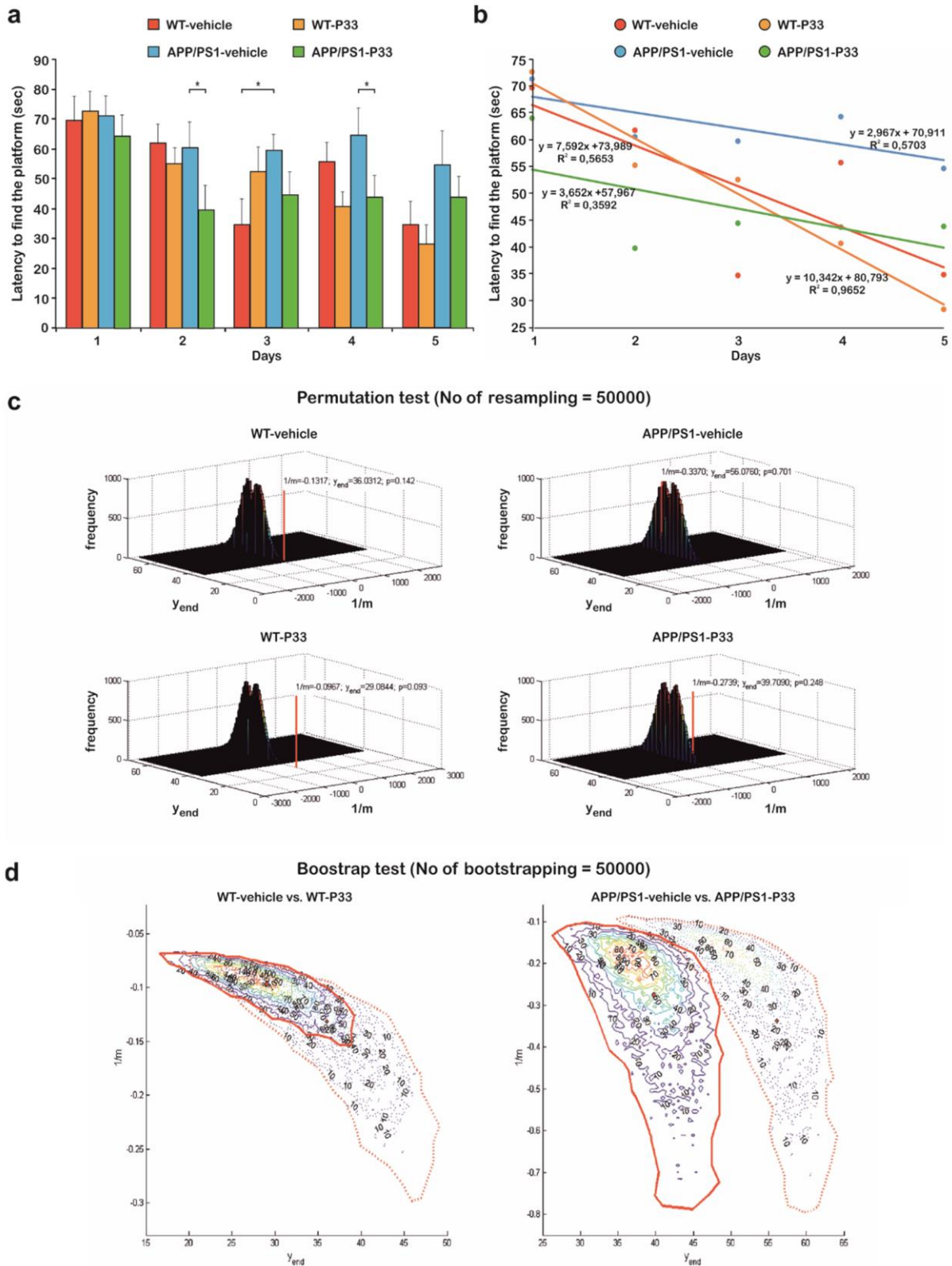
48. ábra Closure kényszerfeltétel illusztrálása

Meg kell említeni még azt a korábban ismert tényt, hogy a closure alkalmazása az egyik profilra (általában a koncentrációra), L1 normát jelent a másik profilra (általában a spektrumra) [P4].

Ugyanebben az évben egyszerű kísérlettervezés alkalmazásával sikerült optimalizálni esszenciális olajtartalmú fertőtlenítők hatását bizonyos baktériumok alkotta biofilmekkel szemben polipropilén felületen [93].

Egy újabb együttműködés keretében a P33 pentapeptid hatását vizsgáltuk a memóriára és a szinaptikus plaszticitásra APP/PS1 transzgenikus egerekben [94]. A feladatom a viselkedés adatok statisztikai elemzése volt (49. ábra). A 49. ábra a) részén egy oszlopdiagramon ábrázoltuk az összes kísérleti adat bizonytalanságát és trendjét a napokhoz viszonyítva (kevert ANOVA szignifikáns: $F_{3, 181} = 3,365$, $p = 0,030$). A b) részén trendvonalakat illesztettünk az átlagos úszási időkre és kiszámítottuk az illesztett pontok R^2 értékeit is. A c) részén a fals pozitív döntési hiba becsléséhez a permutációs tesztek (50000-es újramintázással) háromdimenziós (3D) eloszlási eredményeit tüntettük fel. A WT egerek eredményei jelentősen eltérnek a véletlenszerűségtől (a piros vonalak helyzete messze esnek a véletlenszerűen generált adatok eloszlásától), míg az APP/PS1 hordozóval kezelt egerekkel kapott mérések véletlenszerűnek tekinthetők (a piros vonal a véletlenszerűen generált adatok sűrűjébe esik). A P33-kezelés jelentős javulást eredményezett, mivel a mérések véletlenszerűsége drasztikusan csökkent. A d) részén kétdimenziós (2D) kontúrgramokat ábrázoltunk a vad típusú és transzgenikus csoportok

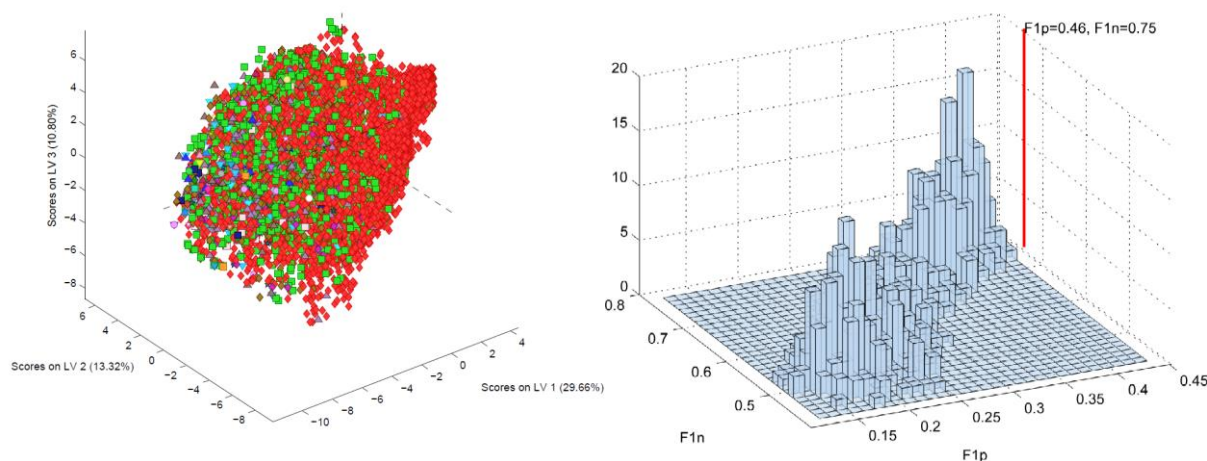
adateloszlásáról bootstrapping (szintén 50000-es újramintázás) alapján. A bal oldali kontúrdiagramok teljes átfedése mutatja, hogy a vad egyedek esetében a P33 kezelés hatástalan volt.



49. ábra Kezelt és kezeletlen egerek viselkedési adatainak statisztikai elemzése

A jobb oldali kontúrdiagramok kis átfedése viszont azt igazolja, hogy a P33-mal kezelt transzgenikus csoport viselkedése szignifikánsan különbözik a nem kezelt csoport viselkedésétől.

A következőkben JAVA osztályokban (programrészletekben) megbúvó programhibák (bugs) kimutatásával foglalkoztunk [95,96]. A feladat bonyolultságát illusztrálja az 50. ábra bal fele, ahol a PLS-DA absztrakt terét láthatjuk: a piros rombuszok a hibátlan osztályokat, a zöld négyzetek egy hibával rendelkező osztályokat, az egyéb jelek több mint egy hibát tartalmazó osztályokat jelentenek. Nem igazán látni az osztályok között hol lehetnek a határok, éppen ezért permutációs tesztet végeztünk, hogy a PLS-DA véletlenszerű eredményt adott-e: az 50. ábra jobb oldalán láthatjuk, hogy a PLS-DA által adott eredmény nem véletlenszerű, a piros vonal jól elkülönül a véletlenszerűen generált adatsűrűségtől. A 3. táblázatból láthatjuk, hogy nincs egyértelműen legjobb módszer, az általunk fejlesztett és javasolt PLS-DA módszer találta meg a legtöbb hibás JAVA osztályt (TP), igaz hogy a fals pozitív (FP) értékei a legrosszabbak. Eredményeinket egyelőre az arXiv kéziratgyűjtő adatbázisba töltöttük fel [95], illetve a Mendeley Data felületen elérhetővé tettük a felhasznált JAVA osztály-adatbázist és a fejlesztett Matlab algoritmusokat [96], jelenleg a kézirat bírálók által kért módosításán és publikálásán dolgozunk.

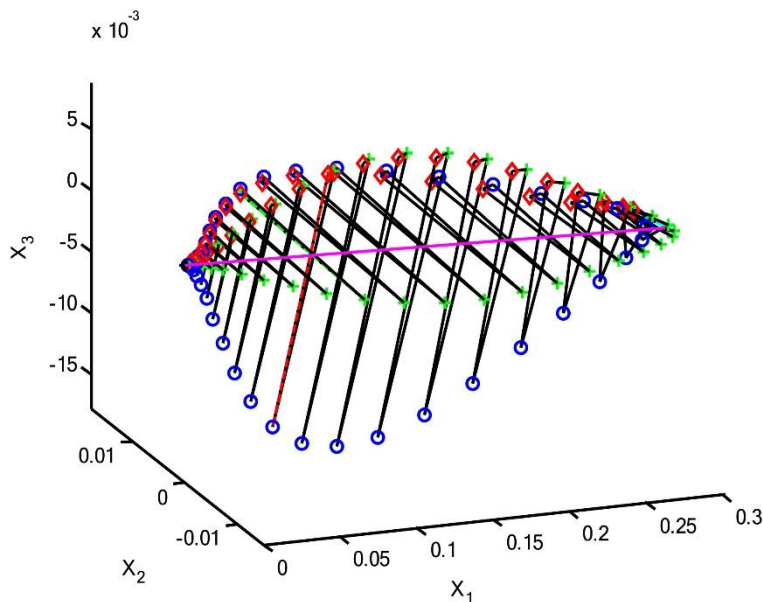


50. ábra

3. táblázat A különböző tanuló algoritmusok/módszerek teljesítményjellemzői újramintázás nélkül (NO) és kitöltő újramintázással (UP)

| Alg. | Precision _p | | Recall _p | | F1 _p | | Precision _n | | Recall _n | | F1 _n | | MCC | | Compl. | | TP | | TN | | FP | | FN | |
|-----------|------------------------|--------------|---------------------|--------------|-----------------|--------------|------------------------|--------------|---------------------|--------------|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|---------------|-------------|--------------|--------------|--------------|
| | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP | NO | UP |
| PLS-DA | 0.374 | 0.338 | 0.593 | 0.721 | 0.458 | 0.460 | 0.894 | 0.915 | 0.774 | 0.681 | 0.830 | 0.780 | 0.300 | 0.300 | 0.693 | 0.794 | 520.9 | 633.1 | 3006.6 | 2642.7 | 876.4 | 1240.3 | 357.1 | 244.9 |
| Std.Dev. | 0.013 | 0.010 | 0.037 | 0.028 | 0.011 | 0.009 | 0.006 | 0.006 | 0.022 | 0.025 | 0.011 | 0.015 | 0.013 | 0.013 | 0.037 | 0.026 | 32.7 | 24.4 | 86.3 | 96.3 | 86.3 | 96.3 | 32.7 | 24.4 |
| DNN | 0.592 | 0.443 | 0.337 | 0.643 | 0.429 | 0.525 | 0.863 | 0.910 | 0.947 | 0.817 | 0.903 | 0.861 | 0.360 | 0.403 | 0.424 | 0.723 | 295.7 | 564.3 | 3679.9 | 3175.2 | 203.9 | 708.6 | 582.3 | 313.7 |
| Std.Dev. | 0.021 | 0.020 | 0.038 | 0.022 | 0.030 | 0.014 | 0.004 | 0.004 | 0.005 | 0.016 | 0.004 | 0.008 | 0.021 | 0.018 | 0.035 | 0.021 | 33.7 | 19.3 | 35.3 | 62.6 | 35.3 | 62.6 | 33.7 | 19.3 |
| Forest | 0.676 | 0.485 | 0.228 | 0.562 | 0.341 | 0.521 | 0.848 | 0.897 | 0.975 | 0.865 | 0.907 | 0.881 | 0.326 | 0.404 | 0.317 | 0.657 | 200.1 | 493.4 | 3787.9 | 3360.5 | 95.9 | 523.3 | 677.9 | 384.6 |
| Std.Dev. | 0.022 | 0.008 | 0.015 | 0.021 | 0.018 | 0.013 | 0.001 | 0.004 | 0.001 | 0.003 | 0.001 | 0.002 | 0.017 | 0.015 | 0.020 | 0.018 | 13.1 | 18.3 | 8.8 | 13.2 | 8.8 | 13.2 | 13.1 | 18.3 |
| KNN | 0.564 | 0.438 | 0.331 | 0.589 | 0.418 | 0.502 | 0.862 | 0.899 | 0.942 | 0.829 | 0.900 | 0.862 | 0.341 | 0.375 | 0.406 | 0.663 | 290.9 | 516.8 | 3659.2 | 3219.8 | 224.6 | 664.0 | 587.1 | 361.2 |
| Std.Dev. | 0.019 | 0.011 | 0.014 | 0.021 | 0.014 | 0.014 | 0.001 | 0.005 | 0.002 | 0.004 | 0.002 | 0.003 | 0.016 | 0.018 | 0.020 | 0.023 | 12.5 | 18.1 | 16.2 | 16.4 | 16.2 | 16.4 | 12.5 | 18.1 |
| SVM | 0.715 | 0.440 | 0.074 | 0.537 | 0.134 | 0.483 | 0.826 | 0.890 | 0.993 | 0.845 | 0.902 | 0.867 | 0.191 | 0.355 | 0.145 | 0.636 | 64.9 | 471.4 | 3857.9 | 3283.2 | 25.9 | 600.6 | 813.1 | 406.6 |
| Std.Dev. | 0.029 | 0.012 | 0.006 | 0.024 | 0.009 | 0.014 | 0.001 | 0.005 | 0.000 | 0.008 | 0.001 | 0.004 | 0.010 | 0.017 | 0.023 | 0.023 | 4.9 | 21.4 | 3.8 | 32.6 | 3.8 | 32.6 | 4.9 | 21.4 |
| Dec. tree | 0.558 | 0.425 | 0.305 | 0.570 | 0.394 | 0.487 | 0.857 | 0.895 | 0.945 | 0.825 | 0.899 | 0.859 | 0.323 | 0.355 | 0.383 | 0.646 | 267.8 | 500.9 | 3671.8 | 3205.2 | 212.0 | 678.6 | 610.2 | 377.1 |
| Std.Dev. | 0.021 | 0.011 | 0.025 | 0.027 | 0.025 | 0.013 | 0.003 | 0.005 | 0.002 | 0.010 | 0.002 | 0.005 | 0.024 | 0.016 | 0.032 | 0.035 | 22.0 | 24.0 | 14.5 | 40.7 | 14.5 | 40.7 | 22.0 | 24.0 |
| Log. reg. | 0.586 | 0.420 | 0.203 | 0.551 | 0.301 | 0.477 | 0.843 | 0.891 | 0.968 | 0.828 | 0.901 | 0.858 | 0.270 | 0.343 | 0.305 | 0.652 | 178.2 | 483.8 | 3757.7 | 3216.8 | 126.1 | 667.0 | 699.8 | 394.2 |
| Std.Dev. | 0.028 | 0.011 | 0.017 | 0.023 | 0.022 | 0.015 | 0.002 | 0.005 | 0.001 | 0.005 | 0.001 | 0.004 | 0.023 | 0.019 | 0.026 | 0.023 | 15.2 | 20.3 | 9.2 | 20.9 | 9.2 | 20.9 | 15.2 | 20.3 |
| Lin. reg. | 0.628 | 0.415 | 0.144 | 0.555 | 0.234 | 0.475 | 0.835 | 0.891 | 0.981 | 0.823 | 0.902 | 0.856 | 0.240 | 0.340 | 0.226 | 0.655 | 126.4 | 487.6 | 3808.9 | 3196.8 | 74.9 | 687.0 | 751.6 | 390.4 |
| Std.Dev. | 0.042 | 0.011 | 0.014 | 0.022 | 0.020 | 0.014 | 0.001 | 0.005 | 0.001 | 0.006 | 0.001 | 0.004 | 0.025 | 0.018 | 0.023 | 0.024 | 12.0 | 19.7 | 8.9 | 21.6 | 8.9 | 21.6 | 12.0 | 19.7 |
| Bayes | 0.354 | 0.346 | 0.577 | 0.605 | 0.439 | 0.440 | 0.889 | 0.892 | 0.762 | 0.742 | 0.821 | 0.810 | 0.287 | 0.288 | 0.685 | 0.708 | 506.9 | 531.2 | 2960.2 | 2880.3 | 923.6 | 1003.5 | 371.1 | 346.8 |
| Std.Dev. | 0.008 | 0.008 | 0.017 | 0.016 | 0.011 | 0.010 | 0.002 | 0.004 | 0.003 | 0.006 | 0.003 | 0.004 | 0.014 | 0.014 | 0.018 | 0.016 | 15.3 | 14.5 | 22.8 | 24.6 | 22.8 | 24.6 | 15.3 | 14.5 |

Korábbi cikkünk [P11] eredményeit sikerült általánosítanunk háromkomponensű trilineáris adattömbökre [97]. A 27. ábra jobb oldalán lévő két komponensű rendszerekre vonatkozó grafikon általánosítása három komponens esetére a 51. ábrán látható. Jelenleg a bármely komponens számra vonatkozó általánosítást keressük.



51. ábra A magenta egyenes egy trilineáris megoldást eredményez, amely átmegegy a kék körök-zöld kereszték-piros rombuszok által meghatározott háromszögeken

A következő nagy projekt az egyértelmű felbontás általános ismérveinek összegyűjtése volt [98]. Főszabályként megállapítottuk, hogy a General Rule for Uniqueness (GRU) alapján minden olyan információ, amely képes rögzíteni egy komponens összes komplementer komponensének alterét az egyik térben, egyértelmű megoldást ad erre a komponensre a másik térben. Példákat hoztunk olyan kényszerfeltételek alkalmazásával, mint a trilinearitás, egyenlőség, zéró-koncentrációjú tartomány, lokális rang és nettó analit jel (net analyte signal, NAT).

Az eredeti adat-alapú egyértelműség elvének megalkotásánál sávmegoldásokkal dolgoztunk [P8]. Kiderült, hogy elegendő csak egyetlen profilbecslés komponensenként, és azokból megfelelő szabályok alkalmazásával eldönthető, hogy a becsült profil egyértelműen meghatározott avagy sem [99]. Tovább dolgozva a témán, matematikai szigorúsággal áttekintettük a helyzetet és számos tételt tudtunk megfogalmazni és bizonyítani [100]. Jelentős megállapításunk volt, hogy kétféle egyértelműség létezik: 1) teljes szigorúsággal tekintett egyértelműség, ekkor egyetlen megoldás szimplex létezik, 2) több megoldás szimplex létezik, de egyikük csúcsai körül sem alakul ki megengedett megoldás tartomány.

Legutóbb a normalizálás problémakörét jártam körül S/MCR alkalmazásokban. Egyelőre csak kézirat formájában jelentettem meg az arXiv-on [101]. Definiáltam a külső és belső normalizálást, számos kinetikai példát hoztam helytelen kémiai rendszer alkalmazására, illetve az irreducibilis mátrixok témakörét érintettem. Mellékeltem egy részletes Matlab szkriptet (jegyzőkönyv, notebook) a Matlab Live Editorával szerkesztve, így interaktívan is kipróbálhatók az algoritmusok.

Látható, hogy a pályázatomban megfogalmazott eredményeket már megalkotásuk közben/után is fel tudtam használni, ill. újabb eredményeket sikerült létrehoznom. A jövőben ez irányú tudományos ténykedésemet szeretném töretlenül folytatni.

8. Irodalomjegyzék

A pályázat alapját képező 20 saját publikáció:

- P1. R. Rajkó, K. István: Analytical solution for determining feasible regions of Self-Modeling Curve Resolution (SMCR) method based on computational geometry. [*Journal of Chemometrics* 19: 448-463, 2005.](#)
- P2. R. Rajkó: Natural duality in minimal constrained self modeling curve resolution. [*Journal of Chemometrics* 20: 164-169, 2006.](#)
- P3. R. Rajkó: Computation of the range (band boundaries) of feasible solutions and measure of the rotational ambiguity in self-modeling/multivariate curve resolution. [*Analytica Chimica Acta* 645: 18-24, 2009.](#)
- P4. R. Rajkó: Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method. [*Journal of Chemometrics* 23: 265-274, 2009.](#)
- P5. R. Rajkó: Additional knowledge for determining and interpreting feasible band boundaries in self-modeling/multivariate curve resolution of two-component systems. [*Analytica Chimica Acta* 661: 129-132, 2010.](#)
- P6. S. Beyramysoltan, R. Rajko, H. Abdollahi: Investigation of the equality constraint effect on the reduction of the rotational ambiguity in three-component system using a novel grid search method. [*Analytica Chimica Acta* 791: 25-35, 2013.](#)
- P7. S. Beyramysoltan, H. Abdollahi, R. Rajkó: Newer developments on self-modeling curve resolution implementing equality and unimodality constraints. [*Analytica Chimica Acta* 827: 1-14, 2014.](#)
- P8. R. Rajko, H. Abdollahi, S. Beyramysoltan, N. Omidikia: Definition and detection of data-based uniqueness in evaluating bilinear (two-way) chemical measurements. [*Analytica Chimica Acta* 855: 21-33, 2015.](#)

- P9. A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, M. Sawall, R. Tauler: A review of recent methods for the determination of ranges of feasible solutions resulting from soft modeling analyses of multivariate data. [*Analytica Chimica Acta* 911: 1-13, 2016.](#)
- P10. R. Rajkó, N. Omidikia, H. Abdollahi, M. Kompany-Zareh: On uniqueness of the non-negative decomposition of two- and three-component three-way data arrays. [*Chemometrics and Intelligent Laboratory Systems* 160: 91-98, 2017.](#)
- P11. N. Omidikia, H. Abdollahi, M. Kompany-Zareh, R. Rajko: Analytical solution and meaning of feasible regions in two-component three-way arrays. [*Analytica Chimica Acta* 939: 42-53, 2016.](#)
- P12. N. Omidikia, M. Ghaffari, R. Rajkó: Sparse non-negative multivariate curve resolution: L0, L1, or L2 norms? [*Chemometrics and Intelligent Laboratory Systems* 199: 103969, 2020.](#)
- P13. N.M. Faber, R. Rajkó: How to avoid over-fitting in multivariate calibration – the conventional validation approach and an alternative. [*Analytica Chimica Acta* 595: 98-106, 2007.](#)
- P14. R. Rajkó: Some surprising properties of Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) algorithms. [*Journal of Chemometrics* 23: 172-178, 2009.](#)
- P15. R. Rajkó: Comments on “Near-Infrared Hyperspectral Unmixing Based on a Minimum Volume Criterion for Fast and Accurate Chemometric Characterization of Counterfeit Tablets”. [*Analytical Chemistry* 82: 8750-8752, 2010.](#)
- P16. K. István, R. Rajkó, G. Keresztury: Towards the solution of the eluent elimination problem in HPLC/IR measurements by chemometric methods. [*Journal of Chromatography A* 1104: 154-163, 2006.](#)
- P17. P.R. Nassab, R. Rajkó, P. Szabó-Révész: Physicochemical characterization of meloxicam-mannitol binary systems. [*Journal of Pharmaceutical and Biomedical Analysis* 41: 1191-1197, 2006.](#)

- P18. R. Rajkó, P. R. Nassab, P. Szabó-Révész: Self-modeling curve resolution method applied for the evaluation of dissolution testing data: A case study of meloxicam-mannitol binary systems. [Talanta 79: 268-274, 2009.](#)
- P19. M. Akbari Lakeh, R. Rajkó, H. Abdollahi: Local rank deficiency caused problems in analyzing chemical data. [Analytical Chemistry 89: 2259-2266, 2017.](#)
- P20. E. Tavakkoli, H. Abdollahi, R. Rajkó: New duality based generalized rank annihilation algorithm for determining analyte concentration with realistically estimated error level for practical data sets. [Chemometrics and Intelligent Laboratory Systems 203: 104058, 2020.](#)

Hivatkozott mások által írt, vagy a pályázatba nem bevont saját publikációk:

- [1] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke: Handbook of Chemometrics and Qualimetrics. Elsevier, Amsterdam, Part A:1997, Part B:1998.
- [2] E.R. Malinowski: Factor Analysis in Chemistry (3rd Ed.). Wiley, New York, 2002.
- [3] S. Brown, R. Tauler, B. Walczak (eds.): Comprehensive Chemometrics. Chemical and Biochemical Data Analysis. Elsevier, Amsterdam, 1ed: 2009, 2ed: 2020.
- [4] Veszprémi T., Fehér M.: A kvantumkémia alapjai és alkalmazásai. Műszaki Könyvkiadó, Budapest, 2002.
- [5] J. Gasteiger, T. Engel (eds.): Chemoinformatics: A Textbook. Wiley, New York, 2004.
- [6] Tasi Gy.: Matematikai kémia. JATEPress, Szeged, 1ed:2009, 2ed:2014.
- [7] Halassy B.: Adatmodellezés. Nemzeti Tankönyvkiadó, Budapest, 2002.
- [8] Rajkó R.: Robusztus eljárások alkalmazása a kémiai kalibráció során. Egyetemi Diplomadolgozat, JATE, Szeged. 1988.

- [9] Rajkó R.: Mérési hibával terhelt koncentráció és jel esetén fuzzy halmazok alkalmazása a kalibráció elvégzéséhez. Tudományos Diákköri Dolgozat, JATE, Szeged. 1988-1989.
- [10] Rajkó R.: Új paraméterbecslő eljárások alkalmazása az analitikai mérésiértékelés gyakorlatában. Egyetemi Doktori Értekezés, JATE, Szeged. 1995.
- [11] Rajkó R.: Regressziós módszerek alkalmazása és fejlesztése a kémiai és élelmiszeripari mérések kiértékelése során. PhD Értekezés, BME, Budapest, 1999.
- [12] Rajkó R.: Kemometriai módszerek szerepe a folyamatelemzéses technológiában, Chemometric methods as process analytical technology (PAT) tools. Habilitációs tantárgyi előadás. SZTE GYTK, Szeged. 2008.
- [13] Rajkó R.: Kemometriai módszerek alkalmazása és fejlesztése, Applications and developments of chemometric methods. Habilitációs tudományos előadás. SZTE GYTK, Szeged. 2008.
- [14] P. Geladi, K. Esbensen: The start and early history of chemometrics: Selected interviews. Part 1. *J. Chemometrics*, 4: 337–354, 1990.
- [15] K. Esbensen, P. Geladi: The start and early history of chemometrics: Selected interviews. Part 2. *J. Chemometrics*, 4: 389–412, 1990.
- [16] Borosy A.P., Héberger K., Horvai Gy., Kolossváry I., Lengyel A., Paksy L., Rajkó, R., Szepesváry P.: Sokváltozós adatelemzés (kemometria). Nemzeti Tankönyvkiadó, Budapest, 2001.
- [17] .D. Rubin: 18 years of multiple imputation. *J. Amer. Stat. Assoc.* 91: 473-486, 1996.
- [18] .S. Wold, K. Esbensen, P. Geladi: Principal Component Analysis. *Chemom. Intell. Lab. Syst.*, 2: 37-??, 1987.
- [19] R. Tauler, A. Smilde, B. Kowalski: Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemometrics*, 9: 31–58, 1995.
- [20] W.H. Lawton, E.A. Sylvestre: Self modeling curve resolution. *Technometrics*, 13: 617–633, 1971.

- [21] R. Tauler: Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.*, 30: 133–146, 1995.
- [22] K.S. Booksh, B.R. Kowalski: *Theory of Analytical Chemistry*. *Anal. Chem.*, 66: 782A–791A, 1994.
- [23] A. Smilde, R. Bro, P. Geladi: *Multi-way Analysis. Applications in the Chemical Sciences*. Wiley, Chichester, 2004.
- [24] A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (eds.): *Fundamentals and Analytical Applications of Multiway Calibration*, Elsevier, Amsterdam, 2015.
- [25] J.D. Carroll, J. Chang: Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart–Young' decomposition. *Psychometrika*, 35: 283–319, 1970.
- [26] R.A. Harshman: Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16: 1–84, 1970.
- [27] J.B. Kruskal: Three-way arrays: rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics, *Linear Algebra Appl.*, 18: 95–138, 1977.
- [28] O.S. Borgen, B.R. Kowalski: An extension of the multivariate component-resolution method to three components. *Anal. Chim. Acta*, 174: 1–26, 1985.
- [29] O.S. Borgen, N. Davidsen, Z. Mingyang, Ø. Øyen: The multivariate N-component resolution problem with minimum assumptions. *Microchim. Acta*, II: 63–73, 1986.
- [30] C.B. Barber, D.P. Dobkin, H.T. Huhdanpaa: The Quickhull algorithm for convex hulls. *ACM Trans. on Mathematical Software*, 22: 469–483, 1996.
- [31] K. Fukuda, A. Prodon: Double description method revisited. In *Combinatorics and Computer Science*, M. Deza, R. Euler, I. Manoussakis (eds). Springer-Verlag, London, 91–111, 1996.

- [32] R.C. Henry, B.M., Kim: Extension of self-modeling curve resolution to mixtures of more than three components. Part 1. Finding the basic feasible region. *Chemom. Intell. Lab. Syst.*, 8: 205–216, 1990.
- [33] B.M., Kim, R.C. Henry: Extension of self-modeling curve resolution to mixtures of more than three components: Part 2. Finding the complete solution. *Chemom. Intell. Lab. Syst.*, 49: 67–77, 1999.
- [34] G.B. Dantzig, B.C. Eaves: Fourier-Motzkin elimination and its dual. *J. Comb. Theory, Ser. A.*, 14: 288–297, 1973.
- [35] K. Fukuda: *cdd.c*: C-implementation of the double description method for computing all vertices and extremal rays of a convex polyhedron given by a system of linear inequalities. Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1993. program available from http://www.ifor.math.ethz.ch/~fukuda/cdd_home/index.html [April 12, 2005].
- [36] The Mathworks Inc., *MATLAB1*, Version 6.1. (R12.1), User's Guide, Natick, MA 2000.
- [37] F. Torrisi, M. Baotic: *CDDMEX*: a Matlab MEX file interface obtained by linking CDD library *CDDLIB* [40]. <http://control.ee.ethz.ch/~hybrid/cdd.php> [April 12, 2005].
- [38] R. C. Henry: Current factor analysis receptor models are ill-posed. *Atmos. Environ.*, 21: 1815–1820, 1987.
- [39] P.J. Gemperline: Computation of the range of feasible solutions in self-modeling curve resolution algorithms. *Anal. Chem.*, 71: 5398–5404, 1999.
- [40] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, *J. Chemometr.* 15: 627–646, 2001.
- [41] H. Abdollahi, M. Maeder, R. Tauler: Calculation and meaning of feasible band boundaries in multivariate curve resolution of a two-component system. *Anal. Chem.* 81: 2115–2122, 2009.

- [42] R. Tauler: Application of non-linear optimization methods to the estimation of multivariate curve resolution solutions and of their feasible band boundaries in the investigation of two chemical and environmental simulated data sets. *Anal. Chim. Acta* 595: 289–298, 2007.
- [43] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, M. Maeder: On rotational ambiguity in model-free analyses of multivariate data. *J. Chemometr.* 20: 302–310, 2006.
- [44] A. Golshan, H. Abdollahi, M. Maeder: Resolution of rotational ambiguity for three-component systems. *Anal. Chem.* 63: 836–841, 2011.
- [45] Li, X: Adaptively choosing neighbourhood bests using species in a particle swarm optimizer for multimodal function optimization. In: K. Deb (eds) *Genetic and Evolutionary Computation – GECCO 2004*. GECCO 2004. Lecture Notes in Computer Science, vol 3102. Springer, Berlin, Heidelberg, 105–116, 2004.
- [46] R. Manne: On the resolution problem in hyphenated chromatography. *Chemometr. Intell. Lab. Syst.*, 27: 89–94, 1995.
- [47] J.M.F. ten Berge, N.D. Sidiropoulos, On uniqueness in CANDECOMP/PARAFAC, *Psychometrika*, 67: 399–409, 2002.
- [48] N. Omidikia, H. Abdollahi, M. Kompany-Zareh, On uniqueness and selectivity in three-component parallel factor analysis, *Anal. Chim. Acta*, 782: 12–20, 2013.
- [49] P. Paatero, P.K. Hopke, X.-H. Song, Z. Ramadan, Understanding and controlling rotations in factor analytic models, *Chemom. Intell. Lab. Syst.*, 60: 253–264, 2002.
- [50] J.B. Kruskal: Rank decomposition, and uniqueness for 3-way and N-way arrays. In: R. Coppi, S. Bolasco (eds.), *Multiway Data Analysis*, Elsevier, North-Holland, Amsterdam, 8–18, 1989.
- [51] M.J. Culzoni, H.C. Goicoechea, A.P. Pagani, M.A. Cabezón, A.C. Olivieri: Evaluation of partial least-squares with second-order advantage for the multi-way

- spectroscopic analysis of complex biological samples in the presence of analyte-background interactions. *Analyst*, 131: 718–723, 2006.
- [52] V.A. Lozano, G.A. Ibanez, A.C. Olivieri: Second-order analyte quantitation under identical profiles in one data dimension. A dependency- adapted partial least-squares/residual bilinearization method. *Anal. Chem.*, 82: 4510–4519, 2010.
- [53] E. Gorrochategui, J. Jaumot, R. Tauler: ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets. *BMC Bioinf.* 20: 1–17, 2019.
- [54] R. Tibshirani: Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58: 267–288, 1996.
- [55] H. Zou, T. Hastie: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67: 301–320, 2005.
- [56] R.D. Snee: Validation of regression models: Methods and examples. *Technometrics*, 19: 415–428, 1977.
- [57] R. Tauler, A. de Juan: Multivariate curve resolution homepage. <http://www.ub.edu/mcr/welcome.html> [November 22, 2008].
- [58] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler: A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemom. Intell. Lab. Syst.*, 76: 101–110, 2005.
- [59] M.H. Van Benthem, M.R. Keenan: Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *J. Chemom.*, 18: 441–450, 2004.
- [60] M.B. Lopes, J.C. Wolff, J.M. Bioucas-Dias, M.A. Figueiredo. Near-infrared hyperspectral unmixing based on a minimum volume criterion for fast and accurate chemometric characterization of counterfeit tablets. *Anal. Chem.*, 82: 1462–1469, 2010.
- [61] M.D. Craig: Minimum-volume transforms for remotely sensed data. *IEEE Trans. Geosci. Remote Sens.*, 32: 542–552, 1994.

- [62] A. Perczel, M. Hollósi, G. Tusnády, G.D. Fasman: Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.*, 4: 669–679, 1991.
- [63] A. Perczel, K. Park, G.D. Fasman: Analysis of the circular dichroism spectrum of proteins using the convex constraint algorithm: A practical guide. *Anal. Biochem.*, 203: 83–93, 1992.
- [64] R.C. Henry: Multivariate receptor modeling by N-dimensional edge detection. *Chemom. Intell. Lab. Syst.*, 65: 179–189, 2003.
- [65] J. Chalmers, P. Griffiths (eds): *Handbook of Vibrational Spectroscopy*. Wiley, Chichester, 2002.
- [66] D. Kuehl, P.R. Griffiths: Novel approaches to interfacing a high performance liquid chromatograph with a fourier transform infrared spectrometer. *J. Chromatogr. Sci.*, 17: 471–476, 1979.
- [67] G.W. Somsen, C. Gooijer, N.H. Velthorst, U.A.Th. Brinkman: Coupling of column liquid chromatography and Fourier transform infrared spectrometry. *J. Chromatogr. A*, 811: 1–34, 1998.
- [68] K.H. Wiberg, U.K. Hultin: Multivariate chemometric approach to fiber-optic dissolution testing. *Anal. Chem.*, 78: 5076–5085, 2006.
- [69] A. de Juan, E. Casassas, R. Tauler: Soft Modeling of Analytical Data. In R.A. Meyers, S.D. Brown (eds): *Encyclopedia of Analytical Chemistry*, Wiley, New York, 2000.
- [70] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbühler: Calculation of equilibrium constants from multiwavelength spectroscopic data—II132, 95.: Specfit: two user-friendly programs in basic and standard fortran 77. *Talanta*, 32: 257–264, 1985.
- [71] H.R. Keller, D.L. Massart: Peak purity control in liquid chromatography with photodiode-array detection by a fixed size moving window evolving factor analysis. *Anal. Chim. Acta*, 246: 379–390, 1991.
- [72] A.C. Whitson, M. Maeder: Exhaustive evolving factor analysis (E-EFA). *J. Chemometrics*, 15: 475–484, 2001.

- [73] Z.-D. Zeng, C.-J. Xu, Y.-Z. Liang, B.-Y. Li: Sectional moving window factor analysis for diagnosing elution chromatographic patterns. *Chemom. Intell. Lab. Syst.*, 69: 89–101, 2003.
- [74] G.M. Escandar, H.C. Goicoechea, A.M. de la Peña, A.C. Olivieri: Second- and higher-order data generation and calibration: A tutorial. *Anal. Chim. Acta*, 806: 8–26, 2014.
- [75] Å. Rinnan, J. Riu, R. Bro: Multi-way prediction in the presence of uncalibrated interferences. *J. Chemometrics*, 21: 76–86, 2007.
- [76] N. Kumar, A. Bansal, G.S. Sarma, R.K. Rawal: Chemometrics tools used in analytical chemistry: An overview. *Talanta*, 123: 186–199, 2014.
- [77] C.N. Ho, G.D. Christian, E.R. Davidson: Application of the method of rank annihilation to quantitative analyses of multicomponent fluorescence data from the video fluorometer. *Anal. Chem.*, 50: 1108–1113, 1978.
- [78] E. Sanchez, B.R. Kowalski: Generalized rank annihilation factor analysis. *Anal. Chem.*, 58: 496–499, 1986.
- [79] E. Sanchez, B.R. Kowalski: Tensorial calibration: II. Second-order calibration. *J. Chemometrics*, 2: 265–280, 1988.
- [80] S.K. Karimvand, M. Akbari Lakeh, E. Tavakkoli, M. Ghaffari, N. Omidikia, S.K.A. Abad, R. Rajkó, H. Abdollahi: A general rule for uniqueness in self-modeling curve resolution methods. *J. Chemometrics*, 34: e3268, 2020.
- [81] M. Akbari, H. Abdollahi: Investigation and visualization of resolution theorems in self modeling curve resolution (SMCR) methods. *J. Chemometrics*, 27: 278–286, 2013.
- [82] M. Alinaghi, R. Rajkó, H. Abdollahi: A systematic study on the effects of multi-set data analysis on the range of feasible solutions, *Chemom. Intell. Lab. Syst.*, 153: 22–32, 2016.
- [83] PLS_Toolbox 8.8. Eigenvector Research, Inc., Manson, WA USA 98831, 2018. software available at <http://www.eigenvector.com>
- [84] [http://www.chemometry.com/Index/Links and downloads/Papers/Derivation GRAM.pdf](http://www.chemometry.com/Index/Links_and_downloads/Papers/Derivation_GRAM.pdf) [January 6, 2017]

- [85] J. Ferré, R. Boqué, N.M. Faber: Multiway data analysis: eigenvector-based methods. Ch. 2.21, In S. Brown, R. Tauler, B. Walczak (eds.): *Comprehensive Chemometrics. Chemical and Biochemical Data Analysis*. Elsevier, Amsterdam, 365–409, 2009.
- [86] A. de Juan, H. Abdollahi, M. Maeder, R. Rajko, R. Tauler: Report on a meeting on rotational ambiguity in soft-modelling analyses. *J. Chemometrics*, 27: 1–2, 2013.
- [87] R. Rajkó, Y. Zheng: Distance Algorithm based procedure for Non-Negative Least Squares. *J. Chemometrics*, 28: 691–695, 2014.
- [88] R. Rajkó, S. Beyramysoltan, H. Abdollahi, J. Eóri, G. Pongor: Fitting-free curve resolution of spectroscopic data: chemometric and physical chemical viewpoints. *Anal. Chim. Acta*, 888: 19–26, 2015.
- [89] P. Erdélyi, R. Rajkó: Using interactive psychrometric charts to visualize and explore psychrometric processes. *J. Chem. Educ.* 93: 391–393, 2016.
- [90] Mihalkó J., Rajkó R: Többváltozós folyamatszabályozás alkalmazásának előnyei és hátrányai az élelmiszeriparban. In: T. Alapi, I. Ilisz (Eds.) *Proceedings of the 22nd International Symposium on Analytical and Environmental Problems : October 10, 2016 Szeged, Magyarország : University of Szeged, Department of Inorganic and Analytical Chemistry (2016)* 438 p. pp. 385–389.
- [91] J. Mihalkó, R. Rajkó: Investigation of multivariate statistical process control in R environment. *Anal. Tech. Szeged.*, 11: 36–40, 2017.
- [92] N. Omidikia, S. Beyramysoltan, J.J. Mohammad Jafari, E. Tavakkoli, M. Akbari Lakeh, M. Alinaghi, M. Ghaffari, S. Khodadadi Karimvand, R. Rajkó, H. Abdollahi: Closure constraint in multivariate curve resolution. *J. Chemometrics*, 32: e2975, 1–15, 2018.
- [93] A. Vidács, E. Kerekes, R. Rajkó, T. Petkovits, N. S. Alharbi, J. M. Khaled, Cs. Vágvölgyi, J. Krisch: Optimization of essential oil-based natural disinfectants against *Listeria monocytogenes* and *Escherichia coli* biofilms formed on polypropylene surfaces. *J. Mol. Liq.*, 255: 257–262, 2018.

- [94] T. Szögi, I. Schuster, E. Borbély, A. Gyebrovszki, Zs. Bozsó, J. Gera, R. Rajkó, M. Sántha, B. Penke, L. Fülöp: Effects of the Pentapeptide P33 on Memory and Synaptic Plasticity in APP/PS1 Transgenic Mice: A Novel Mechanism Presenting the Protein Fe65 as a Target. *Int. J. Mol. Sci.*, 20: 3050, 1–20, 2019.
- [95] R. Ferenc, I. Siket, P. Hegedűs, R. Rajkó: Employing Partial Least Squares Regression with Discriminant Analysis for Bug Prediction. <https://doi.org/10.48550/arXiv.2011.01214> 32 p., 2020.
- [96] R. Rajkó, R. Ferenc, I. Siket, P. Hegedűs: Data for: Employing Partial Least Squares Regression with Discriminant Analysis for Bug Prediction. <https://doi.org/10.17632/CB22T5225N.1> Mendeley Data, V1, 2020.
- [97] N. Omidikia, H. Abdollahi, M. Kompany-Zareh, R. Rajkó: Trilinear self-modeling curve resolution using Borgen-Rajkó plot. *J. Chemometrics*, 34: e3161, 1–17, 2020.
- [98] S. Khodadadi Karimvand, M. Akbari Lakeh, E. Tavakkoli, M. Ghaffari, N. Omidikia, S. Khalili Ali Abad, R. Rajkó, H. Abdollahi: A general rule for uniqueness in self-modeling curve resolution methods. *J. Chemometrics*, 34: e3268, 1–14, 2020.
- [99] M. Akbari Lakeh, H. Abdollahi, R. Rajkó: Predicting the uniqueness of single non-negative profiles estimated by multivariate curve resolution methods. *Anal. Chim. Acta*, 1199: 339575, 1–10, 2022.
- [100] N. Gillis, R. Rajkó: Partial Identifiability for Nonnegative Matrix Factorization. *SIAM J. Matrix Anal. Appl.*, 44: 27–52, 2023.
- [101] R. Rajkó: On problematic practice of using normalization in Self-modeling/Multivariate Curve Resolution (S/MCR): with Supplementary Material. <https://doi.org/10.48550/arXiv.2308.03801> 63 p., 2023.