

**Analýza multispektrálnych dát z konfokálnej mikroskopie.**

**DIPLOMOVÁ PRÁCA**

Kamil Paulíny

**UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY FYZIKY A INFORMATIKY  
KATEDRA APLIKOVANEJ INFORMATIKY**

Študijný odbor: Informatika

Vedúci záverečnej práce: RNDr. Anton Mateašik, PhD.

BRATISLAVA 2007

Čestne vyhlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím citovaných zdrojov.

Bratislava, Máj 2007

Kamil Paulíny

## **Abstrakt**

Konfokálna laserová rastrovacía mikroskopia poskytuje unikátne prostriedky pre štúdium funkčných a štruktúrnych charakteristík živých systémov. Pomocou špecifických fluorescenčných farbív, ktoré sa dokážu naviazať na komponenty študovanej vzorky, a za predpokladu, že poznáme emisné spektrá týchto farbív, môžeme získať presnú distribúciu sledovaných biologických objektov. V praxi môžu nastať situácie, keď nepoznáme spektrálne vlastnosti fluorescenčných látok, ktoré sa nachádzajú vo vzorke. Pre takéto situácie je potrebný algoritmus na separovanie jednotlivých komponentov. V tejto práci sa budeme zaoberať návrhom algoritmu pre problém dekompozície spektrálne odlišených fluorescenčných dát, ktoré boli získané konfokálnym laserovým mikroskopom.

Kľúčové slová: konfokálny mikroskop, fluorescencia, dekompozícia spektrálne odlišených dát, Independent Component Analysis.

# Obsah

1. Úvod.....	5
2. Prehľadová kapitola.....	7
2.1. Konfokálna laserová rastrovacía mikroskopia.....	7
2.2. Dekompozícia spektrálne rozlíšených dát.....	8
2.2.1. Lineárny unmixing.....	9
2.2.2. Dekompozícia bez znalosti spektrálnej matice.....	10
2.3. Metóda hlavných komponentov (PCA).....	11
2.4. Metóda nezávislých komponentov (ICA).....	12
2.4.1. Nezávislosť a nekorelovanosť.....	13
2.4.2. Nezávislé = nie Gaussové.....	13
2.4.3. Miery negaussianity.....	14
2.4.4. Predspracovanie.....	16
2.4.5. Algoritmus FastICA.....	17
2.5. Genetický algoritmus.....	17
3. Ciele diplomovej práce.....	21
4. Výsledky a diskusia.....	22
4.1. Kritériá.....	22
4.2. Všeobecné spektrá.....	24
4.2.1. Implementácia genetického algoritmu.....	24
4.2.2. Test algoritmu.....	25
4.3. Ohraničené spektrá.....	27
4.3.1. Modifikácia algoritmu.....	28
4.3.2. Test algoritmu.....	29
4.4. Uniformné spektrum.....	31
4.5. Zložené spektrá.....	33

4.6. Paralelný výpočet.....	34
4.7. Výsledky na dátach z praxe.....	34
5. Záver.....	42
6. Zoznam skratiek.....	43
7. Používateľská príručka.....	44
8. Použitá literatúra.....	46

# 1. Kapitola

## Úvod

Konfokálna laserová rastrovacia mikroskopia (confocal laser scanning microscopy – CLSM) poskytuje unikátne prostriedky pre štúdium funkčných a štruktúrnych charakteristík živých systémov. Základným princípom CLSM je osvetlenie vzorky laserovým lúčom, ktorý je zaostrý na jediný bod. Následne sa vyžiarené svetlo (fluorescencia alebo odrazené svetlo) premieta na apertúru, ktorá je umiestnená pred snímacím detektorom. Takto, bod po bode, konfokálna mikroskopia umožňuje objemové snímanie mikroštruktúr selektívnym zobrazovaním paralelných rezov [Pawley, 2006]. Najnovšie CLSM zariadenia umožňujú aj spektrálnu dekompozíciu emitovaného svetla prostredníctvom súčasného merania vo viacerých pásmach s nastaviteľnou frekvenciou [Dickinson a kol., 2001].

V biomedicíne je spektrálna informácia často využívaná na zvýšenie informačného obsahu. Pomocou spektrálnej informácie môžeme rozlíšiť rôzne regióny tkaniva. Pomocou precízne vybratých indikátorov, napríklad protilátok, s presne definovanými spektrálnymi vlastnosťami, môžeme detekovať viaceré, vysoko špecifické biologické objekty. Naša schopnosť vnímať a vyhodnotiť spektrálnu informáciu je výrazne obmedzená. Sme limitovaní na viditeľné svetlo a všetky spektrálne informácie, nezáleží na tom ako zložité, spadajú do troch širokých spektrálnych rozpätí. Spektrálne rozlíšená konfokálna laserová mikroskopia nám umožňuje meranie optického spektra v každom bode vzorky. Tým nám umožňuje prekonať tieto nedostatky a pozorovať rozdiely v spektrálnej informácii, ktoré by boli normálne nepostrehnuteľné. Do

vzorky môžeme pridať špeciálne fluorescenčné farbivá s odlišnými emisnými spektrami, ktoré sa naviažu na hľadané biologické komponenty. Sledovaním týchto farbív získame informácie o distribúcii jednotlivých komponentov a ich vzájomnom vzťahu. Pomocou CLSM mikroskopu získame sériu obrázkov, pričom každý obrázok je nasnímaný pri rozdielnej vlnovej dĺžke. Ak sa spektrá farbív neprekrývajú, tak môžeme komponenty ľahko separovať. V opačnom prípade musíme použiť dodatočné algoritmy na extrakciu čo najväčšieho množstva informácie z analyzovanej scény. Pre prípad, keď poznáme spektrá použitých farbív, boli vyvinuté algoritmy, ktoré dokážu separovať pôvodné komponenty, aj keď sa priestorovo aj spektrálne signifikantne prekrývajú [Dickinson a kol., 2001].

Farbivá nemajú nemenné spektrá, na ich tvar má vplyv prostredie, v ktorom sa nachádzajú. Vo vzorke sa tiež môžu nachádzať látky, ktoré majú vlastnosť autofluorescencie. Preto potrebujeme aj algoritmy pre prípad, keď spektrá nepoznáme. Prvé používané metódy sú z triedy štatistických techník druhého rádu. Medzi ne patria napríklad Metóda hlavných komponentov (PCA) a Faktorová analýza. Tie sú postačujúce len pre málo sa prekrývajúce komponenty. Spektrálna mikroskopia nie je jediná oblasť, ktorá sa týmto problémom zaoberá. Diaľkové snímanie pomocou satelitov potrebuje riešiť podobný problém. Táto oblasť sa vyvíja už dlhšiu dobu, za ktorú preskúmali využitie viacerých techník vyšších rádov. Metóda hlavných komponentov (ICA) je momentálne najpoužívanejšou metódou [Nascimento, Dias, 2005]. Biologické dáta, získané CLSM mikroskopom, majú oproti satelitným dátam rozdielne vlastnosti. Hlavnými negatívami je podstatne slabší signál a nižšie spektrálne rozlíšenie. Tieto odlišnosti sťažujú použitie rovnakých metód. Nájdenie nových metód na riešenie tohto problému je cieľom tejto diplomovej práce.

## 2. Kapitola

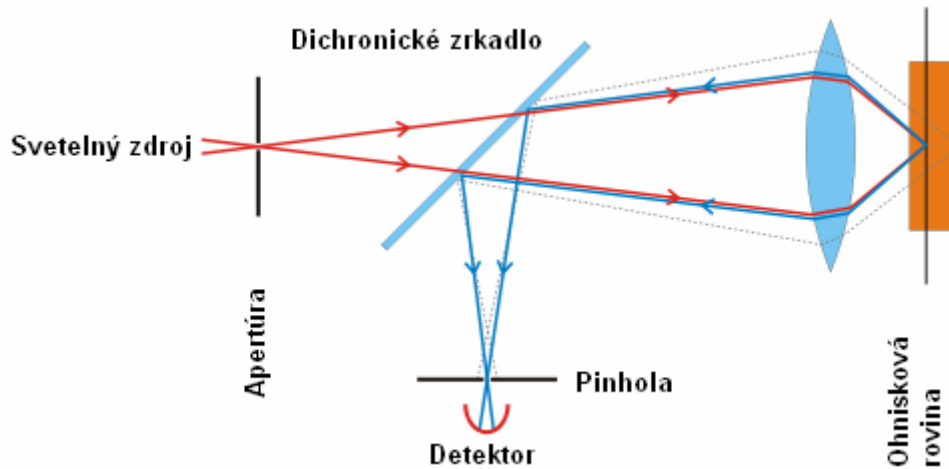
### Prehľadová kapitola

#### 2.1 Konfokálna laserová rastrovacia mikroskopia

V konvenčnej široko pásmovej mikroskopii je vzorka osvetlená horizontálne aj vertikálne v rovnaký čas. Výsledný obraz je ovplyvnený každým osvetleným bodom, aj keď je sústredený v ohnisku. Tieto osvetlené body navzájom interferujú a rušivé svetlo kompromituje kontrast obrazu. Kontrast obrazu, definovaný ako rozdiel medzi maximálnou a minimálnou intenzitou dvoch bodov, má výrazný vplyv na výsledné rozlíšenie. Bez náležitého kontrastu je len malý rozdiel medzi signálom a pozadím a rozlíšenie nedosiahne možnosti šošovky.

Osvetlením len jedného bodu eliminujeme ruch spôsobený nežiadúcim svetlom, a tým dosiahneme lepšiu kvalitu obrazu. Laser je dobrým kandidátom na dosiahnutie takéhoto osvetlenia. Okrem toho laser dokáže poskytnúť zmiešané svetlo diskretných vlnových dĺžok a má vysokú intenzitu, čo je výhodné pre fluorescenčnú excitáciu. V praxi sa na imitáciu bodového svetla používa laserový lúč (obr. 1). Svetlo je pomocou šošovky nasmerované na malú časť vzorky (bod). Reakciou na svetlo je vyžiarovanie fotónov, ktoré sú nasmerované do detektora fotónov. Keďže laserový lúč prechádzal aj bodmi mimo ohnisko, potrebujeme odfiltrovať vyžarovanie z týchto bodov. Preto pred detektor postavíme pinholu, cez ktorú sa tieto nechcené fotóny nedostanú. Získaný obraz je ostrejší a má vyššie rozlíšenie. Tieto operácie výrazne znižujú množstvo prijatých fotónov (približne 50-násobne), čím zvyšujú citlivosť na šum.





Obr. 1. Princíp CLSM [15]

Počet fotónov prijatých z jedného bodu sa pohybuje v desiatkach. Ich počet kolíše v rámci Poissonovho rozdelenia:

$$f(k, \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (2.1)$$

kde  $\lambda$  je očakávaný počet fotónov. Mikroskop sníma vzorku bod po bode. Dlhšie snímanie jedného bodu by oslabilo šum, ale vzorku potrebujeme nasnímať čo najrýchlejšie, kvôli zmenám, ktoré v nej nastávajú. Vplyvom detektoru je ešte ku signálu pridaný slabší aditívny Gaussov šum.

## 2.2 Dekompozícia spektrálne rozlíšených dát

Výstupom spektrálneho mikroskopu je teda sada obrázkov, každý pre rôznu vlnovú dĺžku. Vektor  $\bar{y}$ , ktorý obsahuje pixely so všetkých obrázkov s rovnakými súradnicami, budeme ďalej nazývať pixel. Každý pixel je zložený z lineárnej kombinácie farieb a šumu. Úlohou dekompozície je zistiť zastúpenie farieb v danom pixeli. Formálne to popíšeme pomocou Linear mixture model (LMM) [Li, 2002]:

LMM je popísané nasledovnou lineárnou rovnicou :

$$\bar{y} = A\bar{x}_o + \bar{e}_o \quad (2.2.1)$$

kde

$$\vec{y} = [y_1, y_2, \dots, y_N]^T \quad (2.2.2)$$

reprezentuje pozorovaný zmiešaný pixel;

$$\vec{x}_o = [x_{o1}, x_{o2}, \dots, x_{oM}]^T \quad (2.2.3)$$

reprezentuje relatívne zastúpenia farieb;

$$\vec{e}_o = [e_{o1}, e_{o2}, \dots, e_{oN}]^T \quad (2.2.4)$$

reprezentuje náhodnú chybu; a

$$A = [\vec{a}_1, \vec{a}_2, \dots, \vec{a}_M] \quad (2.2.5)$$

obsahuje spektrá farieb. Každý stĺpec obsahuje jedno spektrum

$$\vec{a}_i = [a_{i1}, a_{i2}, \dots, a_{iN}]^T \quad (2.2.6)$$

V rámci dekompozície spektrálne rozlíšených dát riešime tieto dva problémy:

- a) Výpočet vektora  $\vec{x}_o$ , pričom poznáme pozorovaný pixel  $\vec{y}$  a spektrálnu maticu  $A$ .
- b) Nepoznáme maticu  $A$  a hľadáme zastúpenie farieb vo vzorke len pomocou pozorovaných vektorov.

## 2.2.1 Lineárny unmixing

Lineárny unmixing zahŕňa metódy na riešenie prvého problému. Najpoužívanejšou metódou na zistenie hodnôt vektora  $\vec{x}_o$  je metóda najmenších štvorcov - LSE (least square estimation). Cieľom tejto metódy je, pre konkrétnu maticu  $A$  a vektor  $\vec{y}$ , nájsť optimálny odhad ( $\vec{x}_{LS}$ ) vektora  $\vec{x}_o$  ktorý minimalizuje

$$\|\vec{e}_o\|^2 = \|\vec{y} - A\vec{x}_o\|^2 \quad (2.2.7)$$

Táto rovnica sa dá prepísať na:

$$\begin{aligned} \vec{e}_o^T \vec{e}_o &= (\vec{y} - A\vec{x}_o)^T (\vec{y} - A\vec{x}_o) \\ &= \vec{y}^T \vec{y} - 2\vec{x}_o^T A^T \vec{y} + \vec{x}_o^T A^T A \vec{x}_o \end{aligned}$$

Minimum bude dosiahnuté v bode, kde bude derivácia podľa  $\vec{x}_o$  rovná 0 :

$$-2A^T \vec{y} + 2A^T A \vec{x}_o = \vec{0}$$

Za predpokladu, že existuje  $(A^T A)^{-1}$  dostaneme :

$$\bar{x}_{LS} = (A^T A)^{-1} A^T \bar{y} \quad (2.2.8)$$

Ak sú spektrá látok v matici  $A$  lineárne nezávislé, tak táto inverzná matica vždy existuje [Li, 2002].

## 2.2.2 Dekompozícia bez znalosti spektrálnej matice

Ako už bolo spomenuté v úvode, nie vždy poznáme spektrá použitých farbív. Môže to byť kvôli zmenám vlastností farbív podľa prostredia, v ktorom sa nachádzajú, kvôli autofluorescencii niektorých látok/organizmov alebo prípadne nepoznáme použité farbivá. Jednou z možností je použitie všetkých známych spektier. Počet spektier, ktoré môžeme naraz vyskúšať, je obmedzený počtom snímaných vlnových pásiem. Na korektnú dekompozíciu potrebujeme, aby spektrálna matica obsahovala spektrá všetkých fluorescenčných látok, ktoré sa vo vzorke nachádzajú. Počet vlnových pásiem, snímaných naraz mikroskopom, je príliš malý na použitie tohto postupu. Takýto postup nám tiež nepomôže, ak dané spektrá nepoznáme.

Na zistenie distribúcie jednotlivých komponentov nám už nepostačia údaje z jedného pixela. Budeme brať do úvahy celú vzorku, teda dostaneme sústavu rovníc:

$$\bar{y}_i = A\bar{x}_i + \bar{e}_i \quad (2.2.9)$$

v ktorej poznáme jedine vektory  $\bar{y}_i$ . Naším cieľom bude získať distribúcie jednotlivých komponentov, teda vektory  $\bar{x}_i$ , prípadne aj maticu  $A$ . Riešení je nekonečne veľa, preto si musíme zvoliť nejaké kritérium, podľa ktorého budeme hodnotiť jednotlivé riešenia.

Tento problém bol už podrobne skúmaný v podobnej oblasti, a to pri snímaní zeme pomocou satelitov. Najznámejšie používané metódy sú Metóda hlavných komponentov (PCA) [Hyvärinen, 1999] [Haykin] a Metóda nezávislých komponentov (ICA) [Hyvärinen, Oja, 2000] [Hyvärinen, 1999] [Haykin]. Kritériom PCA je čo najväčšia disperzia distribúcie komponentov. ICA hľadá komponenty so štatisticky nezávislými distribúciami. CLSM mikroskopia sa líši od satelitného snímania a to najmä v týchto bodoch :

1. Signál je podstatne slabší, a preto je silne ovplyvnený Poissonovým šumom.
2. Komponenty sú viac štatisticky závislé.

3. Počet snímaných vlnových pásiem – Pri snímaní zo satelitu sa tento počet pohybuje okolo 200, pričom pri konfokálnej mikroskopii to sú desiatky.
4. Spektrá farbív majú určité vlastnosti, ktoré nám umožňujú výrazne znížiť počet hľadaných riešení.

Prvé dve vlastnosti sťažujú dekompozíciu a zhoršujú výsledky, ktoré dáva algoritmus ICA. Posledné dve vlastnosti nám dávajú možnosť použiť kritéria, ktoré by pri satelitnom snímaní nefungovali, prípadne by boli výpočtovo príliš náročné.

## 2.3 Metóda hlavných komponentov (Principal component analysis - PCA)

PCA je štandardná technika, bežne používaná na redukciu dát pri rozpoznávaní štatistických vzorov a pri spracovaní signálov. Jej hlavnými cieľmi sú extrakcia príznakov a redukcia dimenzie. Základnou ideou PCA je nájdenie takých komponentov  $x_i$ , ktoré majú najvyššiu možnú disperziu. PCA môžeme definovať pomocou rekúzie. Definujeme smer prvého hlavného komponentu:

$$\bar{w}_1 = \arg \max_{\|\bar{w}=1\|} E \left\{ (\bar{w}^T \bar{y})^2 \right\} \quad (2.3.1)$$

Prvý hlavný komponent je teda projekcia do smeru, v ktorom je disperzia maximálna. Keď máme  $k$  komponentov, ďalší vypočítame ako prvý hlavný komponent zvyšku:

$$\bar{w}_{k+1} = \arg \max_{\|\bar{w}=1\|} E \left\{ \left[ \bar{w}^T \left( \bar{y} - \sum_{i=1}^k \bar{w}_i \bar{w}_i^T \bar{y} \right) \right]^2 \right\} \quad (2.3.2)$$

Hlavné komponenty sú potom dané rovnicou  $x_i = \bar{w}_i^T \bar{y}$ . Vektory  $\bar{w}_i$  sa dajú ľahko vypočítať, sú to vlastné vektory kovariančnej matice  $E \{ \bar{x} \bar{x}^T \}$ , zodpovedajúce jej najväčším vlastným číslam. Vstupný vektor transformujeme na vektor rovnakej dimenzie, ktorý má najvýznačnejšie vlastnosti zahrnuté v prvých súradniciach vektora. Takto môžeme ľahko znížiť dimenziu vektora zanedbaním posledných súradníc, pričom stratíme len tie najmenej podstatné informácie, ktoré boli zahrnuté vo vstupnom vektore. Keďže šum má malú disperziu, zníženie dimenzie sa zároveň postará o vyhladenie šumu.

V prípade slabého prekryvu komponentov a ich spektier je táto metóda dostačujúca. V biologických dátach sa komponenty silne prekrývajú a cieľom tejto diplomovej práce je nájsť algoritmus pre dekompozíciu komponentov s prekrývajúcimi sa spektrami. PCA použitá samostatne nevyhovuje našim požiadavkám. Kvôli jej schopnosti znížiť dimenziu vstupu bude použitá pri predspracovaní vstupu pre ICA.

## 2.4 Metóda nezávislých komponentov (Independent component analysis - ICA)

Vezmime si model (2.2.1) a odstráňme z neho chybový vektor:

$$\bar{y} = A\bar{x} \quad (2.4.1)$$

$$\bar{y} = \sum_{i=1}^M \bar{a}_i x_i$$

Rovnica (2.4.1) popisuje ako boli pozorované dáta generované zmixovaním komponentov  $x_i$ . Komponenty  $x_i$  a matica  $A$  sú neznáme. Všetko čo pozorujeme je náhodný vektor  $\bar{y}$ . Naším cieľom je vypočítať  $A$  a  $\bar{x}$ , s použitím čo najvšeobecnejších predpokladov.

Náš základný predpoklad bude, že komponenty  $x_i$  sú štatisticky nezávislé. Pre zjednodušenie budeme predpokladať, že matica  $A$  je štvorcová. Potom môžeme použiť maticu  $W$  inverznú k matici  $A$  a dostaneme takýto model:

$$\bar{x} = W\bar{y} \quad (2.4.2)$$

Keďže nepoznáme  $A$  ani  $\bar{x}$ , tak nemôžeme určiť disperzie nezávislých komponentov. Akýkoľvek násobok jedného zo zdrojov  $x_i$  môžeme napraviť predelením  $\bar{a}_i$  tým istým číslom. Preto potrebujeme ustáliť magnitúdu nezávislých komponentov. Najprirodzenejším riešením je predpokladať, že majú jednotkovú disperziu:  $E\{x_i^2\} = 1$ . Ešte nám ostáva nejednoznačnosť znamienka, kvôli čomu dostaneme pre každý komponent 2 výsledky. Vo väčšine aplikácií nie je problém zistiť, ktorý výsledok je správny.

### 2.4.1 Nezávislosť a nekorelovanosť

Majme 2 náhodné premenné  $s_1$  a  $s_2$ . Neformálne, premenné  $s_1$  a  $s_2$  sú nezávislé, ak informácia o hodnote  $s_1$  nedáva žiadnu informáciu o hodnote  $s_2$ .

Nezávislosť môžeme definovať pomocou hustôt náhodných premenných. Nech  $p(s_1, s_2)$  je funkcia hustoty zloženej pravdepodobnosti  $s_1$  a  $s_2$ .  $p_1(s_1)$  je funkcia hustoty premennej  $s_1$ ,  $p_2(s_2)$  pre  $s_2$ . Premenné  $s_1$  a  $s_2$  sú nezávislé práve vtedy, keď platí:

$$p(s_1, s_2) = p_1(s_1)p_2(s_2) \quad (2.4.3)$$

Definícia sa dá rozšíriť na  $n$  náhodných premenných. Z definície sa dá odvodiť nasledujúca vlastnosť náhodných premenných:

$$E\{h_1(s_1)h_2(s_2)\} = E\{h_1(s_1)\}E\{h_2(s_2)\} \quad (2.4.4)$$

Kde  $h_1$  a  $h_2$  sú ľubovoľné funkcie.

Slabšou formou nezávislosti je nekorelovanosť. Dve náhodné premenné sú nekorelované, ak majú nulovú kovarianciu:

$$E\{y_1 y_2\} - E\{y_1\}E\{y_2\} = 0 \quad (2.4.5)$$

Ak sú premenné nezávislé, sú aj nekorelované. Preto veľa ICA metód hľadá nekorelované odhady nezávislých komponentov. To redukuje počet voľných parametrov a zjednodušuje problém.

### 2.4.2 Nezávislé = nie Gaussové

Podľa centrálnej limitnej vety teórie pravdepodobnosti platí, že suma náhodných premenných, za určitých podmienok, mieri ku Gaussovej distribúcii. Suma náhodných premenných má zvyčajne distribúciu, ktorá je bližšie ku Gaussovej ako ktorákoľvek z pôvodných premenných. Preto môžeme hľadať nezávislé premenné tak, že budeme hľadať premenné čo najviac vzdialené Gaussovmu rozdeleniu.

Na výpočet ICA sa používajú aj iné kritériá. Nezávislosť sa dá počítat' aj pomocou minimalizovania vzájomnej informácie alebo metódy maximálnej pravdepodobnosti (Maximum

likelihood estimation). Nie je cieľom tejto práce poskytnúť detailnú teóriu pre ICA, preto sa budeme venovať len jednému kritériu nezávislosti.

### 2.4.3 Miery negaussianity

Aby sme mohli použiť negaussianitu na hľadanie nezávislých premenných, potrebujeme mieru negaussianity pre náhodnú premennú, napríklad  $s$ . Kvôli zjednodušeniu predpokladajme, že  $s$  má nulovú strednú hodnotu a jednotkovú disperziu.

#### 2.4.3.1 Špicatosť (Kurtosis)

Klasickou mierou negaussianity je špicatosť alebo štvrtý kumulant. Špicatosť  $s$  je definovaná nasledovne:

$$kurt(s) = E\{s^4\} - 3(E\{s^2\})^2 \quad (2.4.6)$$

Keďže sme predpokladali, že  $s$  má jednotkovú disperziu, rovnicu (2.4.6) môžeme upraviť:

$$kurt(s) = E\{s^4\} - 3 \quad (2.4.7)$$

Špicatosť je teda normalizovaná verzia štvrtého momentu. Štvrtý moment gaussovej náhodnej premennej je  $3(E\{s^2\})^2$ . Špicatosť je teda nulová pre gaussovu náhodnú premennú. Špicatosť je nenulová pre väčšinu (ale nie všetky) zo zvyšných náhodných premenných. Negaussovské premenné s nulovou špicatosťou existujú, ale je ich zanedbateľne málo.

Špicatosť, alebo skôr jej absolútna hodnota, sa často používa ako miera negaussianity. Hlavným dôvodom je jej výpočtová jednoduchosť. Špicatosť má ale v praxi určité nevýhody. Hlavnou nevýhodou je, že je veľmi citlivá na chybné dáta [Huber]. Jej hodnota môže záležať na niekoľkých hodnotách z chvosta distribúcie, ktoré môžu byť chybné alebo bezvýznamné. Inými slovami, špicatosť nie je robustnou mierou negaussianity. Preto je v niektorých prípadoch lepšie použiť iné miery negaussianity.

### 2.4.3.2 Negentropia

Entropia náhodnej premennej môže byť interpretovaná ako množstvo informácie, ktoré dostaneme pozorovaním premennej. Čím viac náhodná je premenná, tým má väčšiu entropiu. Entropia  $H$  je pre diskrétnu náhodnú premennú  $S$  definovaná takto:

$$H(S) = -\sum_i P(S = a_i) \log P(S = a_i) \quad (2.4.8)$$

kde  $a_i$  sú možné hodnoty  $Y$ . Táto definícia môže byť zovšeobecnená pre spojité náhodné premenné a vektory:

$$H(\vec{s}) = -\int f(\vec{s}) \log f(\vec{s}) d\vec{s} \quad (2.4.9)$$

kde  $\vec{s}$  je náhodný vektor s hustotou  $f(\vec{s})$ . Dôležitým výsledkom Teórie informácie je, že náhodná premenná s Gaussovou distribúciou má najväčšiu entropiu spomedzi všetkých náhodných premenných rovnakej disperzie [Cover, Thomas, 1991]. To znamená, že entropia môže byť použitá ako kritérium negaussianity. Negentropia  $J$  je definovaná nasledovne:

$$J(\vec{s}) = H(\vec{s}_{gauss}) - H(\vec{s})$$

kde  $\vec{s}_{gauss}$  je Gaussova náhodná premenná s rovnakou kovariančnou maticou ako  $\vec{s}$ .  $J(\vec{s})$  je nula pre premennú s Gaussovým rozdelením a je vždy nezáporná. Počítanie negentropie pomocou definície je pomalé, preto potrebujeme nájsť jej jednoduchšiu aproximáciu.

### 2.4.3.3 Aproximácia negentropie

Klasická metóda na aproximáciu negentropie používa momenty vyšších rádov [Jones, Sibson, 1987]:

$$J(s) \approx \frac{1}{12} E\{s^3\}^2 + \frac{1}{48} kurt(s)^2 \quad (2.4.10)$$

Táto aproximácia má podobné problémy s robustnosťou, aké má špicatnosť. Nové aproximácie, ktoré nemajú tieto problémy boli vyvinuté v [Hyvärinen, 1998]. Aproximácia je založená na princípe maximálnej entropie:

$$J(s) \approx [E\{G(s)\} - E\{G\{v\}\}]^2 \quad (2.4.11)$$



kde  $v$  je Gaussova náhodná premenná, pričom  $v$  aj  $s$  majú nulovú strednú hodnotu a jednotkovú disperziu.  $G$  môže byť skoro akákoľvek nekvadratická funkcia. Nasledovné 2 možnosti sa ukázali byť zvlášť užitočné:

$$G_1(u) = \frac{1}{a} \log \cosh au, \quad G_2(u) = -\exp(u^2/2) \quad (2.4.12)$$

kde  $1 \leq a \leq 2$ .

Získali sme miery negaussianity, ktoré sú veľmi dobrým kompromisom medzi negentropiou a špicatosťou. Sú rýchlo vypočítateľné a majú vhodné štatistické vlastnosti, najmä robustnosť.

## 2.4.4 Predspracovanie

Skôr ako prejdeme ku konkrétnemu ICA algoritmu, musíme uviesť vstupné dáta do stavu, v ktorom budú spĺňať doteraz stanovené podmienky. Najskôr odpočítame od vstupných dát vektor stredných hodnôt  $\bar{m} = E\{\bar{y}\}$ . Touto úpravou sme dosiahli, že  $\bar{y}$  a aj  $\bar{x}$  majú nulovú strednú hodnotu. Ak potrebujeme znížiť dimenziu vstupu, môžeme použiť už spomenutú metódu PCA. O splnenie zvyšných podmienok sa postará bielenie (whitening).

Bielenie transformuje  $\bar{y}$  tak, že jeho komponenty budú nekorelované a budú mať jednotkovú disperziu. Inými slovami, kovariančná matica  $\bar{y}$  bude matica identity:

$$E\{\bar{y}\bar{y}^T\} = I \quad (2.4.13)$$

Táto transformácia je vždy možná. Populárnou metódou je použitie dekompozície kovariančnej matice  $E\{\bar{y}\bar{y}^T\}$  na vlastné hodnoty (eigen-value decomposition, EVD):  $E\{\bar{y}\bar{y}^T\} = EDE^T$ , kde  $E$  je ortogonálna matica vlastných vektorov a  $D$  je diagonálna matica vlastných hodnôt. Vstup transformujeme nasledovným spôsobom:

$$\bar{y}_{new} = ED^{-1/2}E^T x \quad (2.4.14)$$

kde  $D^{-1/2} = \text{diag}(d_1^{-1/2}, \dots, d_n^{-1/2})$ .

## 2.4.5 Algoritmus FastICA

V časti 2.4.3.3 sme si definovali funkciu (2.4.11), ktorá bude našou mierou nezávislosti komponentov. Vstupné dáta sme si predspracovali do požadovaného stavu. Teraz už potrebujeme len algoritmus, ktorý nájde maximum tejto funkcie.

Odvodenie konkrétneho algoritmu nie je pre túto prácu dôležité a vyžaduje znalosti, ktoré nie sú nikde inde v tejto práci potrebné. Preto tu len spomenieme ako algoritmus funguje. Odvodenie algoritmu sa nachádza v [Hyvärinen, Oja, 2000].

Vytvoríme si maticu  $W$  rozmerov  $M \times M$ , s náhodnými koeficientmi. Maticu  $W$  symetricky zortogonalizujeme:

1.  $W = W / \sqrt{\|WW^T\|}$
2.  $W = \frac{3}{2}W - \frac{1}{2}WW^TW$

Krok 2. opakujeme kým matica skonverguje. Potom opakujeme nasledovnú iteráciu, až kým matica neskonverguje:

$$W = W + \text{diag}(\alpha_i) [\text{diag}(\beta_i) + E\{g(\bar{s})\bar{s}^T\}]W$$

kde:

$$\bar{s} = W\bar{y}$$

$$\beta_i = -E\{s_i g(s_i)\}$$

$$\alpha_i = -1/(\beta_i - E\{g'(s_i)\})$$

$$g(u) = u \exp(-u^2 / 2)$$

Matica  $W$  je po každom kroku symetricky ortogonalizovaná.

## 2.5 Genetický algoritmus

V štvrtej kapitole budeme potrebovať optimalizačný algoritmus. Budeme používať genetický algoritmus. Pre pochopenie daného textu sú nutné aspoň minimálne znalosti tohto algoritmu, preto sa v tejto časti venujem jeho stručnému popisu.

Darwinova teória evolúcie [Darwin, 1989] sa zakladá na téze prirodzeného výberu, podľa ktorej prežívajú len najlepšie prispôsobení jedinci populácie. Reprodukciou dvoch silných jedincov dostaneme potomkov, ktorí s vysokou pravdepodobnosťou budú dobre prispôsobení k úspešnému prežitiu. Pri podrobnej analýze sa ukazuje, že samotné pôsobenie reprodukcie nie je dostatočne efektívne na vznik dobre prispôsobených jedincov s novými vlastnosťami, ktoré významne uľahčujú prežitie. Do evolúcie živej hmoty je nutné zapojiť aj mutáciu, ktorá náhodným spôsobom ovplyvňuje genetický materiál populácie jedincov.

Zavedieme si pojem sily (fitness). V biológii je sila definovaná ako relatívna schopnosť prežitia a reprodukcie jedinca v danom prostredí. Podobne sa chápe aj v umelom živote, je to kladné číslo priradené genetickej informácii reprezentujúcej organizmus, ktoré reprezentuje jeho relatívnu úspešnosť pri plnení zadanej úlohy.

Pojem jedinec nahradíme pojmom chromozóm, ktorý reprezentuje lineárnym usporiadaným spôsobom informačný obsah jedinca. Potom hovoríme o populácii chromozómov, ktoré sa reproduktujú s pravdepodobnosťou úmernou ich sile, pričom súčasťou tejto reprodukcie sú mutácie. Mutácie zanášajú do chromozómov novú informáciu, ktorá môže zvýšiť silu chromozómov vznikajúcich reprodukciou pôvodných chromozómov. Nové chromozómy vytesňujú z populácie chromozómy s malou silou. Tento základný reprodukčný cyklus sa v danej populácii chromozómov neustále opakuje. Po určitom čase sa teda v populácii s vysokou pravdepodobnosťou objavia chromozómy s novými vlastnosťami, ktoré podstatne zvyšujú ich silu a vytesňujú pôvodné chromozómy bez týchto vlastností.

Pristúpme k formalizácii základných pojmov evolúcie. Chromozómy sú binárne vektory fixnej dĺžky  $k$ :

$$\alpha \in \{0,1\}^k \quad (2.5.1)$$

Populácia  $P$  je množina týchto chromozómov:

$$P = \{\alpha_1, \alpha_2, \dots, \alpha_p\} \quad (2.5.2)$$

Kde  $p$  vyjadruje počet chromozómov v  $P$ . Nech  $f$  je účelová funkcia definovaná nad množinou binárnych vektorov dĺžky  $k$ :

$$f : \{0,1\}^k \rightarrow R \quad (2.5.3)$$

Našou úlohou bude hľadať globálne minimum tejto funkcie. Funkcia  $f$  reprezentuje prostredie v ktorom existujú chromozómy populácie. Vyjadrené pomocou biologickej terminológie,

chromozóm  $\alpha$  reprezentuje genotyp organizmu, zatiaľ čo funkčná hodnota  $f(\alpha)$  reprezentuje jeho fenotyp. Mierou úspešnosti chromozómu je jeho funkčná hodnota. Pretože hľadáme minimum účelovej funkcie, chromozóm je tým úspešnejší, čím je jeho funkčná hodnota menšia.

Každý chromozóm  $\alpha \in P$  ohodnotíme silou, ktorá je interpretovaná ako zobrazenie chromozómov na kladné reálne čísla:

$$F : P \rightarrow R_+ \quad (2.5.4)$$

$F$  musí spĺňať nasledujúcu podmienku:

$$\forall \alpha_i, \alpha_j \in P : f(\alpha_i) \leq f(\alpha_j) \Rightarrow F(\alpha_i) \geq F(\alpha_j) \quad (2.5.5)$$

Zobrazenie  $f$  na  $F$  sa dá uskutočniť mnohými spôsobmi, ukážeme si jeden z nich. Je to lineárne zobrazenie funkčných hodnôt na silu tak, že maximálnej (minimálnej) funkčnej hodnote je priradená minimálna (maximálna) sila, ostatné funkčné hodnoty sú lineárne interpolované medzi týmito dvoma limitnými hodnotami:

$$F(\alpha) = \frac{1}{f_{\min} - f_{\max}} [(1 - \varepsilon)f(\alpha) + f_{\min}\varepsilon - f_{\max}] \quad (2.5.6)$$

$\varepsilon$  je malé kladné číslo (napr.  $\varepsilon = 0.01$ ). Z praktických dôvodov je vhodné normalizovať  $F$  tak, aby platilo

$$\sum_{\alpha \in P} F(\alpha) = 1 \quad (2.5.7)$$

Takto bude  $F(\alpha)$  udávať pravdepodobnosť výberu  $\alpha$  z  $P$  pri kvázínáhodnom výbere.

Proces reprodukcie chromozómov začína tým, že z populácie sú vybrané dva chromozómy kvázínáhodne v závislosti na ich sile (chromozómy s väčšou silou majú väčšiu šancu byť vybrané). Pod reprodukciou budeme rozumieť taký proces, v ktorom sa dva pôvodné chromozómy  $\alpha_1$  a  $\alpha_2$  reprodujú na dva nové chromozómy  $\alpha'_1$  a  $\alpha'_2$ :

$$(\alpha'_1, \alpha'_2) = O_{repro}(\alpha_1, \alpha_2) \quad (2.5.8)$$

Tento operátor obsahuje dve časti, a to kríženie a mutáciu. Tieto operácie sa vykonávajú len s určitou pravdepodobnosťou (to umožňuje zachovanie pôvodných chromozómov). Operátorom reprodukcie vytvoríme toľko chromozómov, koľko mala pôvodná populácia. Starú populáciu sa potom nahradí novou.

Operátor kríženia je zobrazenie, ktoré dvojici chromozómov  $\alpha = (a_1, \dots, a_k)$ ,  $\beta = (b_1, \dots, b_k)$  priradí dva nové chromozómy o rovnakej dĺžke ako pôvodné:

$$(\alpha', \beta') = O_{cross}(\alpha, \beta) \quad (2.5.9)$$

V základnej verzii vyzerajú nové chromozómy takto:

$$\alpha' = (a_1, \dots, a_j, b_{j+1}, \dots, b_k), \quad \beta' = (b_1, \dots, b_j, a_{j+1}, \dots, a_k) \quad (2.5.10)$$

kde  $j$  je náhodne generované celé číslo z intervalu  $\langle 1, k-1 \rangle$ .

Operátor mutácie zmení chromozóm  $\alpha = (a_1, \dots, a_k)$  tak, že pre každé  $i \in \{1..k\}$  sa s pravdepodobnosťou  $P_{mut}$  invertuje hodnota  $a_i$ . Hodnota  $P_{mut}$  sa zvyčajne pohybuje okolo 0,05.

Keďže prostredie (funkcia  $f$ ) sa nemení, môže byť vhodné hodnotu  $P_{mut}$  postupne znižovať.

V prípade že chromozómy sú vektory reálnych čísel, môže byť mutácia realizovaná pomocou pripočítania náhodného čísla s Gaussovou distribúciou pravdepodobnosti.

Táto časť bola spracovaná podľa [Kvasnicka a kol., 2000].

## **3. Kapitola**

### **Ciele diplomovej práce**

Cieľom tejto diplomovej práce je navrhnúť vhodné algoritmy na dekompozíciu spektrálne odlišených fluorescenčných dát, ktoré boli získané konfokálnym laserovým mikroskopom. Prvou úlohou je získať prehľad o metódach dekompozície spektrálne odlišených dát a zistiť, či je niektorá z nich vhodným riešením. V prípade, že existujúce metódy neposkytujú dostatočne dobré riešenie problému, je úlohou navrhnúť nový postup, ktorý by riešil problém dekompozície spektrálne rozlíšených fluorescenčných dát. Tento algoritmus je potom potrebné prispôbiť pre paralelný beh a implementovať.

## 4. Kapitola

### Výsledky a diskusia

Algoritmus FastICA dáva veľmi dobré výsledky za predpokladu, že distribúcie komponentov vo vzorke sú štatisticky nezávislé. Komponenty v biologických dátach nie sú náhodné distribúcie. Sú distribuované podľa nejakých zákonitostí. Ak majú komponenty tendenciu vyskytovať sa na rovnakých miestach, alebo sa navzájom vyhýbať, tak sa znižuje ich štatistická nezávislosť. Preto potrebujeme nové kritérium, podľa ktorého budeme hodnotiť riešenia sústavy rovníc (2.2.9).

Ak poznáme  $\bar{y}_i$  a maticu  $A$ , tak lineárny unmixing k nim jednoznačne priradí vektory  $\bar{x}_i$  a  $\bar{e}_i$ . Sústavu (2.2.9) si teda môžeme upraviť tak, že jedinými neznámymi budú koeficienty matice  $A$ . Týmto sme výrazne obmedzili priestor riešení. Riešení je ešte stále priveľa, keďže každá matica, ktorá má lineárne nezávislé stĺpce, je riešením rovnice. Potrebujeme s pomedzi nich, pomocou nejakého kritéria, vybrať tú najlepšiu.

#### 4.1 Kritériá

V tejto časti si zadefinujeme niekoľko funkcií, ktorých minimalizovaním budeme hľadať spektrálnu maticu  $A$ .

Majme funkciu  $f_1$  z priestoru všetkých matic rozmerov  $N \times M$  s reálnymi koeficientmi do nezáporných reálnych čísel:

$$f_1(A) = \sum_{i=1}^n \vec{e}_i^T \vec{e}_i \quad (4.1.1)$$

kde  $n$  je počet vstupov a

$$\vec{e}_i = \vec{y}_i - A\vec{x}_{LS\_i}$$

$$\vec{x}_{LS\_i} = (A^T A)^{-1} A^T \vec{y}_i$$

$f_1(A)$  teda udáva súčet chýb, ktoré vzniknú pri počítaní výstupu pomocou metódy LSE, s využitím matice  $A$ . Hodnoty  $y_{ij}$  po načítaní z obrázkov nijak neupravujeme, preto sa pohybujú medzi 0 a 255.

Druhá funkcia bude udávať sumu maximálnych chýb v jednotlivých chybových vektoroch:

$$f_2(A) = \sum_{i=1}^n \max(e_{ij}^2) \quad (4.1.2)$$

Stredná hodnota chybových vektorov by mala byť  $\vec{0}$ . Preto skúsme hľadať taký výsledok, ktorý je k tomu najbližšie:

$$f_3(A) = \sum_{j=1}^N \text{abs}(E_j\{\vec{e}\}) \quad (4.1.3)$$

$$E_j\{\vec{e}\} = \sum_{i=1}^n e_{ij}$$

Zatiaľ sme využívali len vlastnosti chybových vektorov. V poslednej funkcii využijeme aj nezávislosť vektorov  $\vec{x}_i$ . Nemá zmysel hľadať výsledok len na základe ich nezávislosti, keďže tento problém rieši už ICA. Spojíme výsledok  $f_1(A)$  s hodnotou špicatosti výstupov. Keďže špicatosť potrebujeme maximalizovať, použijeme jej prevrátenú hodnotu:

$$f_4(A) = f_1(A) \frac{1}{\sum_{i=1}^M \text{kurt}(x_i)} \quad (4.1.4)$$

Konkrétne algoritmy budeme označovať pomocou funkcie, ktorú využívajú.

Počítať funkciu  $f$  pre celú sústavu je zdĺhavé, žiadny optimalizačný algoritmus by nenašiel riešenie v rozumnom čase. Pri počítaní  $f$  pomocou malej vzorky, bude hodnota  $f$  výrazne kolísat'. Preto potrebujeme optimalizačný algoritmus, ktorý nebude ovplyvnený týmito výkyvmi. Bolo by tiež vhodné, ak by sa dal upraviť pre výpočet na paralelných procesoroch. Spomedzi mne známych algoritmov tieto vlastnosti najlepšie splňa genetický algoritmus.



## 4.2 Všeobecné spektrá

### 4.2.1 Implementácia genetického algoritmu

V prehľadovej kapitole sme popísali všeobecný genetický algoritmus. Teraz ho potrebujeme prispôbiť na riešenie nášho problému. Chromozómy sú matice  $A_i = [\vec{a}_{i1}, \vec{a}_{i2}, \dots, \vec{a}_{iM}]$  rozmerov  $N \times M$  s reálnymi koeficientmi. Prenásobenie stĺpca matice nenulovým číslom nemá vplyv na výsledok, preto budeme stĺpce matíc udržiavať v normovanom tvare:  $\|\vec{a}_{ij}\| = 1$ . Na fungovanie metódy LSE potrebujeme, aby boli stĺpce matice lineárne nezávislé. Je málo pravdepodobné, že počas výpočtu vznikne matica s lineárne závislými stĺpcami, preto tento problém ošetríme len pomocou nulovej hodnoty sily pre takúto maticu.

Vždy na začiatku epochy vyberieme podmnožinu sústavy (2.2.9) a budeme určovať silu chromozómov pomocou nej. Minimálny počet používaných vstupov stanovíme na 100. Zvýšenie tohto počtu zvýši presnosť a zníži pravdepodobnosť uviaznutia v lokálnom minime, ale tiež lineárne zvýši čas behu algoritmu. Na výpočet sily chromozómov použijeme funkciu podobnú funkcii (4.2). Najskôr zmeníme hodnotu  $f_{\max}$ :

$$f_{\max} = C_1 * f_{\min} + (f_{\max} - f_{\min}) / C_2 \quad (4.2.1)$$

Zmeníme hodnotu  $f$  na  $f_{\max}$  všetkým chromozómom, ktoré ju majú vyššiu. Potom vypočítame silu  $F$  pomocou (2.5.6). Ak by sme nechali pôvodnú funkciu, príliš vysoké hodnoty  $f_{\max}$  by spôsobili, že zvyšné chromozómy budú mať skoro rovnakú silu.  $C_1$  môžeme nastaviť napríklad na 1.5. Druhá časť vzorca je pre prípad, keď je minimálna hodnota príliš nízka, čo môže nastať pri slabom šume.  $C_2$  nastavíme na 60.

Kríženie bude tiež trochu pozmenené. Každý prvok chromozómu bude mať 50% šancu na výmenu s korešpondujúcim prvkom v druhom chromozóme.

Koeficienty chromozómov sú reálne čísla, takže pri mutácii k nim budeme pripočítavať náhodné číslo. Náhodné číslo bude mať Gaussovu distribúciu pravdepodobnosti, s nulovou strednou hodnotou, jednotkovou disperziou a bude prenasobené číslom:

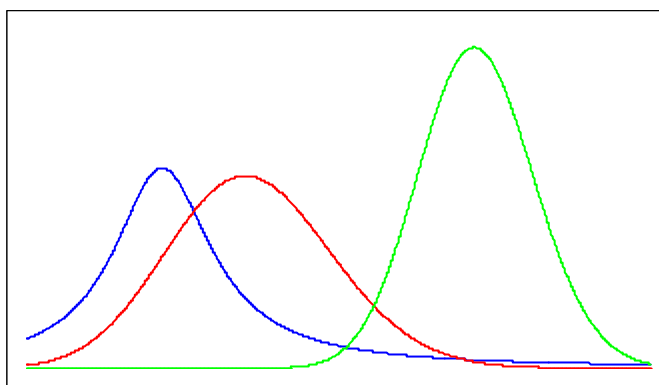
$$0.03 * [1 - (epocha / pocet\_epoch)] \quad (4.2.2)$$

Vysoká mutácia na začiatku slúži k nasmerovaniu správnym smerom a bráni uviaznutiu v lokálnom minime. Nízka mutácia na konci slúži k doladeniu.

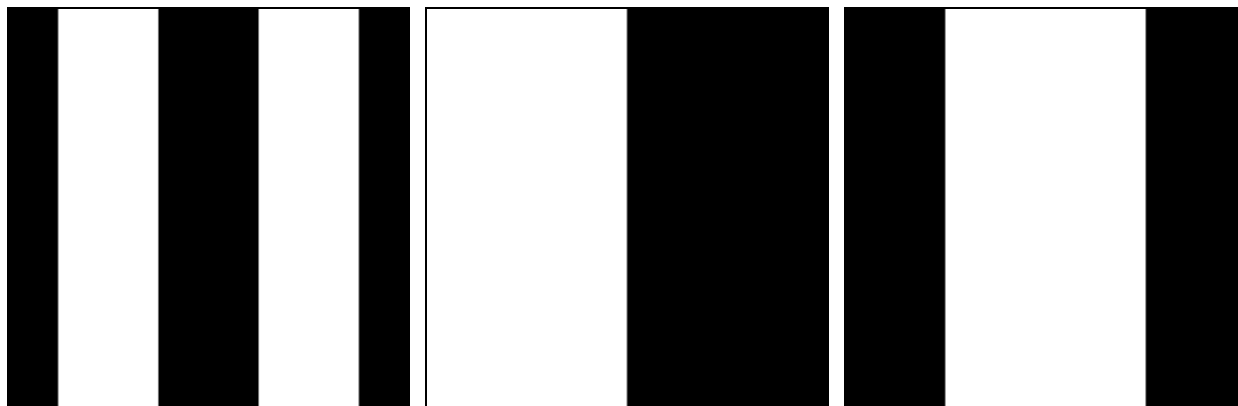
Aby chromozómy stihli skonvergovať, potrebujeme nastaviť počet epoch na aspoň 50 a počet chromozómov na 100.

## 4.2.2 Test algoritmu

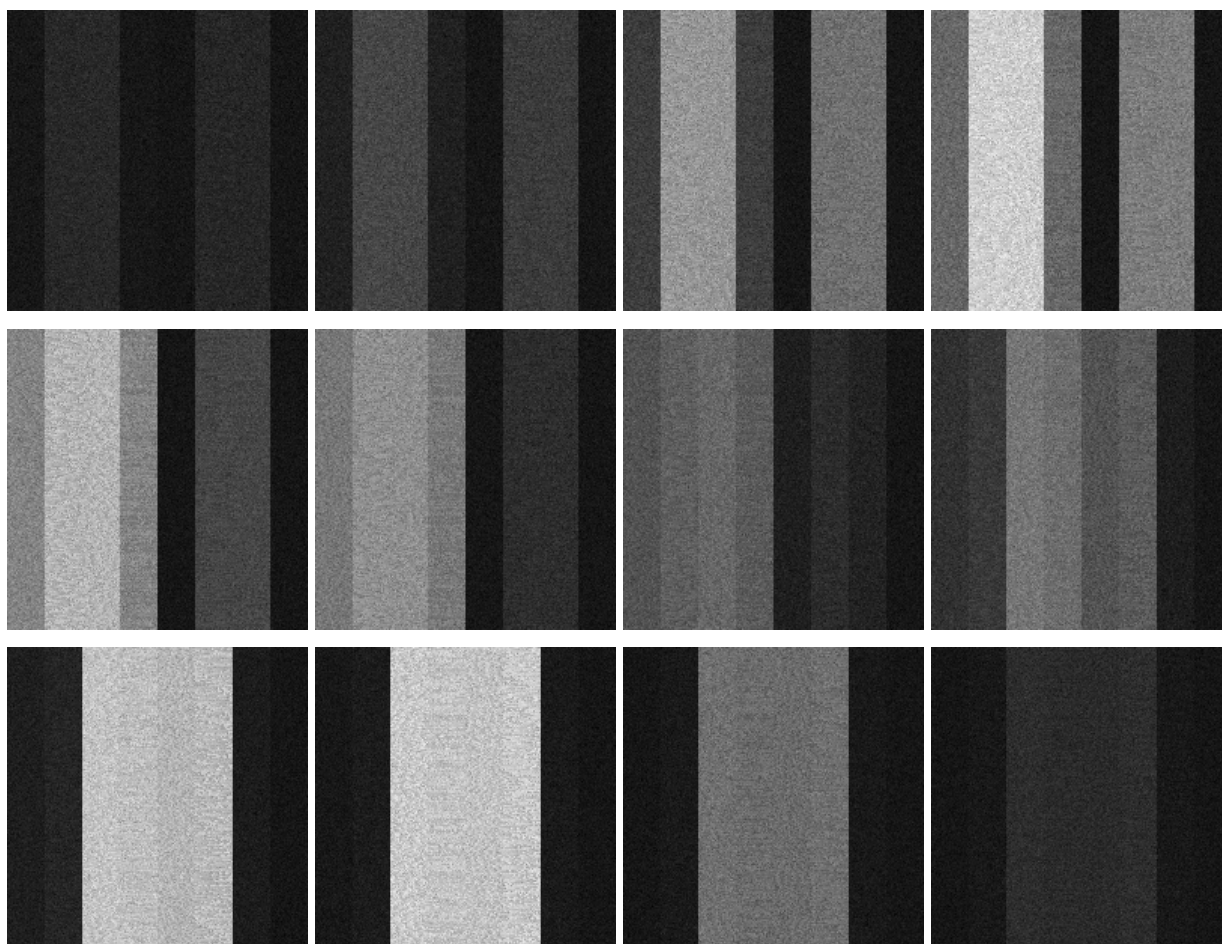
Testovať budeme na umelých dátach (obr. 3), ktoré boli vyrobené tak, aby boli štatisticky nezávislé. To nám umožní porovnať výsledky nových postupov s metódou ICA, pri splnení jej predpokladov. Lineárnou kombináciou komponentov na obr. 3, so spektrami na obr. 2, a pridaním šumu, získame vstupy na obr. 4. Poissonov šum bude generovaný za predpokladu, že maximálny počet očakávaných fotónov z jedného bodu je 100. Na ohodnotenie výsledku algoritmu použijeme 2 kritériá. Prvým bude rozdiel medzi jednotlivými komponentmi a im prislúchajúcimi výsledkami. Konkrétne to bude pomer sily signálu tohto rozdielu a sily signálu pôvodného komponentu. Druhým kritériom bude výsledná hodnota funkcie, ktorú daný algoritmus minimalizuje. Tú porovnáme s hodnotou pre pôvodné spektrá. Každý algoritmus pustíme desať krát a vyberieme výsledok s najnižšou hodnotou minimalizovanej funkcie.



Obr. č. 2: Použité spektrá. Spektrá sú zobrazené na vlnových dĺžkach od 500 do 640 nanometrov.



Obr. č. 3: Nezávislé komponenty



Obr. č.4: Vstupy pre prvý test. Sú lineárnou kombináciou komponentov z obr. 3 a spektier z obr. 2. Poissonov šum generovaný za predpokladu, že signál dosahuje silu 100 fotónov.

		$f(\min)$	$f(A_S)$
$f_1$	38.8%	2.95 E7	4.11 E7
$f_2$	37.6%	1.18 E7	1.39 E7
$f_3$	42.7%	1.2 E6	2.2 E6
$f_4$	34%	5.7 E6	7.2 E6
LSE	29.7%	-	-
ICA	34.2%	-	-
PCA	76%	-	-

Tabuľka č. 1 : Výsledky prvého testu. V prvom stĺpci je vyššie popísaný pomer. V druhom je nájdená minimálna hodnota používanej funkcie. V treťom stĺpci je hodnota funkcie pre použitú spektrálnu maticu. Funkcie uvedené v prvom stĺpci označujú algoritmus, ktorý minimalizuje túto funkciu. Funkcie boli definované v časti 4.1 a ďalej na ne budeme odkazovať týmito názvami.

Percentá nevyjadrujú priamo prekryvanie komponentov, ich hodnota je silne ovplyvnená šumom. Vo výsledku s 30%-ným ohodnotením sa komponenty neprekývajú. Až na minimalizovanie  $f_3$  a algoritmus PCA sú komponenty dobre rozlíšiteľné. Minimalizovanie  $f_3$  malo viesť k výsledku so strednou hodnotou chyby blízkou nule. Výsledky ukazujú, že táto stratégia nie je vhodná. V ďalších testoch sa jej už nebudeme venovať. To isté platí pre algoritmus PCA, ktorý dáva veľmi zlé výsledky.

Výsledky v posledných dvoch stĺpcoch naznačujú, že  $f(A_S)$ , kde  $A_S$  je spektrálna matica použitá pri lineárnej kombinácii komponentov, nie je minimom danej funkcie. My potrebujeme takú funkciu, ktorá bude mať v  $A_S$  globálne minimum.

### 4.3 Ohraničené spektrá

Presunutie minima do bodu  $A_S$  sa pokúsime dosiahnuť tak, že znížime definičný obor funkcií. Ako sme spomenuli v časti (2.2.2), nami používané dáta majú oproti satelitným jednu výhodu.

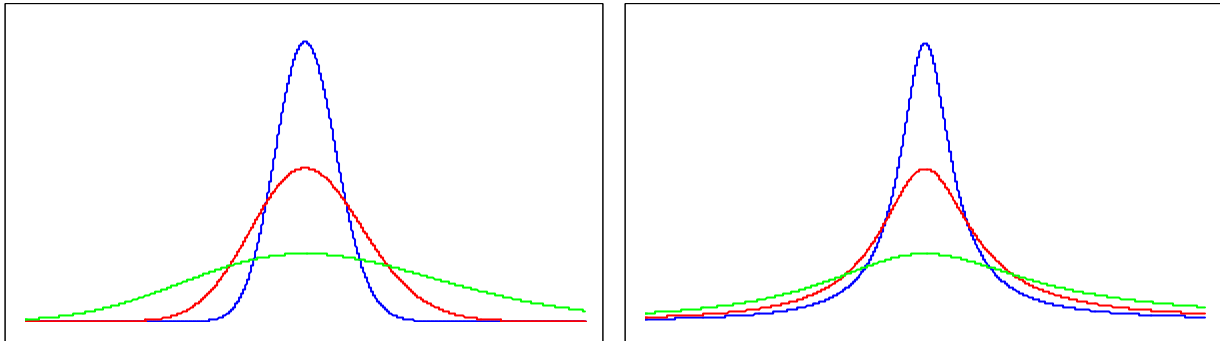
Tou je, že poznáme určité vlastnosti spektier fluorescenčných farbív. Spektrá majú zvyčajne tvar hustôt Gaussovho:

$$f(x; \eta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\eta)^2}{2\sigma^2}} \quad (4.3.1)$$

alebo Lorentzovho náhodného rozdelenia:

$$f(x; \eta, \sigma) = \frac{1}{\pi\sigma \left[ 1 + \left( \frac{x-\eta}{\sigma} \right)^2 \right]} \quad (4.3.2)$$

$x$  v týchto rovniciach nie je vlnová dĺžka, ale frekvencia svetla. Vyžarovanie farbiva po osvetlení svetlom vlnovej dĺžky  $\lambda$  má teda silu  $f\left(\frac{c}{\lambda}; \mu, \sigma\right)$ , kde  $c$  je rýchlosť svetla. Výsledok, pre vlnové dĺžky svetla v rozmedzí 500 až 700 nanometrov, je len mierne odlišný od pôvodných rozdelení (obr. č. 5).



Obr. č. 5: Ukážky hustôt Lorentzovho a Gaussovho náhodného rozdelenia na vlnových dĺžkach od 500 do 700 nm . Rovnaké farby označujú rozdelenia s rovnakými hodnotami  $\eta$  a  $\sigma$

### 4.3.1 Modifikácia algoritmu

Ak predpokladáme, že spektrá majú takýto tvar, nemusia byť chromozómy z priestoru všetkých matíc rozmerov  $N \times M$ . Jeden chromozóm bude tvoriť  $M$  trojíc  $(\eta, \sigma, typ)$ . Typ bude 1 pre Gaussovo rozdelenie a 0 pre Lorentzovo.  $\eta$  a  $\sigma$  majú dosť podobnú funkciu v rámci týchto dvoch rozdelení, takže pri zmene typu nemusíme meniť prvé dve hodnoty.

Kríženie bude pozostávať z dvoch častí. Najskôr sa s 50 percentnou pravdepodobnosťou vymenia rozdelenia, ktoré sú v chromozómoch na rovnakých miestach. Potom si ešte s pravdepodobnosťou 10% vymenia hodnoty  $\sigma$ . Rozdelenia sa medzi sebou krížia málo, lebo výmena jednej hodnoty má príliš veľký vplyv a na nájdenie správnych rozdelení postačí mutácia. Kríženie má takto na starosti nájdenie správnej  $M$ -tice rozdelení.

Mutácia vyzerá podobne ako v predchádzajúcej časti. Pre  $\eta$  a  $\sigma$  treba nastaviť rôzne sily mutácie, aby vyhovovali rozsahu ich hodnôt. Navyše im pridáme malú šancu (napríklad jedno percento), že budú nahradené novou náhodnou hodnotou. Zmutovať bude môcť aj typ. Zmení sa s päť percentnou pravdepodobnosťou.

### 4.3.2 Test algoritmu

Potrebujeme otestovať, či po zmenšení definičného oboru minimalizovaných funkcií je ich minimum v bode  $A_S$ . Testovacie dáta pripravíme pomocou ôsmich rôznych spektrálnych matic. Pre spektrálnu maticu pripravíme 3 verzie vstupov, ktoré sa budú odlišovať v sile šumu. V prvej verzii bude signál dosahovať silu 150 fotónov a Gaussov šum nebude prítomný. V ďalších verziách bude sila signálu znížená na polovicu oproti predošlej verzii a budú ovplyvnené aj primeraným Gaussovým šumom. Každý vstup bude tvorený pätnástimi obrázkami. Obrázky budú snímané v rozmedzí 500 až 640 nm, s odstupom 10 nm medzi jednotlivými obrázkami. Riadkový stupeň matic bude teda 15.

Pre každý vstup bude algoritmus spustený desať krát, v tabuľke bude uvedený minimálny výsledok. Okrem toho bude v tabuľke uvedená hodnota funkcie v  $A_S$ .

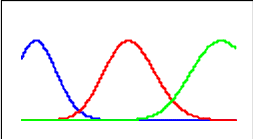
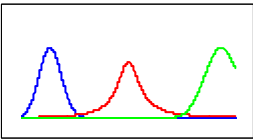
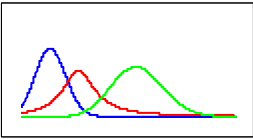
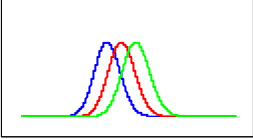
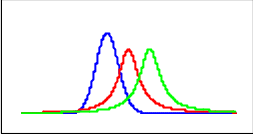
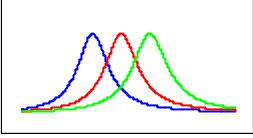
Výsledky sa nachádzajú v tabuľke č. 2. Vo výsledkoch je vidieť, že  $A_S$  je minimum len pre vstupy bez Gaussovho šumu. Zo zvyšných výsledkov je zrejme, že zmenšenie definičného oboru funkcií, nám v tomto smere dostatočne nepomohlo. Problémom je zrejme Gaussov šum. Výsledné spektrá majú oproti pôvodným rovnakú strednú hodnotu, ale vyššiu disperziu. Tým sa snažia pokryť aj miesta, kde sa nachádza len Gaussov šum. Pôvodné spektrá nechávali veľa takéhoto priestoru, preto je suma výsledného šumu vyššia.

Spektrá/funkcie	$f_1 * E07$	$f_1(A_S) * E07$	$f_2 * E07$	$f_2(A_S) * E07$	$f_4 * E06$	$f_4(A_S) * E06$
	4,65	4,6	1,68	1,72	7,92	7,96
	7,66	9,53	2,70	3,23	13,1	16,3
	13,2	19,8	4,43	6,09	23,1	34,7
	1,37	1,25	0,58	0,55	2,22	2,11
	3,77	5,31	1,08	1,37	6,37	9,05
	6,69	12,3	1,90	3,01	11,2	21,5
	1,98	1,93	0,76	0,76	3,32	3,28
	3,43	3,99	1,28	1,46	5,95	6,93
	6,49	8,33	2,41	3,01	12,1	15,0
	0,20	0,19	0,09	0,09	0,36	0,34
	1,05	1,45	0,31	0,34	2,01	2,65
	2,42	4,75	0,71	1,04	5,07	9,33
	0,64	0,54	0,27	0,25	1,03	0,93
	1,98	2,74	0,61	0,73	3,64	4,77
	4,30	7,28	1,35	1,86	9,00	13,2
	1,75	1,72	0,71	0,71	3,02	2,98
	3,96	4,90	1,29	1,50	7,29	8,84
	7,77	11	2,63	3,30	16,1	21,5
	3,25	3,25	1,26	1,28	6,24	6,19
	5,41	7,37	2,08	2,62	14,4	21,3
	9,83	15	3,59	5,21	26,3	51,8
	2,21	2,20	0,89	0,89	4,51	4,48
	3,60	4,49	1,37	1,55	8,49	11,2
	6,75	9,25	2,48	3,13	17,2	25,5

Tabuľka č. 2: Každá trojica riadkov predstavuje výsledky algoritmov pre vstupy, ktoré boli generované pomocou spektrier v prvom stĺpci. Vstupy v jednotlivých riadkoch sa líšia silou použitého šumu. V stĺpcoch sú výsledky pri použití jednotlivých funkcií, a tiež hodnoty týchto funkcií pre spektrálnu maticu.

## 4.4 Uniformné spektrum

Aby sme zabránili rozťahovaniu sa spektier, potrebujeme pri výpočte brať do úvahy Gaussov šum. Na Gaussov šum sa môžeme pozerat' ako na komponent, ktorý má uniformné spektrum. Gaussov šum, na rozdiel od Poissonovho, je aditívny, a teda takýto prístup nijak nenaruša linearitu modelu (2.2.1). Pridanie uniformného spektra by malo odstrániť problém s rozťahovaním sa spektier. Túto hypotézu otestujeme rovnakým testom, aký sme použili v časti 4.3.2).

Spektrá/funkcie	$f_1 * E07$	$f_1(A_S) * E07$	$f_2 * E07$	$f_2(A_S) * E07$	$f_4 * E06$	$f_4(A_S) * E06$
	4,34	4,34	1,58	1,61	6,41	6,45
	6,26	6,24	2,37	2,38	10,1	10,0
	10,5	10,5	3,78	3,81	18,3	18,2
	1,23	1,20	0,54	0,52	1,77	1,72
	1,89	1,87	0,74	0,73	2,96	2,93
	3,46	3,43	1,32	1,33	5,50	5,69
	1,84	1,81	0,71	0,71	2,79	2,79
	2,61	2,60	1,05	1,05	3,94	3,92
	4,98	4,97	1,96	1,98	8,67	8,68
	0,19	0,18	0,09	0,09	0,30	0,29
	0,37	0,37	0,17	0,17	0,64	0,64
	0,69	0,70	0,30	0,30	1,31	1,33
	0,55	0,51	0,24	0,23	0,86	0,78
	0,96	0,94	0,40	0,39	1,60	1,54
	1,62	1,60	0,65	0,64	2,79	2,76
	1,63	1,62	0,68	0,68	2,55	2,53
	2,27	2,26	0,90	0,90	3,74	3,76
	4,51	4,50	1,78	1,80	8,32	8,40



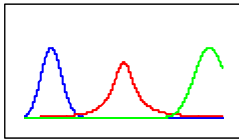
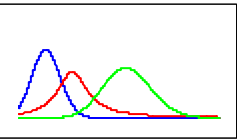
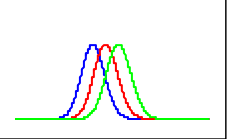
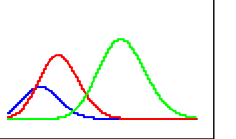
	3,04	3,04	1,15	1,17	5,29	5,25
	4,39	4,38	1,71	1,72	8,78	8,81
	7,93	7,94	2,97	3,01	16,6	18,3
	2,12	2,11	0,85	0,85	3,68	3,67
	3,10	3,09	1,23	1,23	6,10	6,14
	5,69	5,68	2,20	2,20	11,6	12,9

Tabuľka č. 3: Štruktúra je rovnaká ako v tabuľke č. 2.  $A_S$  obsahuje navyše uniformné spektrum.

Z údajov v tabuľke je vidieť, že pridaním uniformného spektra sa situácia výrazne vylepšila. Najlepšie dopadla funkcia  $f_1$ , pri ktorej bol výsledok menší ako  $f_1(A_S)$  len v dvoch prípadoch, a aj to len s minimálnym rozdielom. U zvyšných dvoch sa takéto prípady vyskytujú častejšie a pri  $f_4$  je výsledok v niektorých prípadoch výrazne nižšie.

Pridaním uniformného spektra vznikol jeden nový problém. Občas sa stane, že výsledné spektrum je približne rozdielom uniformného a pôvodného spektra. Vtedy sa výsledný komponent priradí uniformnému spektru. Chyba ktorá takto nastala sa zníži tým, že výslednému spektru sa priradí inverzný komponent. Tomu by sa dalo zabrániť tým, že by sme použili modifikovanú metódu LSE, ktorá by priraďovala hodnotám  $x_{ij}$  len kladné čísla. Takáto metóda je ale príliš pomalá, preto sa viac oplatí pri chybnom výsledku spustiť algoritmus odznova. Tento problém je výraznejší pri použití všeobecných spektier. Vtedy spektrá nie sú obmedzené tvarom, preto vždy existuje inverzné spektrum. Vyriešime to tým, že záporné hodnoty  $x_{ij}$  nahradíme nulou. Má to nevýhodu v tom, že minimum bude nižšie ako  $f(A_S)$ .

To, že má funkcia v bode  $A_S$  minimum, nestačí na to, aby funkcia bola dobrým kritériom. Potrebujeme otestovať ako sa výsledné spektrá a komponenty približujú k pôvodným. Na testovanie použijeme štyri vstupy z predchádzajúceho testu, ktoré dosahujú silu signálu 75 fotónov. Pôvodné aj výsledné spektrá predelíme najvyššou hodnotou, ktorú dosahujú. Takto budú hodnoty spektra z intervalu  $\langle 0,1 \rangle$ . Bodov, v ktorých spektrá meriame, je 15. Výsledok budeme počítať ako súčet absolútnych hodnôt rozdielov pôvodných a výsledných spektier v týchto bodoch. Druhý test bude rovnaký ako sme použili 4.2.2.

								
	spektrá	komp.	spektrá	komp.	spektrá	komp.	spektrá	komp.
$f_1$	0,1	28,2%	0,26	33,0%	2,56	57,0%	1,31	43,0%
$f_2$	0,14	28,0%	0,52	33,7%	1,35	55,2%	1,41	45,0%
$f_4$	0,86	28,2%	1,87	35,6%	0,18	42,6%	0,21	37,1%
$f_1$ -vše.	0,83	29,4%	1,35	37,4%	1,66	45,0%	2,11	39,4%
ICA	-	30,4%	-	40,1%	-	55,7%	-	41,5%
LSE	-	28,3%	-	33,3%	-	42,9%	-	37%

Tabuľka č. 4: Výsledky testu opísaného na predchádzajúcej strane.  $f_1$ -vše. označuje minimalizovanie  $f_1$  pomocou všeobecných spektier. Všetky algoritmy (okrem ICA) používajú uniformné spektrum.

Ako aj pri prvom teste, percentá ktoré dosahuje metóda LSE, zodpovedajú dobre separovaným komponentom. Prvý vstup je vytvorený pomocou neprekrývajúcich sa spektier. S ním nemá problémy žiaden použitý algoritmus. Problémy nie sú ani pri druhom vstupe, kde sa spektrá čiastočne prekrývajú. Tretie spektrá sú pre funkcie  $f_1$  a  $f_2$  príliš blízko. Zvláda ich funkcia  $f_4$ , dokonca s lepšími výsledkami ako v predchádzajúcich prípadoch. Pri poslednom vstupe, kde je jedno spektrum slabšie a celé sa nachádza pod iným spektrom, sa vo výsledkoch pre  $f_1$  a  $f_2$  prvé dva komponenty čiastočne prekrývajú. Pomocou jednej z funkcií  $f_1$  a  $f_4$  sme vždy dostali spektrá skoro identické s pôvodnými. Algoritmus FastICA vo všetkých prípadoch komponenty dobre separoval, vyššie hodnoty sú spôsobené tým, že nepočíta so šumom, a preto je šum vo výsledku silnejší.

## 4.5 Zložené spektrá

Spektrá sú niekedy aj súčtom funkcií hustôt viacerých rozdelení. Prispôbenie algoritmu je jednoduché. Zmeníme len kríženie, kde nebudeme krížiť celé spektrá, ale len jednotlivé

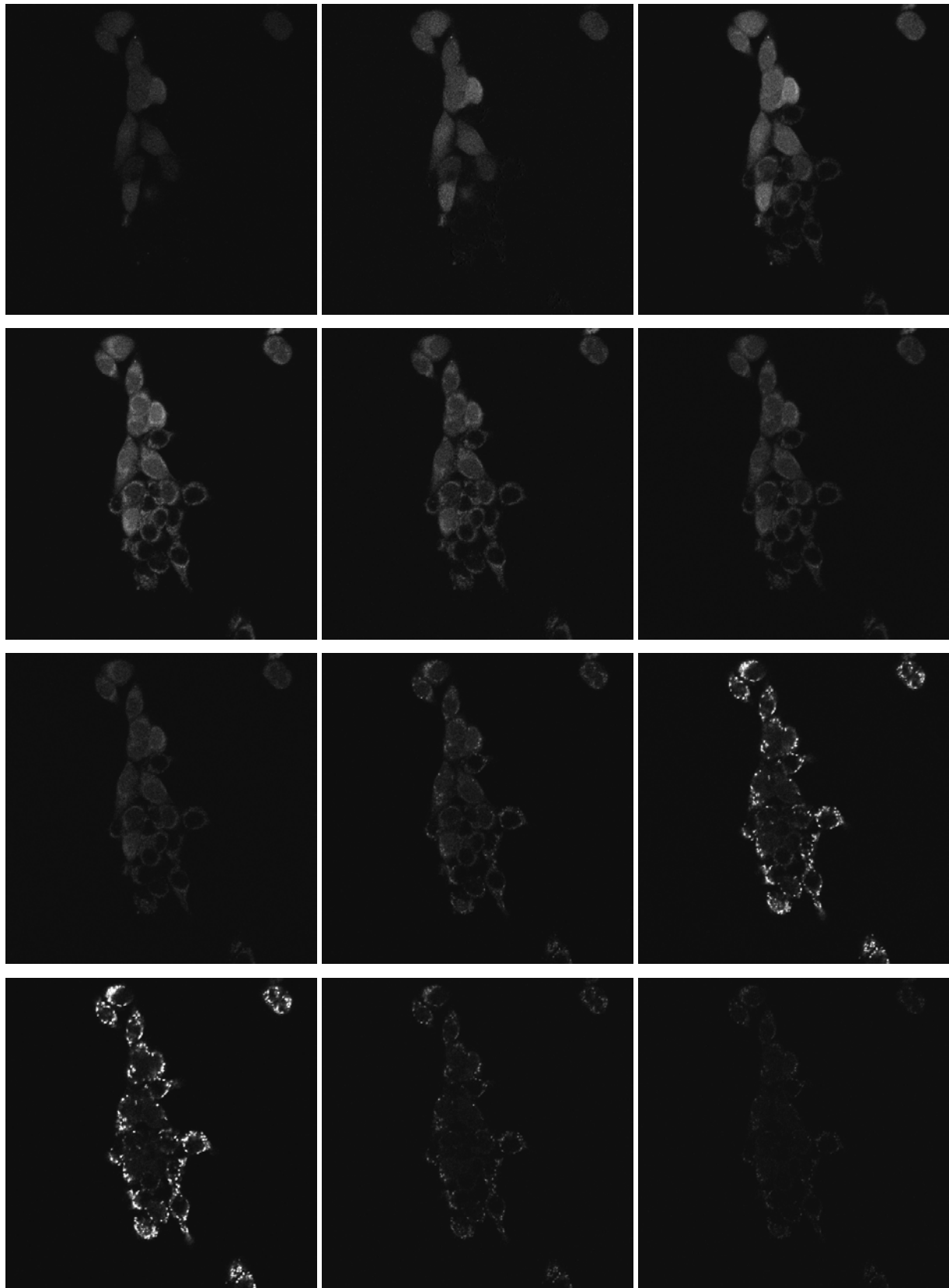
rozdelenia. Mutáciu meniť nemusíme. Ak sú spektrá sumou hustôt viac ako dvoch rozdelení, bude pravdepodobne lepšie využiť všeobecné spektrá. Pre tento prípad nebudeme vypracovávať ďalšie testy na umelých dátach. Budeme sa im venovať v časti 4.7 pri testovaní dát z praxe.

## 4.6 Paralelný výpočet

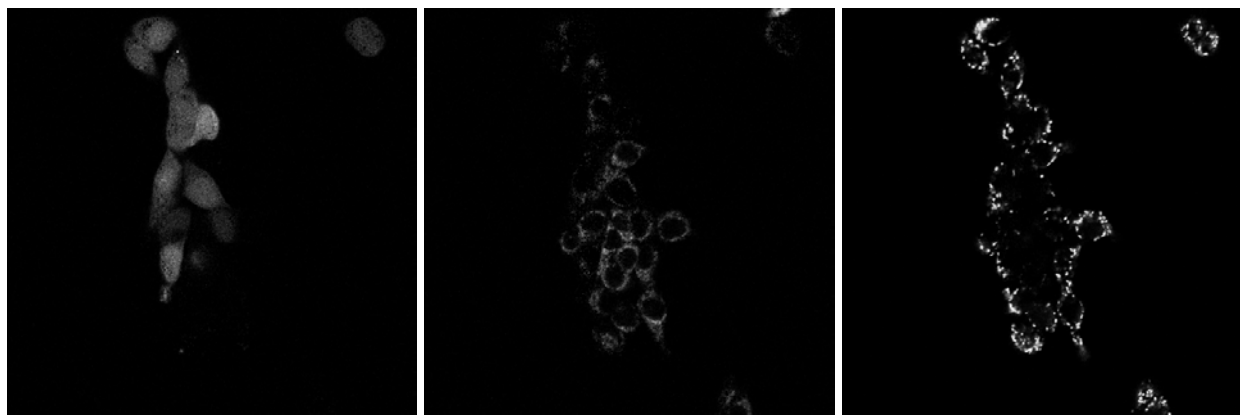
Cieľom tejto diplomovej práce nie je len nájsť vhodný algoritmus na dekompozíciu spektrálne rozlíšených dát fluorescenčnej mikroskopie, ale aj optimalizovať dobu behu tohto algoritmu. Jednou z možností ako algoritmus zrýchliť, je prispôbiť ho pre paralelný beh. Paralelné genetické algoritmy využívajú viacero populácií, ktoré si medzi sebou občas vymieňajú genetickú informáciu. My pôjdeme inou cestou. Najpomalšou časťou nášho algoritmu je výpočet funkcie  $f$ . V prípade, keď počet chromozómov neprevyšuje počet využívaných vstupov, je čas behu algoritmu závislý v podstate len od výpočtu  $f$ . Paralelnú verziu algoritmu vytvoríme tak, že pre každý chromozóm budeme počítat  $f$  v osobitnom procese.

## 4.7 Výsledky na dátach z praxe

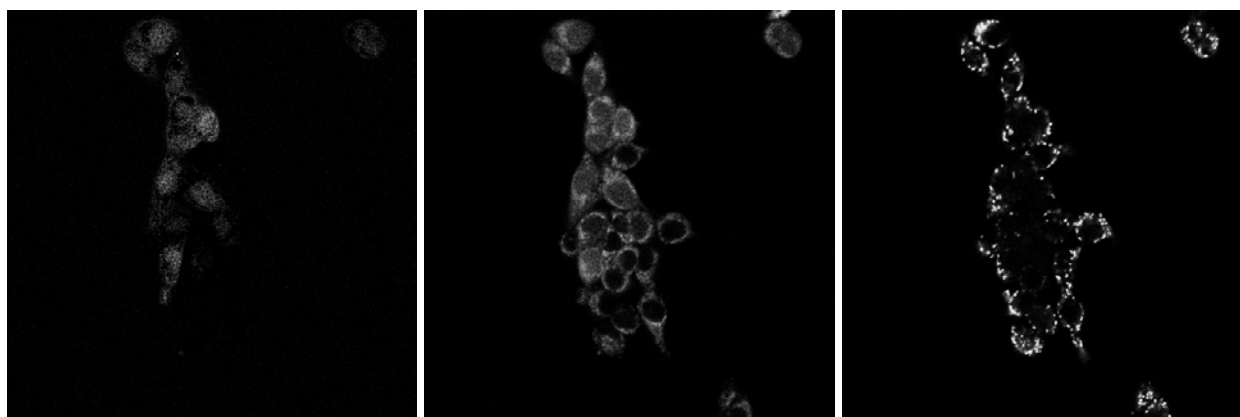
V predchádzajúcich častiach sme ukázali, že funkcie  $f_1$  a  $f_4$ , spolu s pridaním uniformného spektra do spektrálnej matice, by mohli byť dobré kritéria pre dekompozíciu spektrálne rozlíšených fluorescenčných dát. V tejto časti overíme túto hypotézu na dátach z praxe. Dáta boli získané pomocou mikroskopu LSM META 510, ZEISS v Laboratóriu laserovej mikroskopie v Medzinárodnom Laserovom Centre v Bratislave. Na testovanie použijeme 3 vstupy. Prvé dve sú zložené z troch komponent, tretí tvoria 2 komponenty so zloženými spektrami. Pri týchto vstupných dátach poznáme spektrá použitých farbív, takže výsledky môžeme porovnať s výsledkom metódy LSE. Porovnávať budeme výstup metódy LSE a výsledky minimalizovania funkcií  $f_1$  a  $f_4$ . Číselné charakteristiky výsledkov nám veľa nepovedia, preto zobrazíme výsledné obrázky a spektrá.



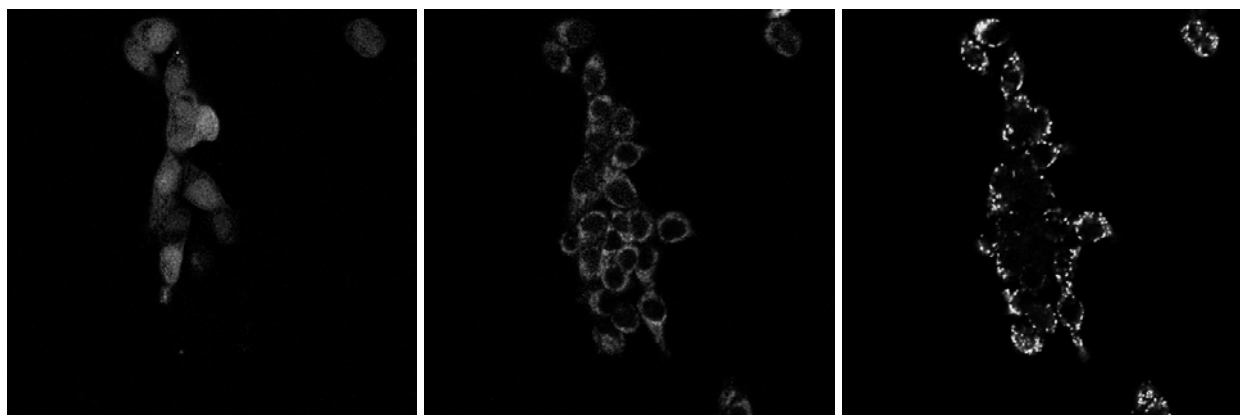
Obr. č. 6: Výstup mikroskopu. Chýbajú tu 3 posledné obrázky, na tých je ale len šum.



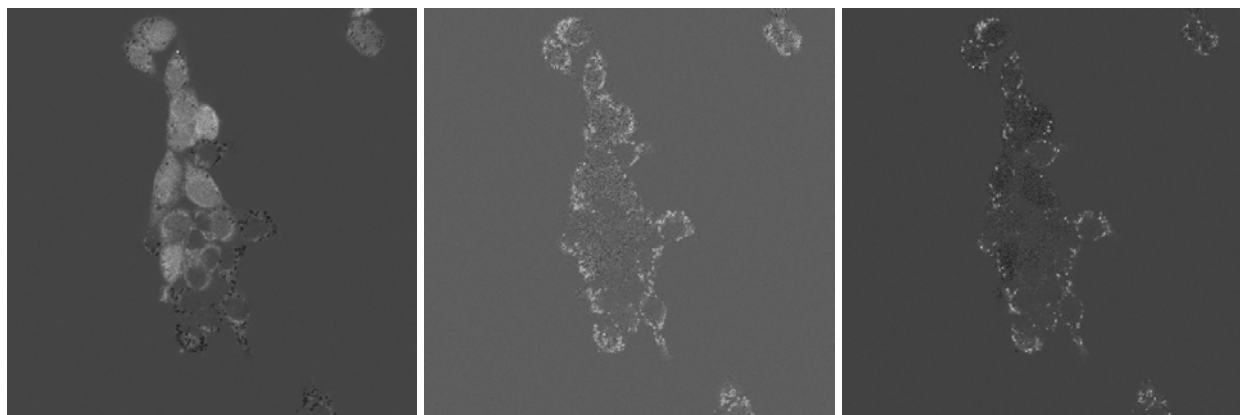
Obr. č. 7: Výsledky metódy LSE za použitia skutočných spektier.



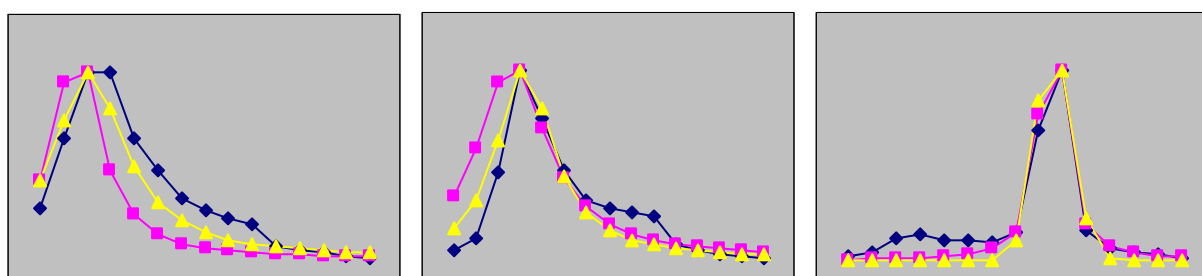
Obr. č. 8: Výsledky získane minimalizovaním funkcie  $f_1$ .



Obr. č. 9: Výsledky získane minimalizovaním funkcie  $f_4$ .



Obr. č. 10: Výsledky algoritmu FastICA.

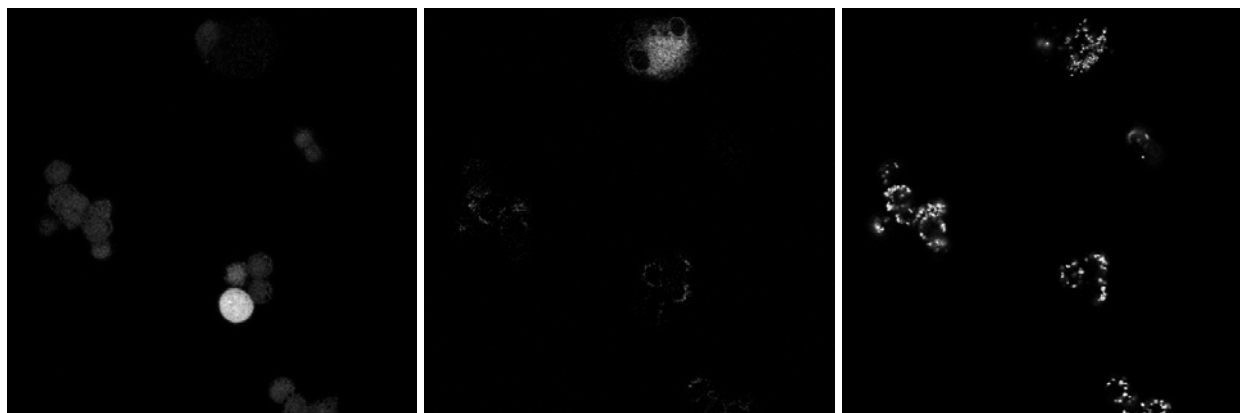


Obr. č. 11: Spektrá farbív. Modrá farba označuje skutočné spektrá. Ružová označuje výsledné spektrá pri minimalizovaní  $f_1$  a žltá označuje výsledné spektrá pri minimalizovaní  $f_4$ . Grafy prislúchajú k obrázkom v rovnakom stĺpci. Spektrá sú zobrazené na vlnových dĺžkach od 500 do 640  $nm$ .

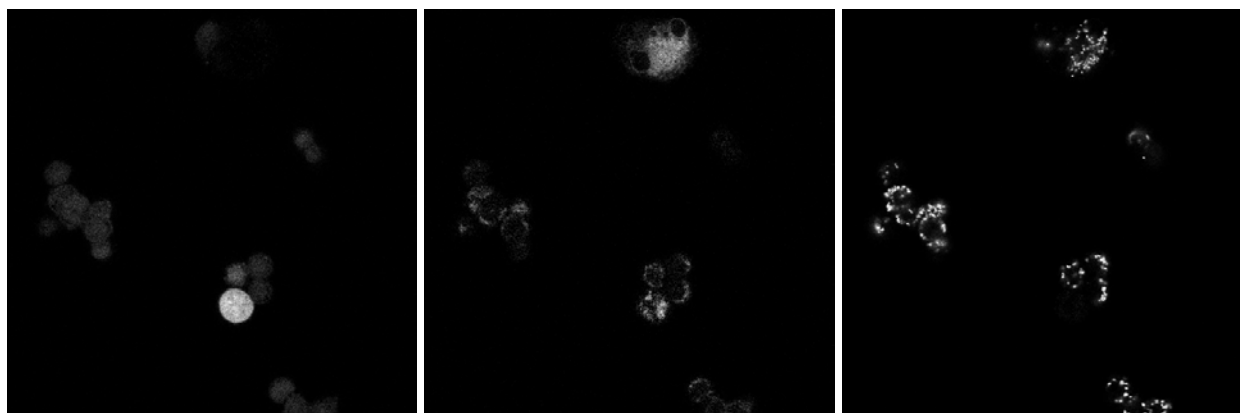
Minimalizovaním  $f_1$  sa nepodarilo úplne separovať prvé 2 komponenty. Je to spôsobené tým, že okolo  $A_5$  je veľká oblasť so skoro rovnakou hodnotou funkcie  $f_1$ . Na obrázku č. 8 je najčastejší výsledok. Pre takýto prípad, keď príliš veľa bodov dosahuje minimum, sme navrhli funkciu  $f_4$ . Aj keď nezávislosť nie je dobré kritérium (obr. č. 10), čo najväčšia nezávislosť v rámci tohto okolia pomáha. Hodnota  $1/kurt$  má podstatne menší pomer minima a maxima, ako má suma chýb. Preto sa pri minimalizovaní  $f_4$  uplatí nezávislosť, až keď sa spektrá dostanú do oblasti s malými rozdielmi v hodnote  $f_1$ . Výsledok (obr. č. 9) je ešte lepší ako pri lineárnom unmixingu pomocou skutočných spektier (obr. č. 7), kde sa časti komponentu strácajú.

Skutočné spektrá sú zložené z funkcií hustôt dvoch náhodných rozdelení. Príspevok druhého je malý, takže veľmi výsledok neovplyvnilo. Pri ďalšom vstupe, pri získavaní ktorého boli použité farbivá s rovnakými spektrami, budeme predpokladať, že spektrá sú zložené z troch

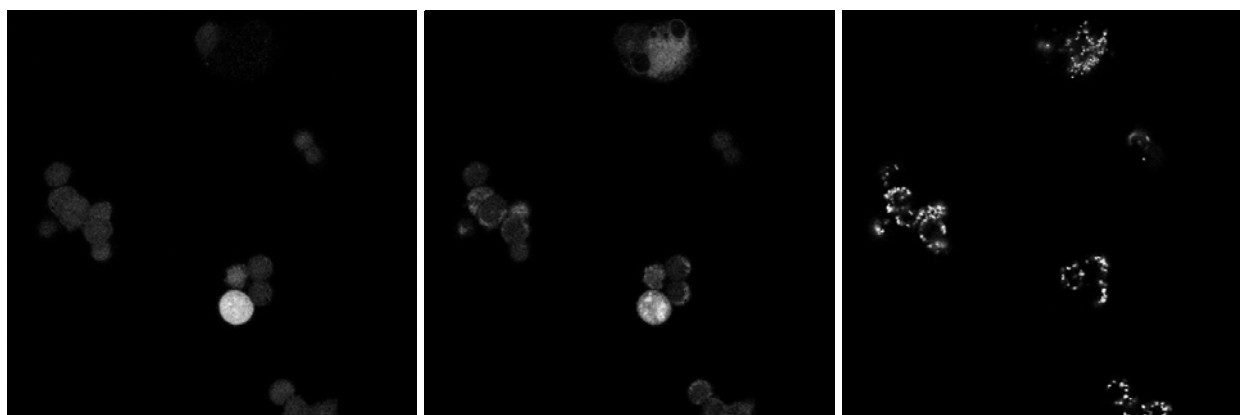
Gaussových alebo Lorentzových rozdelení. Dve by nestačili, rozdelenia by sa snažili len lepšie pokryť najvyššie body spektra.



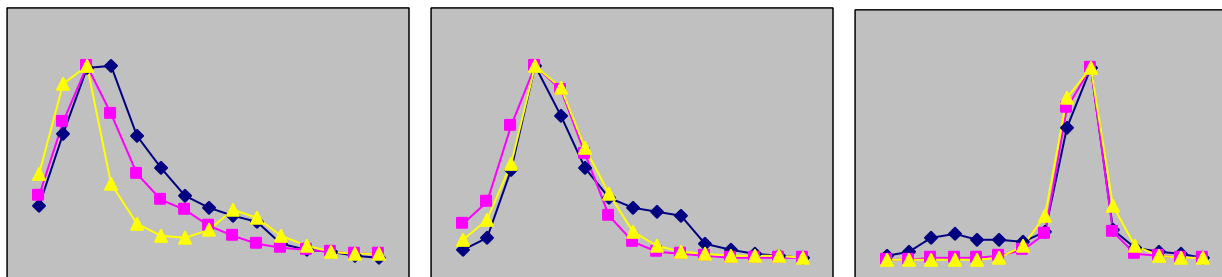
Obr. č. 12: Výsledky metódy LSE za použitia skutočných spektier.



Obr. č. 13: Najlepší výsledok.



Obr. č. 14: Najhorší výsledok.

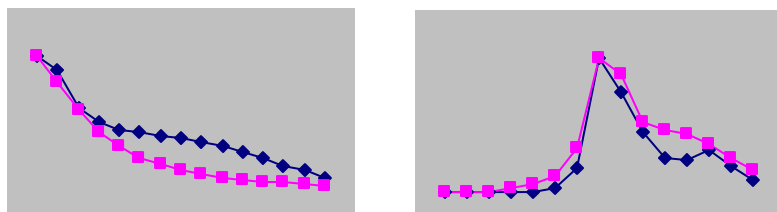


Obr. č. 15: Spektrá farbív. Modrá farba označuje skutočné spektrá. Ružová označuje spektrá výsledku z obr. č. 13 a žltá označuje spektrá výsledku z obr. č. 14. Grafy prislúchajú k obrázkom v rovnakom stĺpci.

Výsledky, pri ktorých sa hodnoty funkcií pohybujú okolo minima, sa pohybujú v rozmedzí výsledkov na obrázkoch číslo 13 a 14. V prípade  $f_4$  je to častejšie ten prvý, v prípade  $f_1$  druhý. Ako aj prvý krát, najlepší výsledok zachránil z druhého komponentu viac ako lineárny unmixing. Použitie zložených spektier výsledné spektrá veľmi nezmenilo.

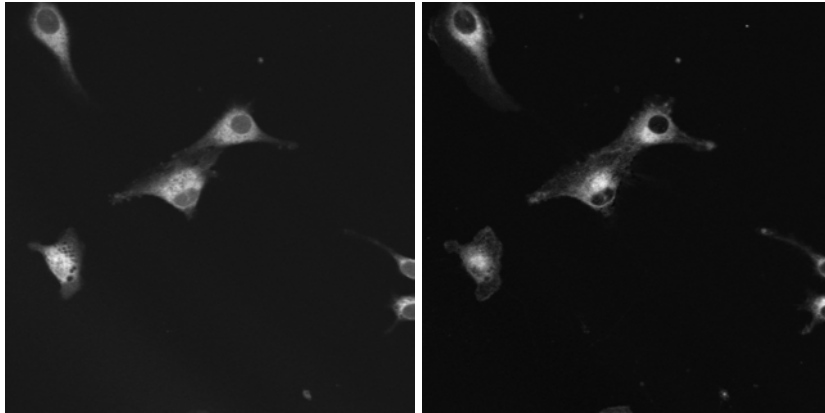
Na poslednom testovacom vstupe sa nachádzajú dva komponenty, ktoré majú zložené spektrá. Príspevok druhého rozdelenia je v tomto prípade silnejší, ako pri predchádzajúcich vstupoch.

Spektrá majú svoje vrcholy ďaleko od seba, takže v tomto prípade nevzniká viac riešení. Problém tohto vstupu je v tom, že spektrum prvého komponentu má vysoké hodnoty na všetkých vlnových dĺžkach. Do komponentu, ktorý prislúcha uniformnému spektru, sa preto dostala časť prvého komponentu. Tento problém môžeme vyriešiť tým, že k vstupu pridáme niekoľko obrázkov, na ktorých bude iba šum. Pri tomto vstupe som schválne nepoužil všetky obrázky, aby som poukázal na tento problém. Po pridaní zvyšných štyroch obrázkov, na ktorých je signál veľmi slabý, je výsledok skoro zhodný s výsledkami na obr. č. 17.

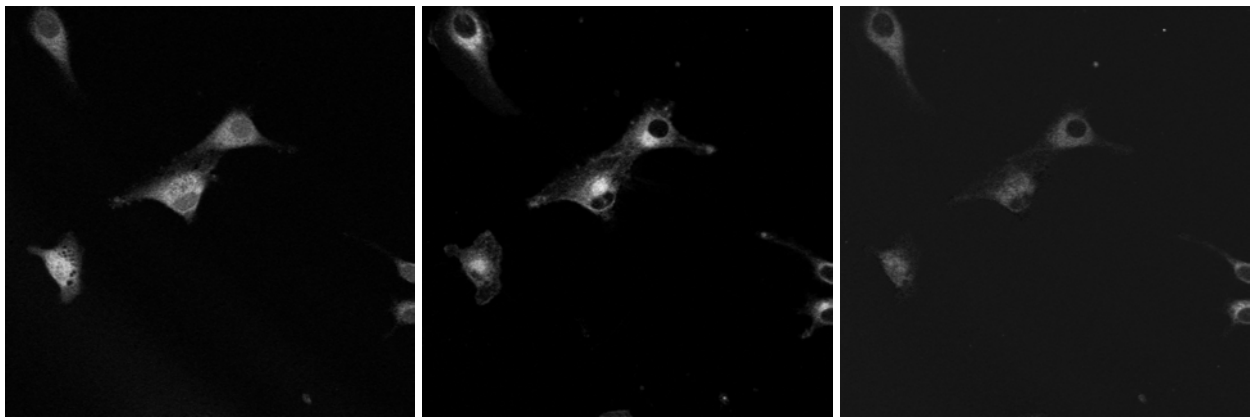


Obr. č. 16: Modrá farba označuje skutočné spektrá, ružová označuje spektra vypočítané.

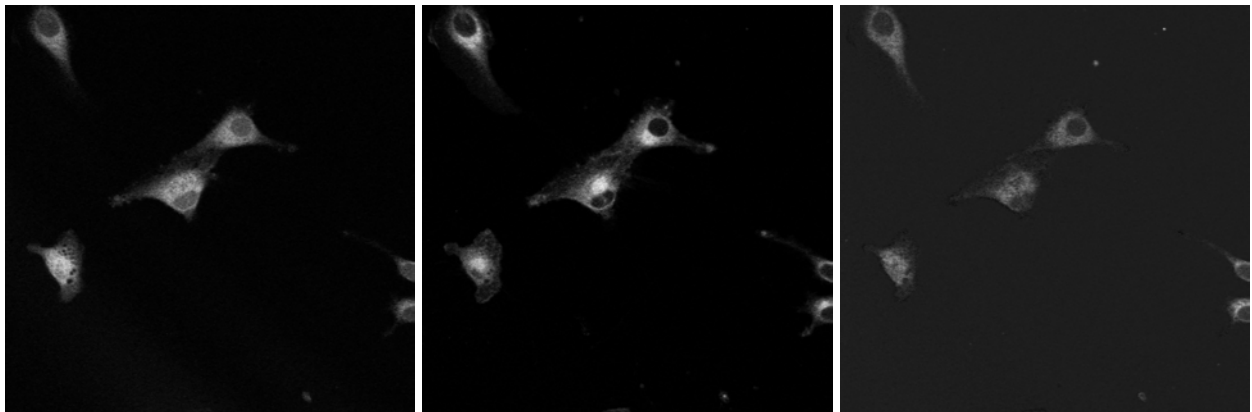




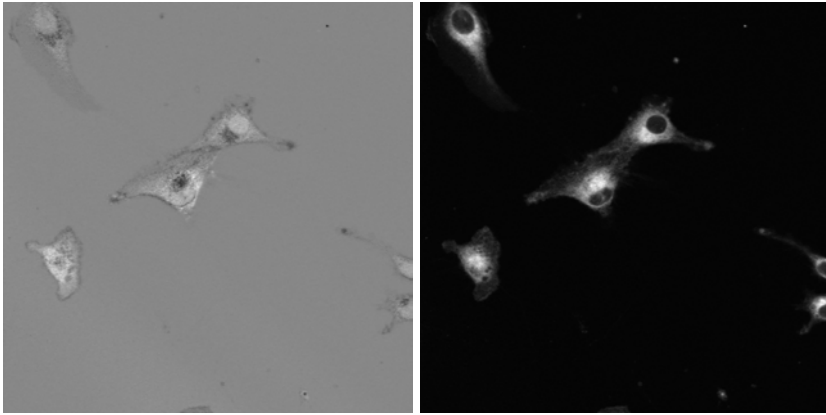
Obr. č. 17: Výsledky metódy LSE za použitia skutočných spektier.



Obr. č. 18: Výsledky metódy LSE pri pridaní uniformného spektra ku skutočným.



Obr. č. 19: Výsledok získaný minimalizovaním funkcie  $f_1$  alebo  $f_4$



Obr. č. 20: Výsledok algoritmu FastICA.

# Záver

V tejto práci sme sa venovali dekompozícii spektrálne odlišených fluorescenčných dát z konfokálnej mikroskopie. Na začiatku sme uviedli krátky prehľad o tom, ako funguje konfokálny mikroskop a aké sú jeho využitia. Potom sme sa venovali niektorým známym metódam na dekompozíciu spektrálne odlišených dát. Ako reprezentanta pre dekompozíciu so znalosťou spektrálnej matice sme vybrali metódu LSE. V prípade slepej dekompozície sme sa venovali Metóde nezávislých komponentov (ICA). Po zistení, že existujúce metódy nie sú dostačujúce, sme sa snažili navrhnúť nové kritérium, podľa ktorého by sme hľadali distribúciu jednotlivých komponentov vo vzorke. To nás priviedlo k funkcii (4.1.1). Po niekoľkých modifikáciách algoritmu, ktorý minimalizuje túto funkciu, sme dostali algoritmus na riešenie nášho problému. V prípade, keď sa spektrá silne prekrývajú, existuje okolo hľadaného výsledku pomerne veľké okolie, na ktorom sa hodnota funkcie (4.1.1) málo líši. Preto sme potrebovali dodatočné kritérium, ktoré by vybralo riešenie spomedzi týchto výsledkov. Hľadanie nezávislých komponentov nevedlo k dobrým výsledkom, kvôli vysokej štatistickej závislosti komponentov. Hľadanie čo najmenej závislých komponentov medzi výsledkami, ktoré získame minimalizovaním funkcie (4.1.1), sa ukázalo ako dobré riešenie.

Spomedzi možných kritérií nezávislosti sme si vybrali špicatosť, kvôli jej najnižším výpočtovým nárokom. Tento výber nemusí byť vhodný pre všetky dáta. Je potrebné otestovať ho na viacerých dátach z praxe a v prípade potreby nahradiť iným kritériom nezávislosti, prípadne použiť iné dodatočné kritérium. Implementované algoritmy sa budú používať na dekompozíciu spektrálnych dát z konfokálneho mikroskopu v Laboratóriu laserovej mikroskopie v Medzinárodnom Laserovom Centre v Bratislave.

## **Zoznam skratiek:**

CLSM – Confocal Laser Scanning Microscopy. Konfokálna laserová rastrovacia mikroskopia.

LSE – Least Square Estimation. Metóda najmenších štvorcov.

PCA – Principal Component Analysis. Metóda hlavných komponentov.

ICA – Independent Component Analysis. Metóda nezávislých komponentov.

FastICA – Konkrétny algoritmus na riešenie ICA.

LMM – Linear Mixture Model.

# Používateľská príručka:

Algoritmy popísané v práci som implementoval v jazyku Java, verzia 1.5. Nachádzajú sa v knižnici spectKonfo.

V triede GenAlg je implementovaný genetický algoritmus, ktorý minimalizuje funkcie (4.1.1) alebo (4.1.4). Na spustenie algoritmu je nutné najskôr zavolať jeden z troch možných konštruktorov:

GenAlg(int m, String[] s, int[] type)

- m udáva počet komponentov.
- s obsahuje cesty k jednotlivým obrázkom. Obrázky musia mať rovnaké rozmery. Podporované typy obrázkov sú .png a .bmp.
- type obsahuje typy spektier jednotlivých komponentov. Na označenie typov spektier sú použité konštanty z triedy Chromosome: UNIVERSAL je všeobecné spektrum. GL je tvorené jedným Gaussovým alebo Lorentzovým rozdelením. GLM2 (resp. GLM3) sa skladá z dvoch (resp. troch) rozdelení. UNIFORM označuje uniformné rozdelenie. Rozdelenie typu UNIFORM môže byť použité maximálne jedno. Ak je type rovné null alebo sa jeho dĺžka nezhoduje s m, tak je použité predvolené nastavenie: m krát GL spektrum a jedno uniformné.

GenAlg(int m, String[] s, int[] type, double minWaveLength, double step, int functionType)

- minWaveLength udáva minimálnu vlnovú dĺžku použitú pri snímaní obrázkov a step udáva rozdiel medzi jednotlivými vlnovými dĺžkami. Hodnoty sa udávajú v nanometroch. Predvolené hodnoty sú 500 a 10.
- functionType udáva funkciu, ktorú má algoritmus minimalizovať. Možné hodnoty sú 1 a 4 podľa funkcií definovaných v tejto práci. Pri použití prvého konštruktoru je použitá  $f_1$ .

GenAlg(int m, String[] s, int[] type, double minWaveLength, double step, int functionType, int chromCount, int inputCount, int epochCount)

-posledný konštruktor pridáva možnosť nastaviť počet chromozómov, počet vstupov a počet epoch. Predvolené hodnoty sú 100, 100, 50.

Po zavolaní konštruktoru je možné algoritmus spustiť dvoma funkciami:

- double[][] run() – spustí algoritmus raz a vráti spektrálnu maticu víťaza.

- `double[][] bestOf(int k)` – spustí algoritmus `k` krát. Pri každom víťazovi spočíta hodnotu minimalizovanej funkcie na celom vstupe a vráti spektrálnu maticu víťaza.

Po skončení je ešte možné zavolať funkciu `saveOutput(double[][] winner)`, ktorá uloží na disk výsledky metódy LSE.

V triedach `FastICA` a `PCA` sú implementované tieto algoritmy. Obidve používajú rovnaké funkcie, preto ich popíšeme spoločne:

`FastICA(int outputDim, String[] fileList)` , `PCA(int outputDim, String[] fileList)`

- konštruktor. `outputDim` označuje dimenziu výstupu a vo `fileList` sa nachádzajú cesty k vstupným obrázkom. Všetky obrázky musia mať rovnaké rozmery.

`FastICA(int outputDim, String[] fileList, double accuracy)`

- druhý konštruktor pre `FastICA` pridáva možnosť hodnotu, podľa ktorej sa zisťuje, či už algoritmus skonvergoval. Predvolená hodnota je 0,001.

`run()` – spustí daný algoritmus a zapíše výsledky do pracovného adresára.

# Použitá literatúra

1. Cover T. M., Thomas J. A., Elements of Information Theory. Wiley. 1991
2. Darwin Ch., Darwinův původ druhů v ilustracích, Panorama, Praha 1989
3. Dickinson M., Bearman G., Tillel S., Lansford R., Fraser S., Multi-Spectral Imaging and Linear Unmixing Add a Whole New Dimension to Laser Scanning Fluorescence Microscopy, BioImaging, Vol. 31, No.6 (2001).
4. Haykin S., Neural Networks A Comprehensive Foundation, Second Edition.
5. Huber P., Projection pursuit, The Annals of Statistics, Vol.13, p. 435-475.
6. Hyvärinen A., New approximations of differential entropy for independent component analysis and projection pursuit, Advances in Neural Information Processing Systems, Vol. 10, p. 273–279, 1998
7. Hyvärinen A., Oja E., Independent Component Analysis: Algorithms and Applications, Neural Networks, Vol. 13, p. 411-430, 2000
8. Hyvärinen A., Survey on Independent Component Analysis, Neural Computing Surveys 2: p. 94-128, 1999.
9. Jones M., Sibson R., What is projection pursuit ? J. of the Royal Statistical Society, Ser. A, Vol. 150, p. 1–36. 1987
10. Kvasnička V., Pospíchal J., Tiňo P., Evolučné algoritmy, STU Bratislava 2000.
11. Levenson R., Hoyt C., Spectral imaging and microscopy, American Laboratory, 2000.
12. Li J., Linear unmixing of hyperspectral signals via wavelet feature extraction, Dissertation, Department of Electrical and Computer Engineering Mississippi State, December 2002

13. Nascimento J.M.P., Dias J.M.B., Does Independent Component Analysis Play a Role in Unmixing Hyperspectral Data?, *Geoscience and Remote Sensing*, Vol. 43, p.175-187, 2005
14. Pawley J. B., *Handbook of Biological Confocal Microscopy*, Third edition, 2006
15. [http://en.wikipedia.org/wiki/Confocal\\_laser\\_scanning\\_microscope](http://en.wikipedia.org/wiki/Confocal_laser_scanning_microscope)