



VILNIAUS GEDIMINO TECHNINIS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINIO MODELIAVIMO KATEDRA

Ignas Dapšys

**ORGANINIŲ MOLEKULIŲ FURJĖ TRANSFORMACIJOS
SPEKTROSKOPIJOS SPEKTRINIŲ POŽYMIŲ NUSTATYMO
METODAI**

**METHODS FOR SPECTRAL FEATURE DETECTION IN
FOURIER TRANSFORM SPECTROSCOPY OF ORGANIC
MOLECULES**

Baigiamasis magistro darbas

Technomatematikos studijų programa, valstybinis kodas 632G16001
Matematikos studijų kryptis

Vilnius, 2018

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINIO MODELIAVIMO KATEDRA

TVIRTINU
Katedros vedėjas

(Parašas)

(Vardas, pavardė)

(Data)

Ignas Dapšys

ORGANINIŲ MOLEKULIŲ FURJĖ TRANSFORMACIJOS
SPEKTROSKOPIJOS SPEKTRINIŲ POŽYMIŲ NUSTATYMO
METODAI

METHODS FOR SPECTRAL FEATURE DETECTION IN
FOURIER TRANSFORM SPECTROSCOPY OF ORGANIC
MOLECULES

Baigiamasis magistro darbas

Technomatematikos studijų programa, valstybinis kodas 632G16001

Matematikos studijų kryptis

Darbo vadovas prof. habil. dr. Raimondas Čiegis

(parašas)

(data)

Konsultantas dr. Mindaugas Karaliūnas

(parašas)

(data)

Lietuvių kalbos konsultantė lekt. Regina Žukienė

(parašas)

(data)

Vilnius, 2018

Vilniaus Gedimino technikos universitetas

Fundamentinių mokslų fakultetas

Matematinio modeliavimo katedra

ISBN ISSN

Egz. sk.

Data-....-....

Technomatikos studijų programos baigiamasis magistro darbas

Pavadinimas **Organinių molekulių Furjė transformacijos spektroskopijos spektrinių požymių nustatymo metodai**

Autorius **Ignas Dapšys**

Vadovas prof. habil. dr. **Raimondas Čiegis**

Kalba

☒ x

lietuvių

☐

užsienio

Anotacija

Baigiamajame magistro darbe nagrinėjama spektrinių komponentų išskyrimo iš mišinių spektrų problema, jos sprendimo būdai ir jų taikymai terahercinių (THz) bangų Furjė transformacijos spektroskopijai. Aprašytas matematinis spektrinių komponentų modelis, pasiūlytas mažiausių kvadratų aproksimacijos principu pagrįstas komponentų išskyrimo metodas. Aptarti kitų mokslininkų išbandyti komponentų išskyrimo metodai: pagrindinių komponentų analizė (PCA), atraminių vektorių mašinos (SVM), mažiausių kvadratų metodas (LS); genetinis algoritmas (GA) ir kiti optimizavimo metodai. Mūsų siūlomas metodas išbandytas su sintetiniais ir realiais duomenimis, pasiūlyta metodo modifikacija realių spektrų analizei, pateikiamos baigiamojo darbo išvados.

Darbą sudaro 7 dalys: įvadas, uždavinio aprašymas, spektroskopijos eksperimento aprašymas, sprendimo metodai, metodų bandymai ir jų rezultatai, išvados ir siūlymai, literatūros sąrašas.

Darbo apimtis – 38 p. teksto be priedų, 17 iliustr., 1 lent., 14 bibliografinių šaltinių.

Prasminiai žodžiai: terahercinės bangos, spektroskopija, matematinis modeliavimas, Furjė transformacija, mažiausių kvadratų metodas, pagrindinių komponentų analizė, komponentų išskyrimo problema, aproksimavimas Gauso funkcijomis.

Vilnius Gediminas Technical University
Fundamental Sciences faculty
Mathematical Modelling department

ISBN ISSN
Copies No.
Date-....-....

Technomathematics study programme master thesis

Title **Methods for spectral feature detection in Fourier transform spectroscopy of organic molecules**

Author **Ignas Dapšys**

Academic supervisor prof. habil. dr. **Raimondas Čiegis**

Thesis language

☒

x

Lithuanian

☐

Foreign (English)

Annotation

In this master thesis we examine the spectral feature separation problem, its solutions, and their application to terahertz (THz) Fourier transform spectroscopy. We describe a mathematical model of spectral components and propose a solution based on least squares Gaussian approximation. A few component separation methods described in scientific literature are examined, namely: principal component analysis (PCA), support vector machines (SVM), method of least squares (LS), as well as genetic algorithm (GA) and other optimization methods. Our proposed method is applied to synthetic and real data; we also propose an alteration to this method to make it suitable for analysis of real spectra. Finally, the main conclusions of this thesis are stated.

Structure: introduction, problem description, spectroscopy experiment setup, methods of solution, method trials and their results, conclusions and suggestions, bibliography.

Thesis consists of: 38 p. of text without appendices, 17 figures, 1 table, 14 bibliographical entries.

Keywords: terahertz radiation, spectroscopy, mathematical modelling, Fourier transform, method of least squares, principal component analysis, component separation problem, Gaussian approximation.

Turinys

Įvadas	5
1 Uždavinio aprašymas	8
1.1 Uždavinio formuluotė	8
1.2 Literatūros apžvalga	10
2 Eksperimento aprašymas ir gaunami duomenys	12
2.1 Eksperimento aprašymas	12
2.2 Pradinis duomenų apdorojimas Savitzky–Golay filtru	16
3 Sprendimo metodai	18
3.1 Optimizavimo metodai	18
3.2 Atraminių vektorių mašinos	19
3.3 Pagrindinių komponentų analizė	22
3.4 Diskriminantinė analizė	23
3.5 Mažiausių kvadratų metodas	23
4 Metodų taikymas duomenims	25
4.1 Pagrindinių komponentų analizė	25
4.2 Atraminių vektorių mašina	27
4.3 Optimizavimo uždavinio sprendimas	29
Išvados	38
Literatūra	39

Iliustracijų sąrašas

2.1	Furjė transformacijos spektroskopijos eksperimento sandara	13
2.2	Eksperimento metu gaunamų spektrų pavyzdys. Pavaizduoti kofeino (raudonas) ir azoto dujų aplinkos (juodas) FFT spektrai	13
2.3	Eksperimento metu gaunamų spektrų pavyzdys. Pavaizduotas kofeino pralaidumo spektras $T(\omega)$, gautas padalijus kofeino FFT spektrą iš aplinkos spektro	14
2.4	Transformuotas kofeino pralaidumo spektras $\varphi(\omega)$	15
4.1	Spektro #3 grafikas po normavimo	26
4.2	Spektro #3 grafikas po singuliariojo išskaidymo ir komponentių išskyrimo	26
4.3	Spektro #3 grafikas (mėlyna) ir jo pagrindinės komponentės (žalia) kartu	27
4.4	Tiesiškai atskiriamų klasių pavyzdys, čia žalia spalva reiškia klasę 0; raudona – klasę 1	28
4.5	Tiesiškai neatskiriamų klasių pavyzdys, čia raudona spalva reiškia klasę 0; mėlyna – klasę 1	29
4.6	Sintetinis spektras, naudotas skaitiniam tyrimui. Jo komponentių parametrai yra, iš kairės į dešinę atitinkamai: viršūnių aukščiai yra (2, 0, 5, 1, 5); spektriniai plotčiai – (1, 10, 1); viršūnių padėtys – (10, 25, 40)	30
4.7	Deformuoti spektrai. Čia mėlynas spektras yra originalus, raudonas – deformuotas sinusu, geltonas – deformuotas Gauso triukšmu	31
4.8	Spektrai, gaunami tam tikrose BFGS optimizavimo metodo iteracijose, kai aproksimuojamas spektras yra: a) pradinis, b) deformuotas 1, c) deformuotas 2, iš 4.1 lentelės. Legendose skaičiai reiškia iteracijos numerį; O – tikrasis sprendinys	32
4.9	Spektrai, gaunami tam tikrose SA optimizavimo metodo iteracijose. Žymėjimai tokie patys, kaip ir 4.8 paveiksle	33
4.10	Realaus spektro pavyzdys. Mėlyna linija yra originalus spektras, žalia – sugludintas spektras	34

4.11 Regresinis spektras (raudona) ir jo tendencijos linija (mėlyna)	35
4.12 Diferencinis spektras	36
4.13 Filtruotas diferencinis spektras (mėlyna) ir jo Gauso aproksimacija (violetinė)	36
4.14 LDA duomenų projekcija. Juodos spalvos apskritimai reprezentuoja spekt- rus be kofeino, raudonos – spektrus, turinčius kofeino	37

Lentelių sąrašas

4.1	Optimizavimo pradinių artinių lentelė. Čia deformuotas 1 yra pradinis spektras su sinusiniu iškraipymu, o deformuotas 2 – su Gauso triukšmu. Pirmojoje eilutėje pateikti originalaus spektro parametrai	31
-----	---	----

Ivadas

Spektroskopija – tai medžiagos savybių tyrimas naudojant elektromagnetinę spinduliuotę. Tyrimo metu gaunamas spektras – fizikinio dydžio priklausomybė nuo dažnio, iš kurio galima spręsti apie tiriamojo bandinio sandarą ir savybes. Spektroskopija yra plati fizikos sritis, nes naudojama plati elektromagnetinė spinduliotė, turtinga sąveikos su medžiaga savybėmis. Kelios plačiausiai naudojamų spektroskopijos rūšių yra Ramano (regimosios šviesos), infraraudonųjų spindulių, Rentgeno spindulių difrakcijos ir terahercinių bangų.

Ramano spektroskopijoje (angl. *Raman spectroscopy*) naudojama regimoji šviesa, kuri, sąveikaudama su mėginiu, sukelia neelastinį išsklaidymą (angl. *inelastic scattering*) – į mėginį krentantys fotonai atsispindi turėdami arba daugiau, arba mažiau energijos. Jos skirtumas tarp krentančių ir atsispindėjusių fotonų turi informacijos apie mėginio dažnius – iš jos gaunamas spektras. Ramano metodas naudingas tuo, kad naudoja nejonizuojančią spinduliuotę, ji nekenkia gyviems organizmams ir jautriems objektams. Norint ištirti objektą, nebūtina jo sunaikinti, o gaunami spektrai atitinka tiriamas medžiagas. Artimųjų infraraudonųjų spindulių spektroskopija (angl. *near-infrared spectroscopy*) yra panaši į Ramano, tik vietoje regimosios šviesos naudojamos infraraudonosios bangos – šimto milijonų hercų (Hz) eilės; elektromagnetinių bangų skalėje priešais regimąją šviesą.

Rentgeno spindulių difrakcija (angl. *X-ray diffraction*) naudojama kristalų cheminei struktūrai nustatyti. Šiam metodui naudojamos bangos yra 10^{18} Hz eilės, o jų ilgis yra angstromų eilės, todėl jos gali atskirti pavienius atomus; elektromagnetinių bangų skalėje jos yra už regimosios šviesos ir ultravioletinių spindulių. Tai yra jonizuojanti spinduliuotė, kuri gali paveikti ryšius tarp bandinio atomų. Metodas paremtas difrakcijos reiškiniu: apšvitinus mėginį tiesiu Rentgeno spindulių pluoštu, mėginio atomų elektronai šį pluoštą iškraipo. Iškraipymai sudaro dvimatį difrakcinį vaizdą. Keičiant stebimo kristalo padėtį pluošto atžvilgiu, gaunama difrakcinių vaizdų serija, kuriai pritaikius Furjė transformacijos metodą, galime gauti trimatį kristalo vaizdą. Šiam metodui reikalingas specialus mėginio paruošimas – turime kristalizuoti reikiamą medžiagą.

Mus labiausiai domina terahercinių bangų spektroskopija [1]. Tai metodas, kuriame mėginiai apšviečiami elektromagnetinėmis bangomis, trilijono hercų eilės dažnio. Šių bangų ilgis yra $30\ \mu\text{m}$ - $3\ \text{mm}$ eilės; EM skalėje jos yra tarp infraraudonųjų spindulių ir mikrobangų. THz spektroskopija, kaip ir Ramano, naudoja nejonizuojančią spinduliotę, o mėginio nebūtina sunaikinti. Tačiau THz spinduliai prasiskverbia pro daugelį medžiagų, tokių, kaip tekstilės pluoštai, plastikas, popierius, ir t.t., ko negali Ramano ir infraraudonųjų spindulių spektroskopijoje naudojamos bangos.

Terahercinės (ar kokios nors kitos) spektroskopijos eksperimentui atlikti reikia keturių dalykų: šviesos šaltinio, bandinio, dispersinio elemento (interferometro) ir šviesos detektoriaus. THz bangas galima generuoti puslaidininkinį kristalą sužadinant femtosekundiniu (fs) lazeriu. Generuota šviesa apšviečiamas bandinys, kuris ją veikia tokiais būdais: (1) praleidžia; (2) atspindi; (3) sugeria; (4) išsklaido.

Mes galime išmatuoti tik pirmuoju ir antruoju būdu paveiktą šviesą. Tai padarome interferometru – interferenciją matuojančiu prietaisu. Šviesos impulsą užrašome kaip interferencijos mato reikšmę kiekvienam laiko momentui. Atlikus išmatuotų impulsų Furjė transformaciją, gauname spektrą – šviesos energijos pasiskirstymą bangos dažnio atžvilgiu. Turėdami praleidimo ir atspindžio spektrus, galime apskaičiuoti pralaidumo ir sugerties spektrus. Jie atitinkamai vaizduoja lūžio rodiklio n ir sugerties koeficiento α pasiskirstymus pagal dažnį. Jei detektorius pakeisime vaizdo kamera, gausime nebe spektrą, o mėginio vaizdą. Šis metodas vadinamas terahercinių bangų vaizdinimu.

THz spektroskopijai ir vaizdinimui numatomi taikymai daugelyje sričių – nuo saugumo iki meno išsaugojimo [2]. Saugumo srityje THz spektroskopija ir vaizdinimas gali būti naudojami pavojingų daiktų [3] ir medžiagų [4] aptikimui. Maisto pramonėje šie spinduliai gali padėti surasti svetimkūnius maisto produktuose [5], o taip pat nustatyti suklastotus maisto produktus [6].

Taikant spektroskopiją cheminių medžiagų aptikimui, tenka atsakyti į klausimą, ar sudėtinėje įvairių komponentų medžiagoje yra ieškomų cheminių junginių. Tai daroma lyginant sudėtinės medžiagos spektrą su žinomu chemiškai grynos ieškomos organinės medžiagos spektru. Jei ieškomos medžiagos mišinyje yra, tada pagal superpozicijos principą, jos spektriniai požymiai turėtų matytis mišinio spektre. Tačiau organinių molekulių aptikimui natūraliose aplinkose, tokiose kaip augalų pluoštai arba ekstraktai, reikia atsižvelgti į spektrus iškraipiančius faktorius:

1. Natūraliose aplinkose esančių organinių molekulių sugerties linijos yra išplitusios ir persikloja su kitų aplinkos sudedamųjų dalių spektriniais požymiais;
2. Dėl mišinio komponentų sąveikos atsiranda papildomi spektriniai požymiai, nebūdingi pavieniams komponentams. Be to, ieškomos medžiagos sugerties linijos gali išplisti arba pasislinkti dažnių skalėje.

Del šių priežasčių sunku vienareikšmiškai atpažinti ieškomų cheminių medžiagų spektrinius požymius.

Minėtas problemas galima spręsti dviem būdais: pirmasis — teoriškai apskaičiuoti spektrų iškraipymus naudojantis kvantinės chemijos metodais, tačiau tam reikia tiksliai aprašyti nežinomą sudėtingą sistemą, t. y. ieškomas molekules ir jų aplinką. Antrasis būdas – naudotis duomenų analizės metodais (pavyzdžiui, pagrindine komponentų analize arba atraminių vektorių mašinomis). Abiem atvejais tai yra sudėtingas uždavinys, reikalaujantis didelių skaičiavimo resursų dėl didelio kiekio duomenų ir laiko patikrinti rezultatų patikimumą.

Šio magistro darbo tikslas – išspręsti komponentų išskyrimo iš mišinio problemą, o skaičiavimus pritaikyti kofeino turinčių bandinių THz spektrų analizei siekiant nustatyti kofeino pėdsakus spektre.

1 skyrius

Uždavinio aprašymas

1.1. Uždavinio formuluotė

Šiame magistro darbe siūlome tokį problemos sprendimą: turėdami išmatuotą spektrą $\varphi(\omega)$, jį aproksimuojame tokiu pavidalu:

$$\varphi(\omega) \approx \sum_{j=1}^n \alpha_j f_j(\omega) + A(\omega), \quad (1.1)$$

čia $f_j(\omega)$, $1 \leq j \leq n$ – spektrinės komponentės, α_j – komponentių amplitudės, $A(\omega)$ – triukšmas. Jei jis yra mažas, galime jį atmesti.

Aproksimavimo esmė – minimizuoti turimos funkcijos ir ją aproksimuojančios funkcijos skirtumo normos kvadrata. Kitaip tariant, turime tokį optimizavimo uždavinį: rasti

$$\arg \min_{\alpha_j} \left\| P\varphi(\omega) - \sum_{j=1}^n \alpha_j f_j(\omega) \right\|^2, \quad (1.2)$$

kad būtų tenkinami apribojimai

$$\sum_{j=1}^n \alpha_j = 1 \quad (1.3)$$

$$\alpha_j \geq 0, \quad 1 \leq j \leq n. \quad (1.4)$$

P – operatorius, reprezentuojantis papildomą spektro apdorojimą. Šio uždavinio dimensija yra $3n$.

Kaip minėjome anksčiau, spektrinės komponentės gali plėstis ir slankioti. Todėl šiuo atveju reikės spręsti tokį uždavinį: rasti

$$\arg \min_{\alpha_j, \beta_j, \gamma_j} \left\| P\varphi(\omega) - \sum_{j=1}^n \alpha_j f_j \left(\frac{\omega - \gamma_j}{\beta_j} \right) \right\|^2, \quad (1.5)$$

čia γ_j yra komponentių poslinkiai (arba *centrai*), o β_j – jų pločiai. Apribojimai tie patys (1.3 – 1.4), tik dar bus vienas papildomas:

$$\beta_j > 0, \quad 1 \leq j \leq n. \quad (1.6)$$

Darome prielaidą, kad spektrinės komponentės galima aprašyti Gauso funkcija, t.y.:

$$f_j \left(\frac{\omega - \gamma_j}{\beta_j} \right) = e^{-\left(\frac{\omega - \gamma_j}{\beta_j} \right)^2}. \quad (1.7)$$

(1.5)-(1.6) optimizavimo uždavinys yra gana sudėtingas. Viena iš priežasčių yra ta, kad komponentių centrai optimizavimo metu slankioja, todėl algoritmų vykdymo laikas ilgas ir didesnė tikimybė jam „užstrigti“ – nutraukti vykdymą anksčiau laiko, neradus minimumo. Norint išvengti šių problemų, reikia rasti komponentių centrus prieš sprendžiant (1.5)-(1.6) uždavinį.

Tam mes atliekame papildomą optimizavimą – imame spektrą, ir ieškome jo maksimumo taškų. Mums reikia rasti *visus* šiuos taškus, todėl darome taip:

1. Suskaidome spektro apibrėžimo sritį $[0, \Omega]$ taškais:

$$P_1 = \{\omega_i : 0 \leq \omega_i \leq \omega_j \leq \Omega, i \leq j\}; \quad (1.8)$$

2. Kiekvieną $\omega_i \in P_1$ naudojame kaip pradinį tašką lokalaus optimizavimo metodui. Tada kiekvienam skaidinio taškui turėsime vieną maksimumo tašką. Jie gali kartotis, todėl išrenkame tik unikalios jų reikšmes.

Taip darydami turime įsitikinti, kad spektrai yra *suglodinti* (apie duomenų glodinimą rašoma 2.2 skyrelyje). Atlikus centrų radimo procedūrą, (1.5)-(1.6) uždavinio dimensija sumažėja iki $2n$

1.2. Literatūros apžvalga

Publikuotų straipsnių iš terahercų spektroskopijos srities yra nemažai. Šiame skyrelyje paminėsime kelis straipsnius, kurie nagrinėti rengiant šį magistro darbą.

M.Yin ir kitų straipsnyje [6], lyginami keli skirtingi metodai – genetinis algoritmas, naudojamas kartu su mažiausių kvadratų metodu, ir du skirtingi intervalų analizės metodai. Jų visų tikslas yra parinkti bangos ilgį, pagal kuriuos galima identifikuoti kitus mėginius. Pirmiausiai autoriai aprašo genetiniu algoritmu paremtą būdą, kuris vykdomas pagal tokią žingsnių seką:

1. Pradinė populiacija; atsitiktinai sugeneruojama keletas dvejetainių sekų, kuriose 1 reiškia, kad kintamasis (bangos ilgis) parinktas, 0 – neparinktas;
2. Individo tinkamumo tikrinimas; pagal kiekvieną populiacijos individą parenkame kintamuosius, ir jiems pritaikome dalinių mažiausių kvadratų regresiją. Tinkamumo matas – kvadratų vidurkio šaknies paklaida (angl. *root-mean-square error*, RMSE), kurią reikia minimizuoti;
3. Parenkame tinkamiausius individus kitai kartai; atliekame rekombinaciją ir mutaciją;
4. Likusią populiacijos dalį naudojame 2 žingsnyje. Procesą kartojame N kartų – pabaigoje gauname identifikacinius bangos ilgį.

Šis metodas lyginamas su dar dviem: intervaliniu PLS ir atgaliniu intervaliniu PLS. Abiem atvejais renkame ne atskirus bangos ilgį, bet jų intervalus. Skirtumas tarp jų slypi tame, kaip intervalai parenkami: pirmuoju atveju spektrą padalijame į kelis vienodo ilgio intervalus, kiekvienam taikome mažiausių kvadratų metodą, išrenkame tuos, kurie duoda mažiausią RMSE. Antruoju atveju, intervalai atmetami pagal RMSE – baigus algoritmą lieka tiksliausiai spektrą identifikuojantys intervalai. Straipsnio autoriai teigia, kad genetiniu algoritmu paremtas analizės metodas yra tiksliausias, tačiau rezultatą reikėtų vertinti atsargiai, mat straipsnyje nurodytas metodo tikslumas siekia 100%, o taip beveik niekada nebūna.

Y.Jiang [7] ir kitų straipsnyje apdorojami duomenys gauti tyrinėjant kviečių grūdų spektrus, siekiant aptikti dygimo ženklus. Iš pradžių gautiems duomenims pritaikytas PCA algoritmas – taip duomenų dimensija sumažinta nuo 512 iki 5. Gauti duomenys suklasifikuoti į 7 klases, kurios reprezentuoja skirtingus dygimo laikus. Taip gauti pradiniai duomenys klasifikatorių treniravimui. Jų naudota trys: dalinių mažiausių kvadratų regresija (PLSR), mažiausių kvadratų atraminių vektorių mašina (LS-SVM) su Gauso branduoline funkcija ir atgalinio sužadinimo neuroninis tinklas. Anot autorių, didžiausią tikslumą davė LS-SVM klasifikatorius – apie 95%.

Hashimoto ir kitų straipsnyje [8] tyrinėjama kava infraraudonųjų spindulių spektroskopijos metodu. Straipsnio tikslas – rasti paprastą, greitą ir tikslų paruoštos kavos įvertinimo metodą. Buvo tiriamos keturios kavos rūšys: 3 arabikos iš Kolumbijos, Gvatemalos ir Brazilijos bei viena robusta iš Indonezijos. Kiekvienam mėginiui išmatuoti atspindžio spektrai, tada jie suglodinti panaudojus Savitzky-Golay metodą ir apskaičiuotos jų antrosios eilės išvestinės. Paaikškėjo, kad metodas sugeba atskirti skirtingas kavos rūšis.

2 skyrius

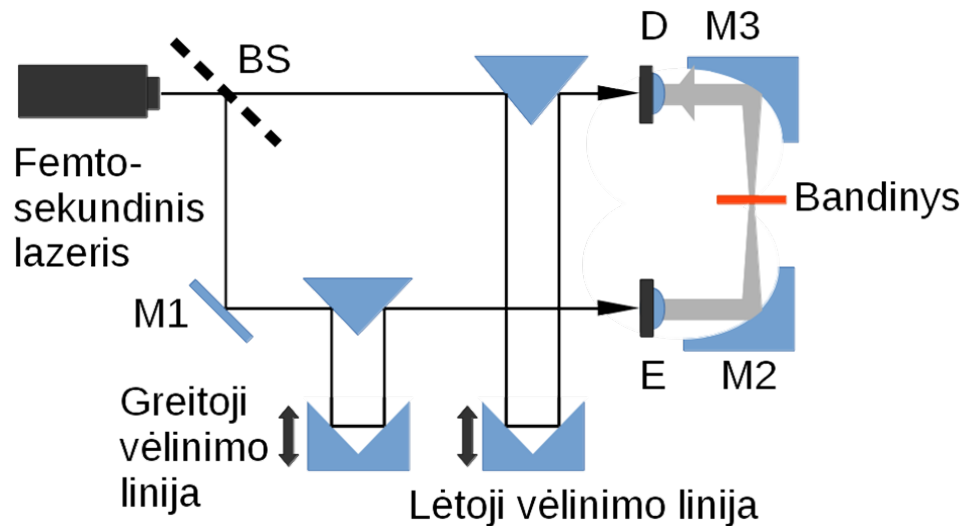
Eksperimento aprašymas ir gaunami duomenys

2.1. Eksperimento aprašymas

Furjė transformacijos spektroskopijos eksperimentui atlikti reikia šių komponentų:

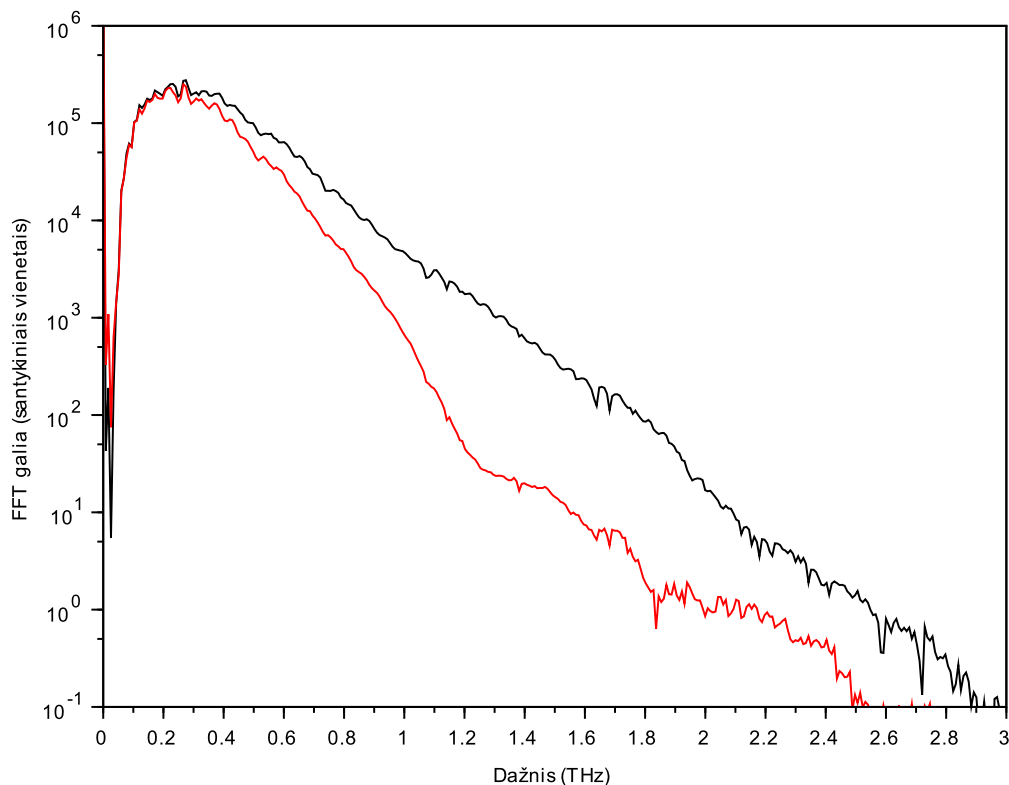
1. Lazerio, kuris generuoja femtosekundinius (10^{-15} s eilės) impulsus. Tokie trumpi impulsai reikalingi tam, kad kai jie pasiekia emiterį, šis generuotų THz spinduliuotę. Lazerio spindulys reikalingas ir detektoriui – dėl to reikia impulsus padalinti;
2. Optinio daliklio (BS), kuris padalina lazerio impulsus į žadinantį (emiteriui) ir skenuojantį (detektoriui);
3. Vėlinimo linijų, kurios užlaiko impulsus žadinančioje ir skenuojančioje linijoje vienas kito atžvilgiu, nes bandinį praėję THz spinduliai ir skenuojantis impulsas turi pasiekti detektorius tuo pačiu metu;
4. Žemos temperatūros galio arsenido (LT-GaAs) emiterio (E). Žadinantysis impulsas pasiekęs emiterį (E) generuoja THz spinduliuotę. Taip nutinka dėl to, kad femtosekundiniai impulsai savyje turi daug harmoninių dažnių ir dalis jų yra THz ruože – šie dažniai mums reikalingi, todėl juos išskiriame THz spinduliuotę praleidžiančiu filtru. THz spinduliuotė kolimuojama (surenkama į vientisą spindulį) ir fokusuojama paraboliniu veidrodžiu (M2) į bandinį;
5. LT-GaAs detektoriaus (D). Bandinį praėjusi THz spinduliuotė nukreipiama į detektorius, kuris, aktyvuotas skenuojančiu lazerio impulsu, verčia THz spinduliuotę matuojamu elektriniu signalu.

Eksperimento schema pateikiama 2.1 paveiksle.

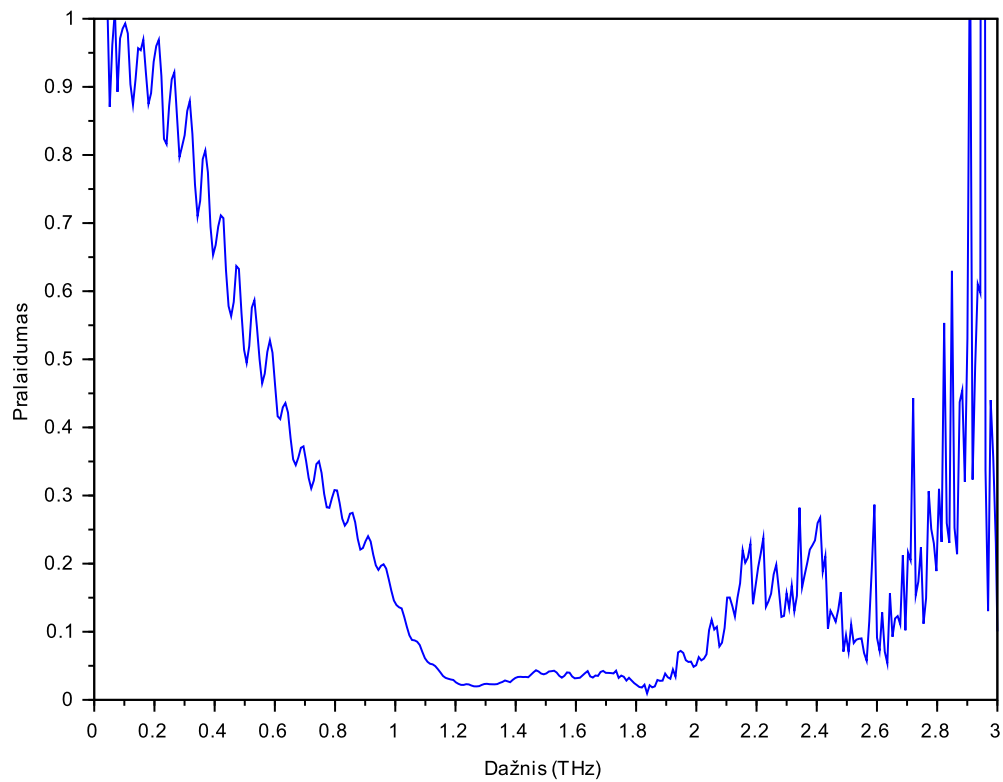


2.1 pav. Furjė transformacijos spektroskopijos eksperimento sandara

Eksperimento metu gaunamas spinduliuotės elektrinio lauko stiprio skirstinys laike. Jam atliekame Furjė transformaciją, ir gauname spektrą – THz impulso reprezentaciją dažnio srityje. Spektrų pavyzdžiai pateikiami 2.2-2.4 paveiksluose.

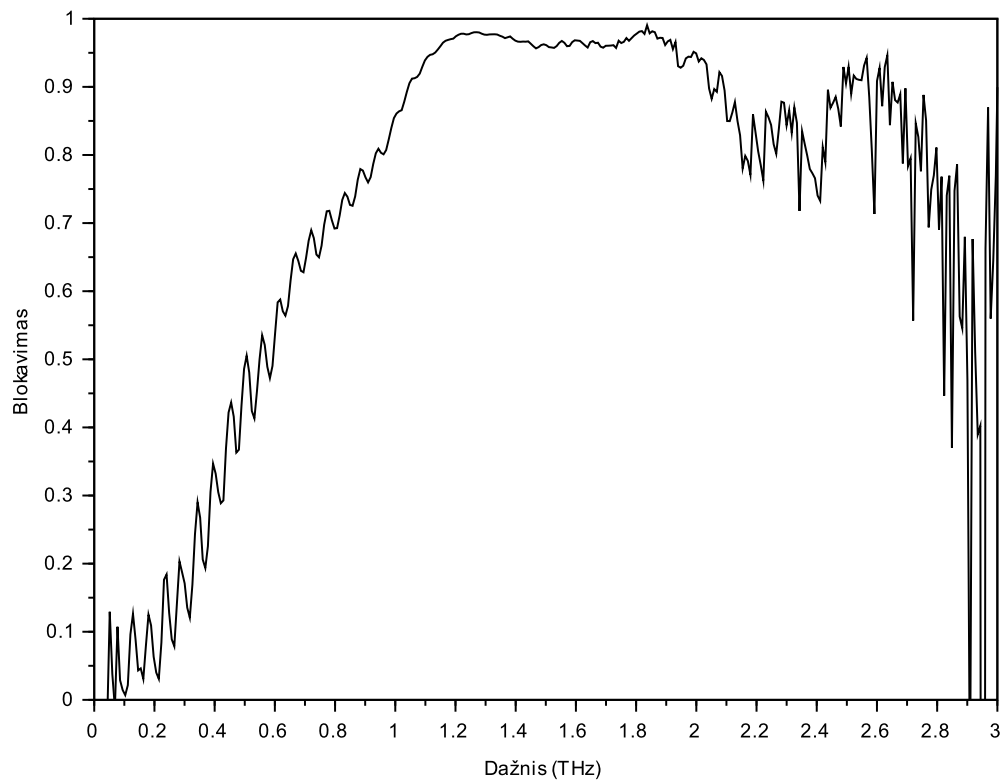


2.2 pav. Eksperimento metu gaunamų spektrų pavyzdys. Pavaizduoti kofeino (raudonas) ir azoto dujų aplinkos (juodas) FFT spektrai



2.3 pav. Eksperimento metu gaunamų spektrų pavyzdys. Pavaizduotas kofeino pralaidumo spektras $T(\omega)$, gautas padalijus kofeino FFT spektrą iš aplinkos spektro

Iš paveikslėlio dar nesimato, kad spektrą sudaro Gauso skirstiniai. Todėl atliekame transformaciją $\varphi(\omega) = 1 - T(\omega)$; transformuotas spektras pavaizduotas 2.4 paveiksle.



2.4 pav. Transformuotas kofeino pralaidumo spektras $\varphi(\omega)$

Dabar matosi, kad spektrą sudaro keletas Gauso skirstinių. Todėl išsikeltas uždavinys yra adekvatus spektrų išskaidymo problemai spręsti.

2.2. Pradinis duomenų apdorojimas Savitzky–Golay filtru

Matuodami spektrą, paprastai gauname funkcijos reikšmes lentelės pavidalu. Dėl to yra sudėtinga apskaičiuoti spektro išvestines, kurių gali prireikti optimizuojant. Be to, matavimo metu prietaisus veikia įvairūs trikdžiai – išsklaidyta šviesa, elektroninis triukšmas ir t.t. Todėl gaunamas spektras bus triukšmingas (tą matome 2.4 pav.), dėl ko jį gali būti sunku interpretuoti. Tyrimams reikia, kad signalo ir triukšmo santykis būtų kiek įmanoma didesnis, tačiau ne per didelis, nes tada yra rizika prarasti dalį informacijos. Dėl išvardintų priežasčių, reikia algoritmo, kuris padidintų spektro glodumą ir sumažintų triukšmo įtaką.

Plačiai naudojamas algoritmas yra Savitzky–Golay filtras [9]. Jo esmė – duomenis aproksimuoti polinomų seka. Kiekvienas jos narys duomenis aproksimuoja tam tikrame intervale. Tarkime, duomenų rinkinys yra n taškų (x_j, y_j) , $j = 1, 2, \dots, n$. Šie duomenys Ox ašyje turi būti nutolę vienodu atstumu h . Pradžioje atliekame transformaciją:

$$z = \frac{x - \bar{x}}{h}, \quad (2.1)$$

čia $\bar{x}$ – centrinis taškas. z yra sveiki skaičiai. Paimkime m tokių skaičių, taip, kad $z = \frac{1-m}{2}, \dots, -1, 0, 1, \dots, \frac{m-1}{2}$ (m turi būti nelyginis). Kiekvienam centriniam taškui imame k laipsnio polinomą:

$$Y = a_0 + a_1 z + a_2 z^2 + \dots + a_k z^k. \quad (2.2)$$

Polinomo koeficientai gaunami iš formulės:

$$\mathbf{a} = (a_0, a_1, a_2, \dots, a_k)^T = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \mathbf{y} = \mathbf{C} \mathbf{y}. \quad (2.3)$$

Matome, kad koeficientų vektorius $\mathbf{a}$ yra lygties

$$\mathbf{J} \mathbf{a} = \mathbf{y} \quad (2.4)$$

sprendinys mažiausių kvadratų metodu, čia

$$\mathbf{J} = \begin{pmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^k \\ 1 & z_2 & z_2^2 & \cdots & z_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_m & z_m^2 & \cdots & z_m^k \end{pmatrix}. \quad (2.5)$$

Turėdami matricą $\mathbf{C}$, galime duomenis suglodinti pagal formulę:

$$Y_w = \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} \mathbf{C}_{1i} y_{w+i}, \quad \frac{m+1}{2} \leq w \leq n - \frac{m-1}{2}. \quad (2.6)$$

Duomenų $p = 1, \dots, k$ eilės išvestines skaičiuojame taip:

$$Y_w^{(p)} = \frac{p!}{h^p} \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} \mathbf{C}_{p+1,i} y_{w+i}. \quad (2.7)$$

Suglodinti duomenys ir jų išvestinės iki p eilės apibrėžia dvi duomenų lenteles: atitinkamai (x_w, Y_w) ir $(x_w, Y_w^{(p)})$. Tačiau reikšmės taškams $x_j, j = 1, \dots, \frac{m-1}{2}$ ir $j = n - \frac{m+1}{2}, \dots, n$ nesuglodinamos – šią problemą galima apeiti pridedant fiktyvius taškus prie duomenų lentelės pradžios ir pabaigos arba imant pirmo ir paskutinio polinomo reikšmes neglodinamiems taškams.

$$\mathbf{C} = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \quad (2.8)$$

3 skyrius

Sprendimo metodai

3.1. Optimizavimo metodai

Kadangi siūlomas uždavinio sprendimas yra aproksimavimas, kuriam reikia optimizuoti paklaidą, pravartu aprašyti optimizavimo metodus [10]. Jų yra daug, todėl aptarsime pačius pagrindinius metodus: gradientinį nusileidimą, Niutono iteracijų metodą ir Lagranžo metodą.

Gradientinio nusileidimo esmė – vykdoma lokalaus minimumo paieška funkcijos $f(\mathbf{x})$ antigradiento $-\nabla f(\mathbf{x})$ kryptimi tam tikru žingsniu α . Parinkus pradinį tašką $\mathbf{x}^{(0)}$, sekančios iteracijos skaičiuojamos pagal formulę:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \alpha_i \nabla f(\mathbf{x}^{(i)}), \quad i = 0, 1, \dots \quad (3.1)$$

Skaičiuojant iteracijas, yra labai svarbu tinkamai parinkti žingsnį α_i . Jei jis yra pastovus, jį reikia parinkti nei per didelį, nei per mažą: jei jis yra per didelis, funkcijos reikšmė gali nesikeisti arba net didėti, o jei žingsnio ilgis per mažas, tada algoritmas lėtai konverguos į minimumą. Tačiau jei funkcija turi griovių, tada net ir tinkamai parinkus žingsnį, konvergavimas bus lėtas. Tada kiekvienoje iteracijoje galima žingsnį keisti adaptyviai. Pavyzdžiui, žingsnį galima parinkti sprendžiant vienmatį optimizavimo uždavinį:

$$\min_{\alpha} f(\mathbf{x}^{(i)} + \alpha \nabla f(\mathbf{x}^{(i)})) \quad (3.2)$$

Kartais būna taip, kad funkcija nėra glodi, arba jos išvestinės nežinomos. Tada gradientus galima skaičiuoti skaitiškai – baigtinių skirtumų metodu. Bet tai gali sukelti problemų, nes skaitinis diferenciacijavimas yra skaitiškai nestabilus – dėl apvalinimo paklaidų, skaitinio diferenciacijavimo rezultatų tikslumas mažesnis nei argumento tikslumo.

Niutono iteracijų metodas panašus į gradientinį nusileidimą, tik nusileidimo kryptis

skaičiuojama sudėtingiau. Metodas paremtas tuo, kad ieškome lygties $\nabla f(\mathbf{x}) = 0$ sprendinio Niutono metodu. Todėl iteracijas skaičiuojame taip:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \alpha_i [\mathbf{H}f(\mathbf{x}^{(i)})]^{-1} \nabla f(\mathbf{x}^{(i)}), \quad i = 0, 1, \dots, \quad (3.3)$$

čia $\mathbf{H}f(f(\mathbf{x}))$ – funkcijos $f(\mathbf{x})$ Hesianas taške $f(\mathbf{x})$. Tai yra funkcijos antrųjų išvestinių matrica, todėl reikalaujama, kad funkcija būtų dukart diferencijuojama. Ją reikia invertuoti, o tam reikia atlikti daug skaičiavimų, todėl dažniausiai nusileidimo kryptį $\Delta \mathbf{x}_i$ gauname sprenddami lygtį:

$$[\mathbf{H}f(\mathbf{x}^{(i)})] \Delta \mathbf{x}_i = -\nabla f(\mathbf{x}^{(i)}). \quad (3.4)$$

Sprendimo metodas priklauso nuo Hesiano apibrėžtumo, o tai priklauso nuo to, kokių funkcijos ekstremumų ieškome. Jei ieškome minimumo, Hesianas turi būti teigiamai apibrėžtas, ir lygtį sprendžiame metodais, kurie tinka tokioms matricoms (pvz. Choleskio metodu). Jei ieškome balno taško (to reikia Lagranžo metodui), Hesianas simetrinis, neapibrėžtas, tada naudojame atitinkamą metodą (pvz. Choleskio $\mathbf{LDL}^T$ išskaidymą).

Lagranžo metodas skirtas optimizavimui su lygybės tipo apribojimais, kitaip tariant, tokio uždavinio sprendimui:

$$\text{rasti } \min_{\mathbf{x}} f(\mathbf{x}), \text{ su apribojimais } \mathbf{g}(\mathbf{x}) = \mathbf{c}. \quad (3.5)$$

Taikydami metodą, sudarome Lagranžo funkciją:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda} \cdot (\mathbf{g}(\mathbf{x}) - \mathbf{c}), \quad (3.6)$$

čia $\boldsymbol{\lambda}$ – vektorius, kurio koordinatės vadinamos Lagranžo daugikliais. Turėdami funkciją, ieškome jos *balno* taškų. Optimizuojant skaitiškai, tai sukelia problemų, nes dažniausiai skaitiniai metodai ieško minimumo (maksimumo). Tai galime išspręsti naudodami jau minėtą Niutono metodą balno taško paieškai. Dar viena galimybė – optimizuoti Lagranžo funkcijos gradiento kvadratą $\|\nabla \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})\|^2$. Tada balno taškai paverčiami minimumais, ir galima naudoti reikiamus minimizavimo metodus.

3.2. Atraminių vektorių mašinos

Yra dar vienas būdas mišinio išskaidymo uždaviniui spręsti. Užuoat ieškojus spektrinių požymių aproksimuojant, galima sudaryti spektrų rinkinį, kuriame būtų įvairūs spektrai mišinių, turinčių dominančios organinės medžiagos; kartu su jais turi būti ir spektrai, neturintys tos medžiagos. Tada, naudojantis šiuo duomenų rinkiniu, galime parinkti kla-

sifikavimo algoritmo parametrus taip, kad jis galėtų pakankamai tiksliai suklasifikuoti naujus duomenis.

Šiame skyrelyje aptarsime vieną plačiai naudojamą klasifikavimo algoritmą – atrami-
nių vektorių mašiną [11] (angl. *support vector machine*, SVM). Jo esmė – duomenų rinkinį
(x_i, y_i), $y_i = \pm 1$, $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$ atskiriame hiperplokštuma taip, kad vienoje jos pu-
sėje būtų tie x_i , kuriems $y_i = -1$, o kitoje – tie x_i , kuriems $y_i = 1$. Be to, hiperplokštumos
atstumas nuo bet kurios klasės turi būti kiek įmanoma didesnis.

Hiperplokštumos lygtis yra $\omega^T x - b = 0$. Nuo jos atidedame dar dvi lygiagrečias
plokštumas: $\omega^T x - (b - 1) = 0$ ir $\omega^T x - (b + 1) = 0$. Tarp šių plokštumų yra erdvės sritis,
vadinama *marža*. Maržos plotis yra atstumas tarp tų plokštumų, ir yra lygus $\frac{2}{\|\omega\|}$. Jis
turi būti kuo didesnis, t. y. $\|\omega\|$ turi būti kuo mažesnis. Dar yra svarbu, kad jokie taškai
nepakliūtų į maržą. Todėl reikia paisyti šių apribojimų:

$$\begin{aligned} \omega^T x_i - (b + 1) &\geq 0, \quad y_i = 1; \\ \omega^T x_i - (b - 1) &\leq 0, \quad y_i = -1. \end{aligned} \quad (3.7)$$

Kadangi $y_i = \pm 1$, apribojimus galima užrašyti trumpiau: $y_i(\omega^T x_i - b) \geq 1$, $\forall i = 1, \dots, n$.
Matome, kad turime optimizavimo uždavinį su apribojimais: rasti

$$\min_{\omega} \|\omega\|, \quad (3.8)$$

kad būtų tenkinami apribojimai

$$y_i(\omega^T x_i - b) \geq 1, \quad \forall i = 1, \dots, n. \quad (3.9)$$

(3.8-3.9) uždavinys tinka tik tuo atveju, kai klasės yra tiesiškai atskiriamos, t. y.
nesusipynusios. Jei taip nėra, į optimizavimo uždavinį turime įskaičiuoti paklaidas, ku-
rios atsiranda dėl taškų patekimo į netinkamą maržos pusę. Tai padarome, įveddami
praradimo funkcijas:

$$\zeta_i = \max(0, 1 - y_i(\omega^T x_i - b)), \quad \forall i = 1, \dots, n. \quad (3.10)$$

Funkcija lygi 0, jei tenkinami (3.9) apribojimai, ir $\|\omega\| = \text{dist}(x_i, \{\omega^T x - b = 0\})$, jei
apribojimai netenkinami. Tada optimizavimo uždavinys bus toks: rasti

$$\min_{\omega, \zeta_i} \sum_{i=1}^n \zeta_i + \mu \|\omega\|, \quad (3.11)$$

kad būtų tenkinamos sąlygos

$$\begin{aligned}\zeta_i &\geq 0; \\ y_i(\omega^T x_i - b) &\geq 1 - \zeta_i.\end{aligned}\tag{3.12}$$

Sąlygos gaunamos iš maksimumo savybių. Parametras μ vadinamas normalizavimo parametru, juo reguliuojama algoritmo paklaida.

(3.11-3.12) uždavinį galima supaprastinti, apskaičiavus Lagranžo dualųjį uždavinį:

$$\max_{c_i} \sum_{i=1}^n c_i + \sum_{i=1}^n \sum_{j=1}^n c_i y_i (x_i \cdot x_j) y_j c_j,\tag{3.13}$$

su apribojimais

$$\begin{aligned}\sum_{i=1}^n y_i c_i &= 0; \\ 0 &\leq c_i \leq \frac{1}{2n\mu}.\end{aligned}\tag{3.14}$$

Kai μ yra pakankamai didelis, šis uždavinys randa plokštumą, skiriančią tiesiškai atskiriamas klases. Uždavinio sprendinys parametą ω apibrėžia taip:

$$\omega = \sum_{i=1}^n y_i c_i x_i.\tag{3.15}$$

Parametrą b randame, paimdami kokį nors x_i ant maržos krašto, ir apskaičiuodami klasifikatoriaus rezultata. Ant maržos krašto esantys taškai vadinami *atraminiais vektoriais*. Nauji taškai klasifikuojami pagal formulę: $\text{sgn}(\omega^T x - b)$.

Jei turime tiesiškai neatskiriamas klases, galime apibrėžti ir netiesinį klasifikatorių. Imkime hiperplokštumą tokiu pavidalu: $\omega^T \varphi(x) - b = 0$. Funkciją $\varphi(x)$ apibrėžia branduolinė funkcija K , taip: $K(x, y) = \varphi(x) \varphi(y)$. Parametro ω ieškome sprenddami (3.13-3.14) uždavinį, tik vietoje skaliarinės sandaugos $x_i \cdot x_j$ imame branduolinę funkciją $K(x_i, x_j)$. Tada:

$$\omega = \sum_{i=1}^n y_i c_i \varphi(x_i).\tag{3.16}$$

Neretai funkcija φ neturi analitinės išraiškos. Tada galima pasinaudoti branduolinės funkcijos savybe, ir iš karto gauti $\omega^T \varphi(x)$ išraišką:

$$\omega^T \varphi(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x).\tag{3.17}$$

Nauji taškai klasifikuojami analogiškai skaliarinės sandaugos atvejui.

3.3. Pagrindinių komponentų analizė

Spektroskopijos metu gaunami spektrai, kuriuos galime užrašyti kaip daugiamačius vektorius – kiekviena koordinatė reiškia intensyvumą konkrečiam bangos ilgiui. Šių vektorių dimensija yra didelė – 100 eilės, o skaičiavimai tokiais vektoriais yra sudėtingi ir užima daug laiko. Todėl reikia duomenų dimensiją sumažinti, bet mažindami duomenų dimensiją, neišvengiamai prarandame informacijos. Reikalingas toks metodas, kuris prarastų mažiausiai informacijos. Vienas iš tokių yra *pagrindinių komponentų analizė* (angl. *principal component analysis*, PCA) – tiesinė ortogonalinė transformacija, kuri sumažina duomenų dimensiją ir išsaugo kiek įmanoma daugiau variacijos [11]. Procesas vyksta taip: tarkime, duomenys surašyti $n \times p$ matricoje $\mathbf{X}$, kur n – mėginių, o p – atributų skaičius. Jos stulpelius centruojame – iš kiekvieno stulpelio atimame to stulpelio vidurkį, kad kiekvieno stulpelio vidurkis būtų 0. Dar galima stulpelius normuoti, tačiau algoritmo rezultatas priklauso nuo normavimo būdo, todėl jį reikia rinktis atsargiai. Iš $\mathbf{X}$ gauname kovariacijų matricą $\mathbf{X}^T \mathbf{X}$. Apskaičiavus šios matricos tikrines reikšmes ir vektorius, gauname $p \times p$ matricą $\mathbf{W}$, kurios stulpeliai – mūsų gautieji tikriniai vektoriai, normuoti, kad jų ilgis būtų 1. Tikrinės reikšmės naudojamos patikrinti, kokią duomenų variacijos dalį reikšmę atitinkantis vektorius reprezentuoja. Ta dalis vektoriui i skaičiuojama pagal formulę:

$$\frac{\alpha_i}{\sum \alpha_i}, \quad (3.18)$$

čia α_i – tikrinį vektorių i atitinkanti tikrinė reikšmė.

Dimensiją sumažiname, palikdami L , $1 \leq L \leq p$, stulpelių matricoje $\mathbf{W}$ – tiek, kiek reikia, kad būtų išsaugota didžioji dalis variacijos. Tokią redukuotą matricą žymėsime $\mathbf{W}_L$. Mažesnės dimensijos duomenis, kurių matricą žymėsime $\mathbf{T}$, gausime, apskaičiavę:

$$\mathbf{T} = \mathbf{X} \mathbf{W}_L. \quad (3.19)$$

Galimas ir kitas būdas, kuriam nereikia skaičiuoti kovariacijų matricos – singularusis išskaidymas (angl. *singular value decomposition*, SVD). Matricą $\mathbf{X}$ galime išskaidyti taip:

$$\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{W}^T, \quad (3.20)$$

čia $\mathbf{U}$ – $n \times n$ matrica, $\mathbf{W}$ – $p \times p$ matrica, jų stulpeliai – atitinkamai *kairieji* ir *dešinieji singularieji vektoriai*; $\mathbf{\Sigma}$ – $n \times p$ diagonalinė matrica, kurios įstrižainėje yra *singularios reikšmės*. Atlikus išskaidymą, matricose $\mathbf{U}$ ir $\mathbf{\Sigma}$ paliekame L didžiausių singulariųjų reikšmių ir jas atitinkančių kairiųjų singulariųjų vektorių, tada mažesnės dimensijos duomenų matrica bus:

$$\mathbf{T} = \mathbf{U}_L \mathbf{\Sigma}_L. \quad (3.21)$$

3.4. Diskriminantinė analizė

Naudodami PCA, sumažiname duomenų dimensiją, palikdami kiek įmanoma daugiau variacijos. Tačiau kartais reikia sumažinti duomenų dimensiją taip, kad aiškiai matytųsi atskiros duomenų klasės. Tokiu atveju naudojame diskriminantinės analizės metodą (angl. *discriminant analysis*, DA) [11]. Tai klasifikavimo metodas, kuris mažina duomenų dimensiją kartu padidindamas atskirtį tarp klasių. Kaip ir SVM atveju, šiam metodui reikia treniravimo rinkinio: duomenų matricos $\mathbf{X}$ ir klasių žymių vektoriaus y (jo koordinatės gali būti arba 0, arba 1). Daroma prielaida, kad kiekvienos klasės duomenys pasiskirstę pagal normalųjį skirstinį; abi klasės turi parametrus (μ_i, Σ_i) , čia $i = 0, 1$, μ - vidurkių vektorius, Σ - kovariacijų matrica. NAGRINĖSIME tik tiesinės diskriminantinės analizės (LDA) atvejį. Tada duomenys projektuojami į hiperplokštumą $\omega^T x - c = 0$; dydis $\omega^T x$ vadinamas *tiesine diskriminantine funkcija*. Jei tenkinama sąlyga $\omega^T x > c$, tada duomuo x priklauso klasei 1. Plokštumos parametrus randame taip: ω gauname sprendami optimizavimo uždavinį 3.22.

$$\max S(\omega) = \max \frac{\sigma_{bw}^2}{\sigma_{wt}^2} = \max \frac{(\omega(\mu_1 - \mu_0))^2}{\omega^T(\Sigma_1 + \Sigma_0)\omega}, \quad (3.22)$$

čia σ_{bw}^2 - dispersija tarp klasių, σ_{wt}^2 - dispersija klasių viduje. Tikslo funkcijos prasmė tokia: norime padidinti skaitiklyje esančią σ_{bw}^2 reikšmę ir sumažinti vardiklyje esantį narį σ_{wt}^2 ; kitaip tariant, mes didiname atskirtį tarp klasių vidurkių ir mažiname dispersiją klasių viduje. Tiesiniu atveju, tikslo funkcija pasiekia maksimumą tada, kai:

$$\omega \propto (\Sigma_1 + \Sigma_0)^{-1}(\mu_1 - \mu_0). \quad (3.23)$$

Kitas parametras, c , randamas suprojektavus duomenis į hiperplokštumą ir išanalizavus jų projekcijas. Jei abiejų klasių projekcijų pasiskirstymai beveik vienodi, tada:

$$c = \frac{1}{2}\omega(\mu_0 - \mu_1). \quad (3.24)$$

3.5. Mažiausių kvadratų metodas

Turint tam tikrus duomenis, neretai prireikia rasti dėsni, siejantį juos, ir juo naudojantis numatyti tam tikros jų savybės reikšmę naujiems duomenims. Kalbant matematiškai, reikia sugebėti aproksimuoti tam tikrą funkciją, priklausančią nuo vektoriaus – šis procesas vadinamas *regresija* [12]. Aproksimavimo būdas priklauso nuo duomenų ir paklaidos matavimo būdo. Vieną iš tokių dabar ir aptarsime – tai yra *mažiausių kvadratų metodas* (angl. *method of least squares*, MLS). Turėdami mėginių matricą $\mathbf{X}$, kiekvienam iš jų

priskiriame tam tikrą funkcijos reikšmę – visas jas suvedame į n -vektorių $\mathbf{y}$.

Mūsų tikslas – rasti funkcijos aproksimaciją pavidalu:

$$f(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}, \quad (3.25)$$

čia $\boldsymbol{\beta}$ - nežinomų parametrų vektorius ($p \times 1$). Taigi, uždavinys susiveda į $\boldsymbol{\beta}$ paiešką. Mažiausių kvadratų metode, parametrų vektorius randamas sprendžiant optimizavimo uždavinį:

$$\boldsymbol{\beta} = \arg \min \frac{1}{2n} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2. \quad (3.26)$$

Paprastai tariant, mums reikia tokio $\boldsymbol{\beta}$, kuris duotiems duomenims minimizuotų vidutinę paklaidą tarp $\mathbf{X}\boldsymbol{\beta}$ ir $\mathbf{y}$.

Funkcija $f(\mathbf{x})$ gali būti ir daugiamatis polinomas – optimizavimo uždavinio tikslo funkcija išliks tokia pati, tik aukštesniųjų laipsnių nariai bus laikomi papildomais požymiais.

Naudodami šį metodą, susiduriame su tendencingumo-variacijos (angl. *bias-variance*) problema. Jei naudojame pernelyg paprastą aproksimavimo dėsnį, palyginus su duomenų struktūra (pavyzdžiui, tiesinę funkciją, kai duomenys turi didelę kreivį), gauname *didelio tendencingumo* (angl. *high bias*) aproksimatorių. O jeigu dėsnis pernelyg sudėtingas (pvz., turime nedaug duomenų, o naudojame didelio laipsnio daugianarį), tada aproksimatorius bus *didelės variacijos* (angl. *high variance*). Dėl to atsiranda nepageidaujamos paklaidos, dėl kurių aproksimatoriaus rezultatai smarkiai skirsis nuo laukiamų. Todėl reikalingas metodas aproksimatoriaus paklaidai reguliuoti – jis vadinamas *regularizacija*. Jo esmė – prie tikslo funkcijos pridedame regularizavimo narį:

$$C \sum_{k=0}^p \beta_k, \quad (3.27)$$

čia C - regularizavimo parametras. Keisdami jo reikšmę, reguliuojame aproksimatoriaus paklaidą. Jei C - didelis, β_k reikšmės smarkiai sumažinamos, ir gauname didelio tendencingumo aproksimatorių. Jei C - mažas, β_k neribojami, ir gauname didelės variacijos aproksimatorių. Tinkamai parinkus C galima sumažinti aproksimavimo paklaidą iki pakankamos reikšmės laukiamiems rezultatams gauti.

4 skyrius

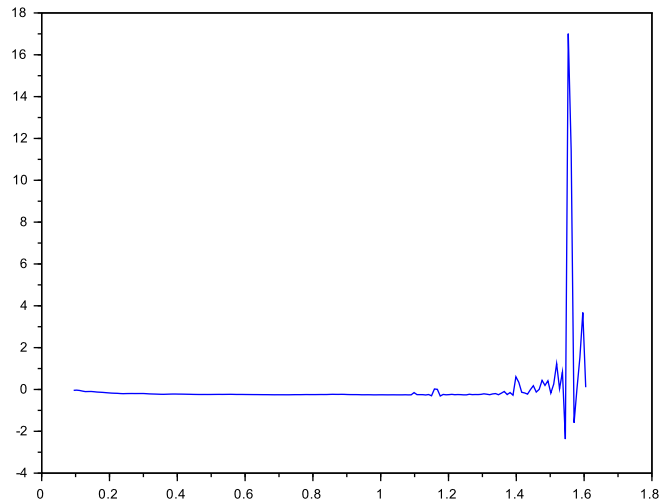
Metodų taikymas duomenims

Šiame skyriuje aprašomi skaičiavimų, taikant minėtus sprendimo metodus, rezultatai. Tyrimams naudoti dviejų tipų duomenys – realūs ir dirbtiniai.

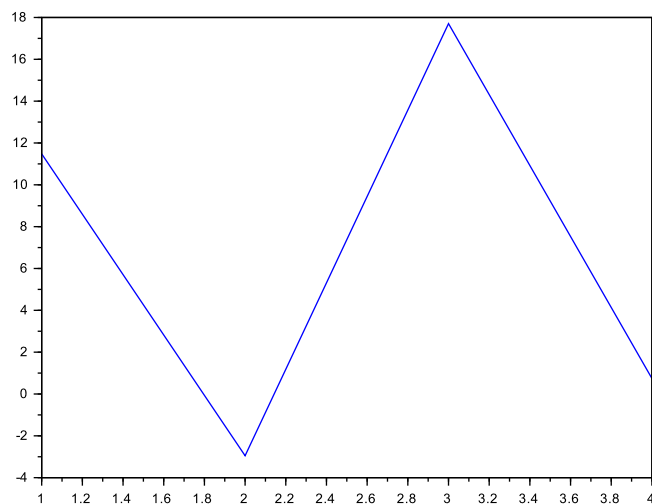
4.1. Pagrindinių komponentių analizė

Pradžioje bandyta pritaikyti pagrindinių komponentių analizės metodą kokiems nors duomenims. Mūsų atveju duomenys yra realūs, gauti tikrinant mėginius spektrometru; tai yra 24 spektrai, sudaryti iš 177 komponentių (nepriklausomo kintamojo reikšmių). Uždavinio tikslas – supaprastinti spektrus taip, kad prarastume kiek įmanoma mažiau informacijos.

Duomenų matrica yra centruojama, paskui išskaidoma – tam panaudojome singuliarųjį išskaidymą. Algoritmas išskyrė keturias komponentes: kiekviena iš jų išsaugo atitinkamai 95,91%, 2,29%, 0,94% ir 0,34% variacijos – iš viso 99,48%. Rezultatai vienam spektrui pateikiami 4.1 ir 4.2 paveiksluose.

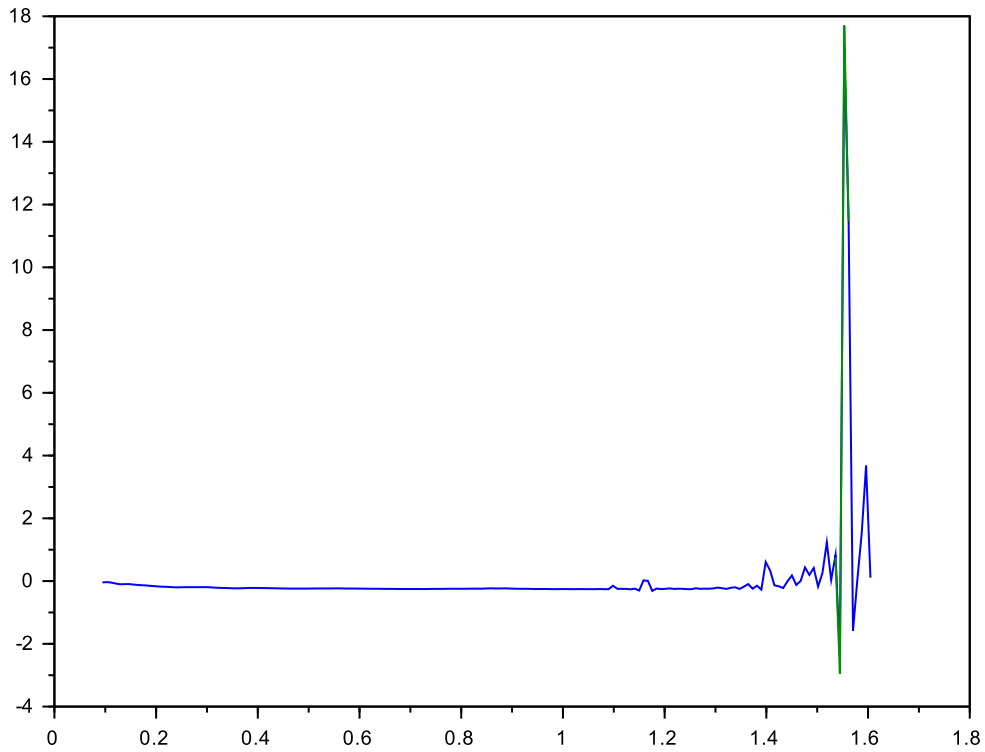


4.1 pav. Spektro #3 grafikas po normavimo



4.2 pav. Spektro #3 grafikas po singuliariojo išskaidymo ir komponentių išskyrimo

Toliau siekta komponentes interpretuoti, kas nėra lengva. Kiekvienam spektrui pagrindinės komponentės atvaizduoja pačią žymiausią spektro vietą. 4.3 paveiksle pavaizduota komponentių interpretacija trečiajam spektrui. Didžiausias keblumas yra tas, kad nežinome, kokias nepriklausomojo kintamojo reikšmės atitinka pagrindinės komponentės. Mūsų atveju komponentių ir kintamojo reikšmių ryšys rastas lyginant spektrų amplitudes su komponentių reikšmėmis.



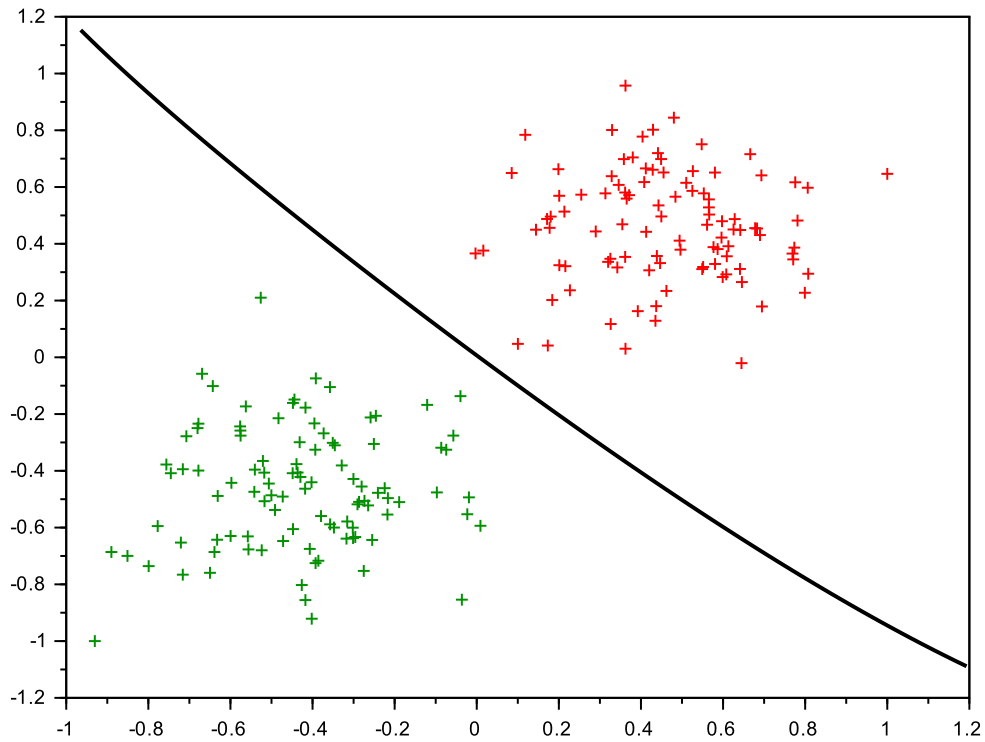
4.3 pav. Spektro #3 grafikas (mėlyna) ir jo pagrindinės komponentės (žalia) kartu

4.2. Atraminių vektorių mašina

Atraminių vektorių algoritmas buvo išbandytas su sintetiniais duomenimis. Skaičiavimai atlikti Scilab 5.5.2 matematiniu paketu [13], naudojantis LIBSVM [14] biblioteka.

Pirmiausia generuotas tiesiškai atskiriamų klasių pavyzdys. Abi klases sudarė nepriklausomi, vienodai pasiskirstę Gauso vektoriai: pirmos klasės vidurkis buvo $m_1 = (-9, -9)$, dispersija $\text{diag}(\Sigma_1) = (16, 16)$; antros klasės vidurkis $m_2 = (9, 9)$, dispersija $\text{diag}(\Sigma_2) = \text{diag}(\Sigma_1)$. Abi klases sudarė po 100 taškų. Klasės normuotos taip, kad kiekvienos koordinatės reikšmė būtų intervale $[-1, 1]$. Pirmai klasei priskirtas klasės numeris 0, antrai – 1.

Naudotas netiesinis SVM klasifikatorius su Gauso branduoline funkcija: $K(u, v) = \exp\{-\gamma \|u - v\|^2\}$. Pirmame pavyzdyje parinkti tokie parametrai: $\mu = 1$, $\gamma = 0,5$. Klasės ir klasifikavimo rezultatai pavaizduoti paveiksle 4.4.

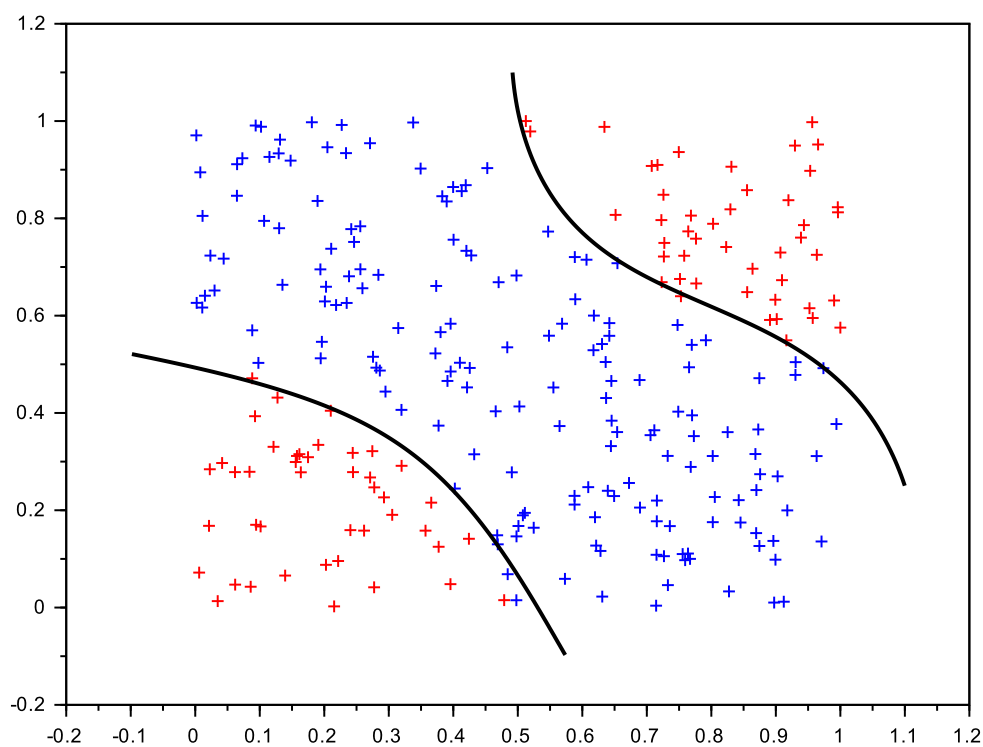


4.4 pav. Tiesiškai atskiriamų klasių pavyzdys, čia žalia spalva reiškia klasę 0; raudona – klasę 1

Paskui generuotas tiesiškai neatskiriamų klasių pavyzdys. Pradžioje generuoti trys klasteriai, kuriuos sudarė nepriklausomi, vienodai pasiskirstę tolygieji vektoriai, srityje $[0, 2] \times [0, 2]$:

1. (X_1, Y_1) , tokie, kad $Y_1 < 1 - X_1^2$;
2. (X_2, Y_2) , tokie, kad $1 - X_2^2 < Y_2 < 1 + (X_2 - 2)^2$;
3. (X_3, Y_3) , tokie, kad $Y_3 > 1 + (X_3 - 2)^2$.

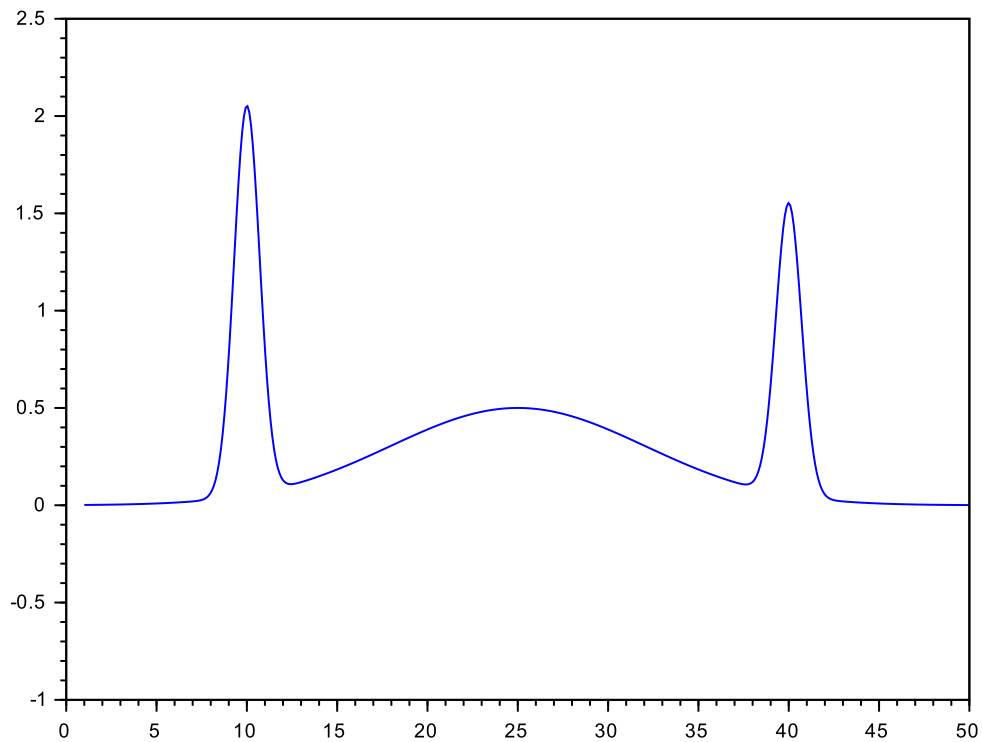
Klasterius 1 ir 3 sujungėme į vieną klasę su žymeniu 0, o klasteriui 2 priskyrėme žymenį 1. Kaip ir pirmame pavyzdyje, duomenis klasifikavome naudodami Gauso SVM, su tokiais parametrais: $\mu = 10^5$, $\gamma = 4$. Rezultatas pavaizduotas 4.5 paveiksle.



4.5 pav. Tiesiškai neatskiriamų klasių pavyzdys, čia raudona spalva reiškia klasę 0; mėlyna – klasę 1

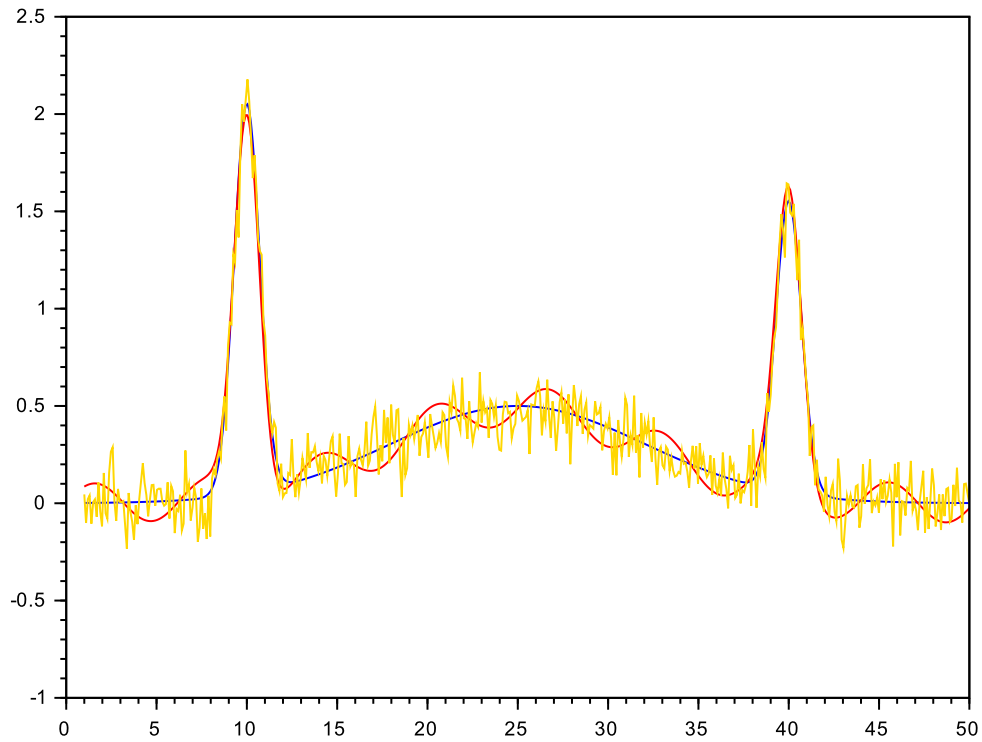
4.3. Optimizavimo uždavinio sprendimas

Šiame skyrelyje aprašomi skaičiavimai (1.5)-(1.6) uždavinio sprendimui. Iš pradžių spektras $\varphi(\omega)$ apibrėžtas kaip trijų komponentų suma. Šis spektras pavaizduotas 4.6 paveiksle.



4.6 pav. Sintetinis spektras, naudotas skaitiniam tyrimui. Jo komponentų parametrai yra, iš kairės į dešinę atitinkamai: viršūnių aukščiai yra $(2, 0, 5, 1, 5)$; spektriniai pločiai – $(1, 10, 1)$; viršūnių padėtys – $(10, 25, 40)$

Padarėme prielaidą, kad komponentų parametrai nežinomi, ir taikėme optimizavimo metodus su įvairiais pradiniais artiniais. Naudoti du metodai: lokalus kvaziniutono BFGS metodas ir globalus atkaitinimo modeliavimo (SA) metodas. Iš pradžių netaikyti apribojimai, nes norėta patikrinti, ar jie reikalingi. Optimizavimas atliktas originaliam spektrui ir dviem kitiems, gautiems deformavus originalų. Vienam pridėtas sinusinis iškraipymas su amplitudė 0,1 ir dažniu 1, o kitam – Gauso triukšmas su nuliniu vidurkiu ir dispersija 0,01. Deformuoti spektrai pavaizduoti 4.7 paveiksle.



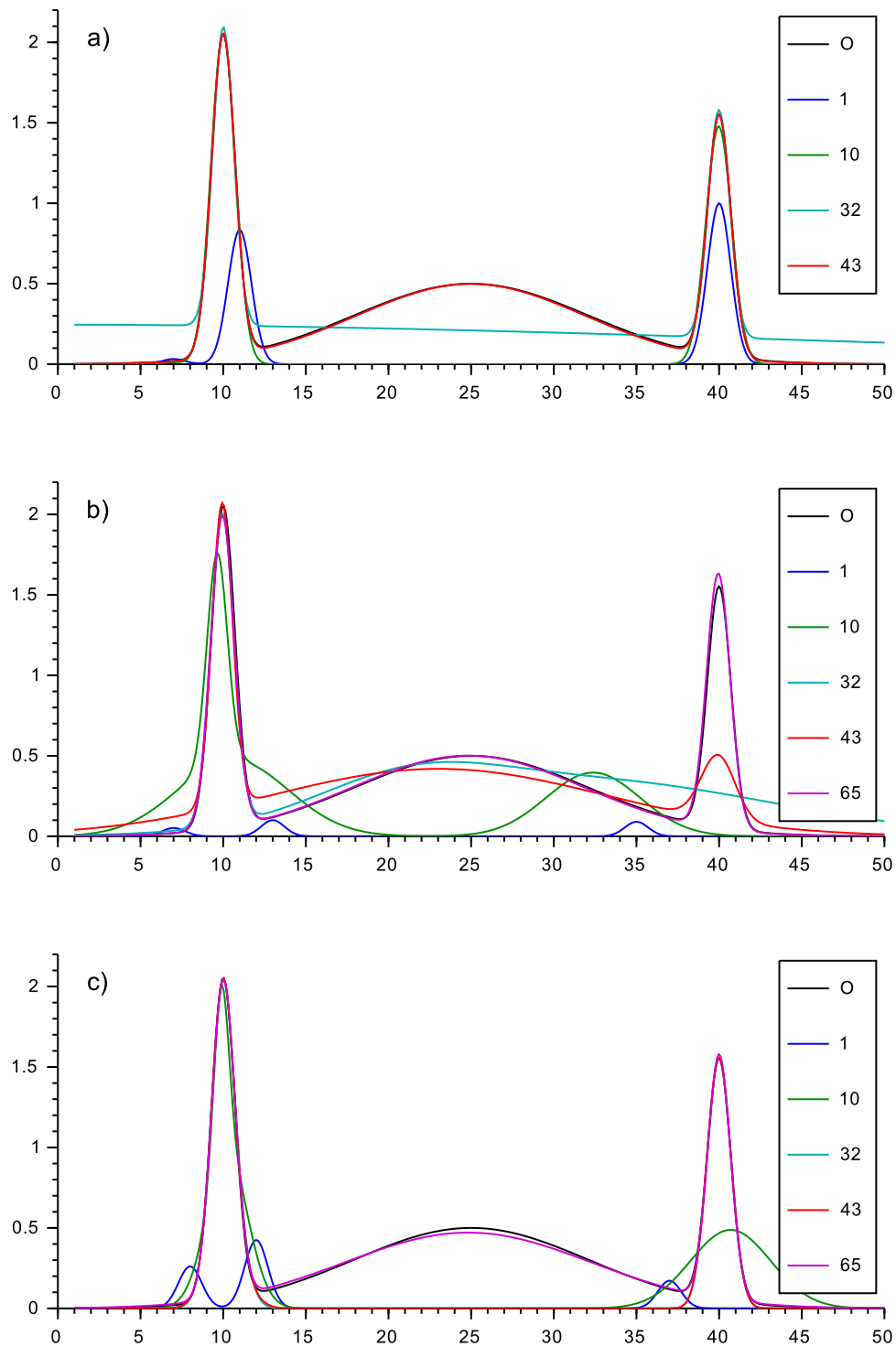
4.7 pav. Deformuoti spektrai. Čia mėlynas spektras yra originalus, raudonas – deformuotas sinusu, geltonas – deformuotas Gauso triukšmu

Pradiniai artiniai parinkti taip, kad jie būtų kuo arčiau atskaitos taško, t.y., kiekvienas α_i ir γ_i būtų kiek įmanoma arčiau 0, o β_i – arčiau 1; be to, artiniai dar parinkti taip, kad optimizavimo algoritmai iš jų artėtų į globalų minimumą. Skaičiavimų rezultatai pateikiami 4.1 lentelėje.

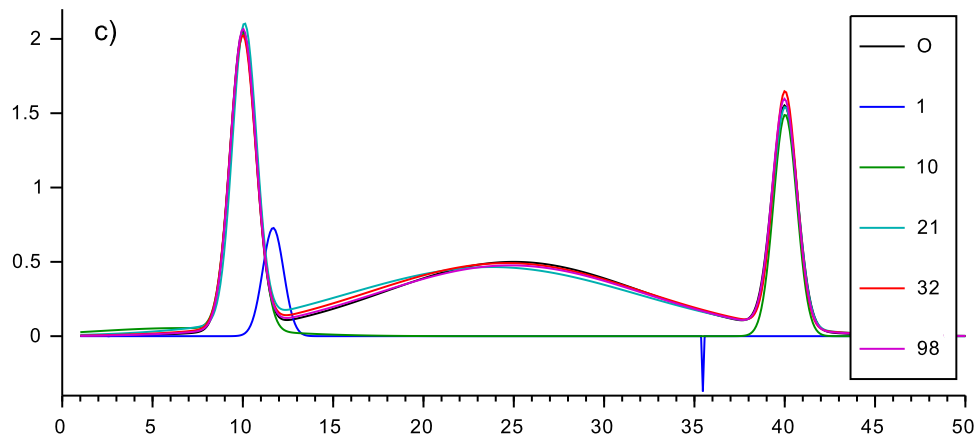
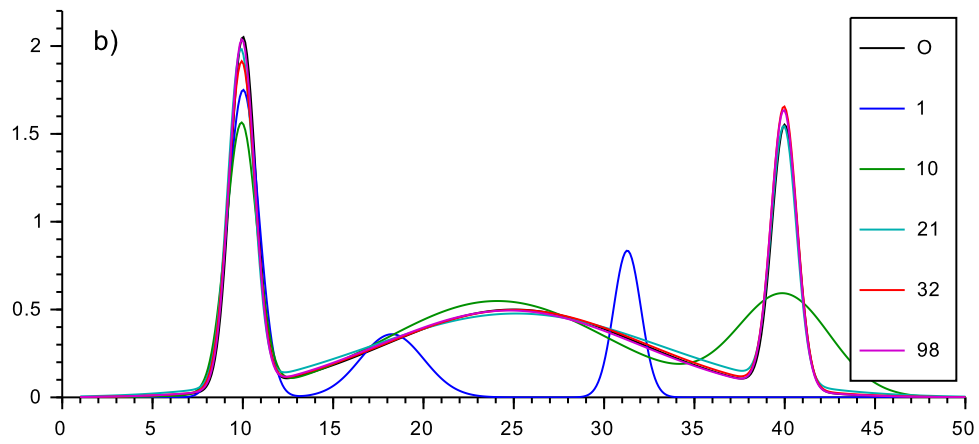
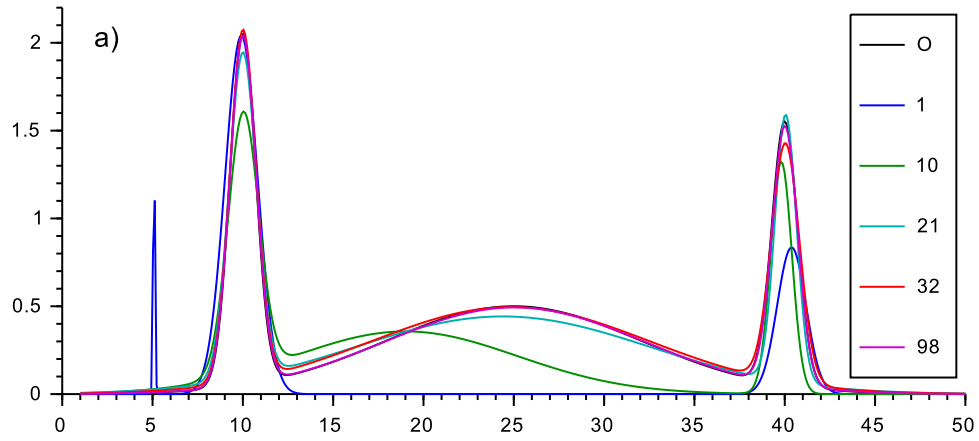
4.1 lentelė. Optimizavimo pradinių artinių lentelė. Čia deformuotas 1 yra pradinis spektras su sinusiniu iškraipymu, o deformuotas 2 – su Gauso triukšmu. Pirmojoje eilutėje pateikti originalaus spektro parametrai

Spektras	Metodas	α	β	γ
Pradinis	–	(2; 0, 5; 1, 5)	(1; 10; 1)	(10; 25; 40)
Pradinis	BFGS	(0; 0; 0)	(1; 1; 1)	(7; 11; 40)
	SA	(0; 0; 0)	(1; 1; 1)	(5; 10; 40)
Deformuotas 1	BFGS	(0; 0; 0)	(1; 1; 1)	(7; 13; 35)
	SA	(0; 0; 0)	(1; 1; 1)	(6; 19; 30)
Deformuotas 2	BFGS	(0; 0; 0)	(1; 1; 1)	(8; 12; 37)
	SA	(0; 0; 0)	(1; 1; 1)	(6; 12; 36)

Toliau pateikiami paveikslai parodo, kaip spektrai artėja prie sprendinio.

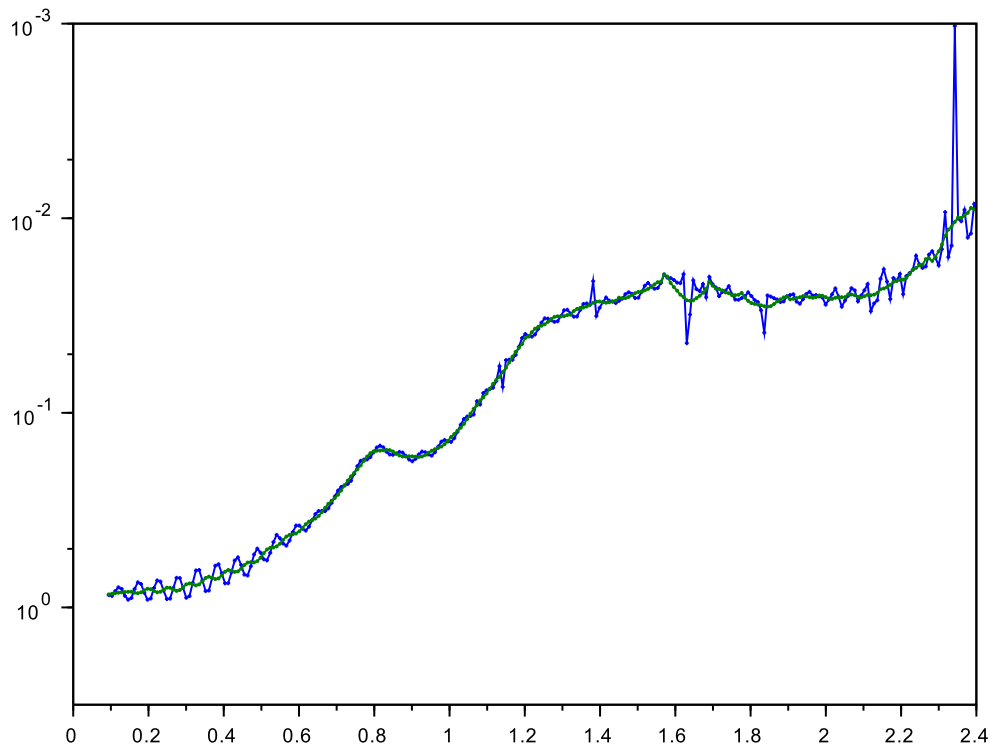


4.8 pav. Spektrai, gaunami tam tikrose BFGS optimizavimo metodo iteracijose, kai aproksimuojamas spektras yra: a) pradinis, b) deformuotas 1, c) deformuotas 2, iš 4.1 lentelės. Legendose skaičiai reiškia iteracijos numerį; O – tikrasis sprendinys



4.9 pav. Spektrai, gaunami tam tikrose SA optimizavimo metodo iteracijose. Žymėjimai tokie patys, kaip ir 4.8 paveiksle

Paskui bandyta (1.5)-(1.6) uždavinį spręsti su realiais duomenimis. Toks uždavinys sudėtingesnis, nes tada jau nebeaišku, ar spektrą sudaro Gauso funkcijų tiesinė kombinacija. Paveiksle pavaizduotas realaus spektro pavyzdys. Matome, kad net jį suglodinusi sunku įžiūrėti Gauso funkcijas.

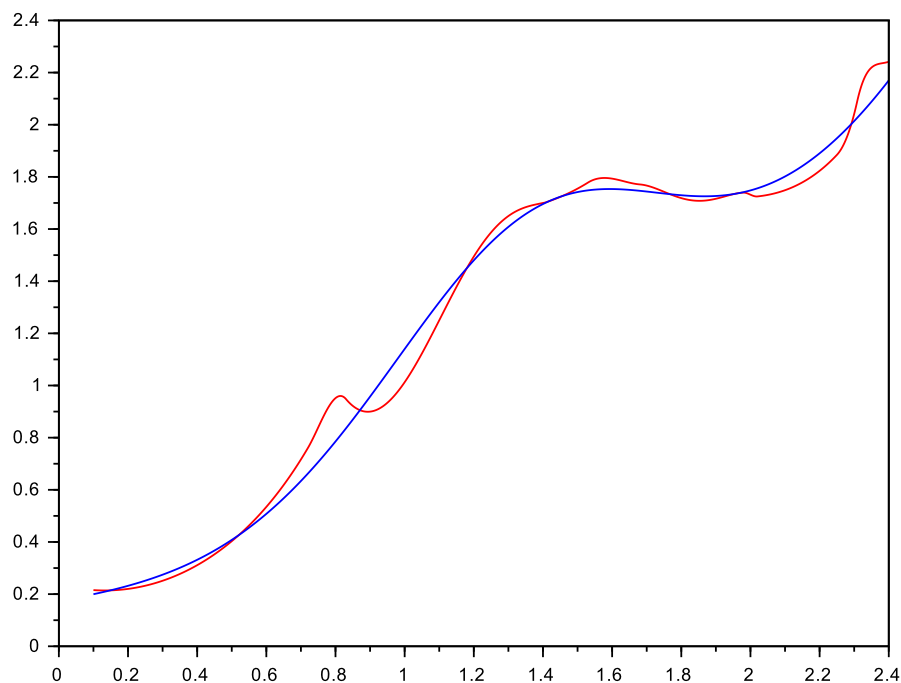


4.10 pav. Realaus spektro pavyzdys. Mėlyna linija yra originalus spektras, žalia – suglodintas spektras

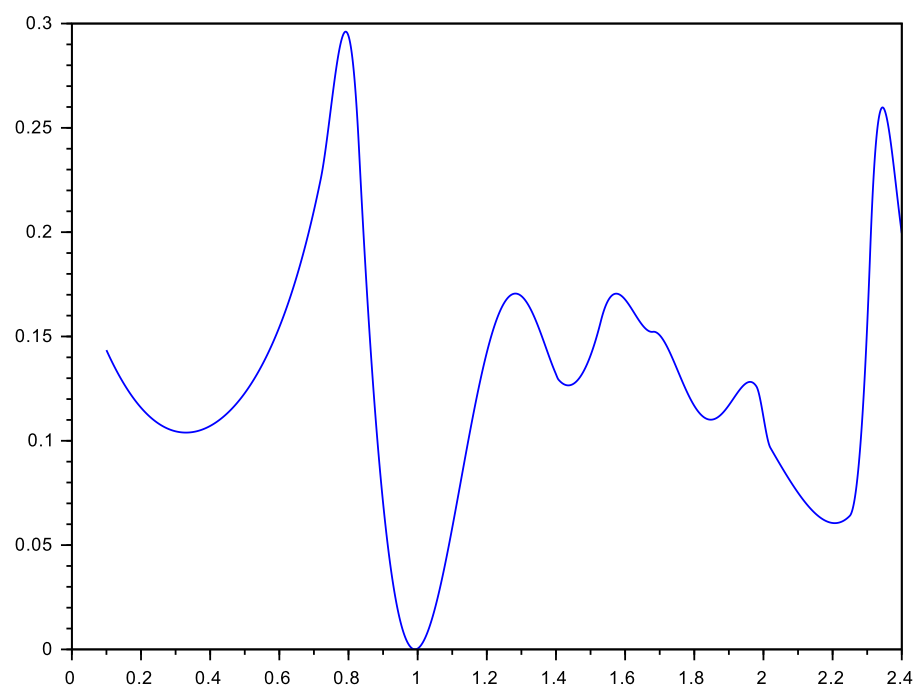
Tokiems spektrams tirti sukūrėme specialią procedūrą, kurią pritaikius galima išskirti spektre esančias paslėptas komponentes. Procedūrą sudaro šie žingsniai:

1. **Polinominė regresija.** Ją naudodami pašaliname triukšmą, kuris gali trukdyti spektro aproksimavimui Gauso funkcijomis. Tačiau jei taikome regresiją visame THz bangų ruože, prarandame informaciją apie spektro pikus, nes regresijos metu mažiname *vidurkinę* paklaidą, o norint tiksliai aproksimuoti spektrą su visais pikais, turime mažinti paklaidą *visuose intervalo taškuose* – gauname daugiakriterinio optimizavimo uždavinį, kuris žymiai sudėtingesnis nei mūsų pagrindinis uždavinys. Praktiniams taikymams reikia paprastesnio sprendimo. Vienas iš tokių yra spektro apibrėžimo intervalo skaidymas į dalis. Kiekvienoje dalyje atliekame polinominę regresiją – gautoji aproksimacija tikslesnė. Ją vadinsime *regresiniu spektru* (žr. 4.11 pav.);

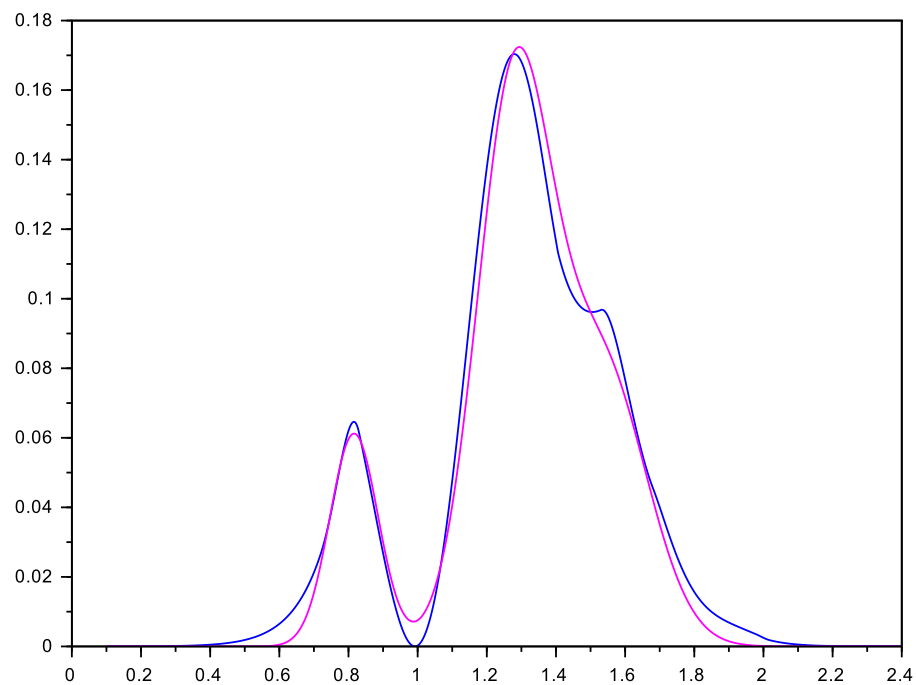
2. **Gauso tendencijos linijos radimas.** Regresiniam spektrui pritaikius (1.5)-(1.6) uždavinį, dar negauname spektro aproksimacijos Gauso funkcijomis, dėl tos pačios priežasties kaip ir polinominės regresijos netikslumo. Vietoje to, gauname tendencijos liniją (4.11 pav.), kurią atimame iš regresinio spektro – gauname *diferencinį spektrą* (4.12 pav.). Jo pavyzdys pavaizduotas paveiksle. Kaip matome, spektras netinkamas aproksimavimui Gauso funkcijomis, nes (1) jo kraštai neartėja prie 0 ir (2) yra neigiamų reikšmių;
3. **Diferencinio spektro filtravimas.** Anksčiau minėtas problemas išsprendžiame, diferencinį spektrą pastumdami taip, kad mažiausia jo reikšmė būtų 0. Tada jam pritaikome juostinį filtrą (angl. *band-pass filter*): tai tiesiog yra Gauso funkcija, kurios centras yra intervalo tarp pirmojo ir paskutiniojo lokaliųjų minimumų vidurio taškas; funkcijos plotis – toks, kad minėto intervalo galai nuo funkcijos centro būtų nutolę per 5σ , čia σ – iš pločio gaunamas standartinis nuokrypis, jei Gauso funkciją laikytume tikimybinio tankiu. Tokia funkcija yra beveik lygi 0 intervalo kraštuose. Ją sudauginame su diferenciniu spektru;
4. **Gauso aproksimavimas.** Sutvarkytam diferenciniam spektrui (4.13 pav.) atliekame Gauso aproksimavimą pagal (1.5)-(1.6) uždavinį. Gautoji aproksimacija pavaizduota 4.13 paveiksle.



4.11 pav. Regresinis spektras (raudona) ir jo tendencijos linija (mėlyna)

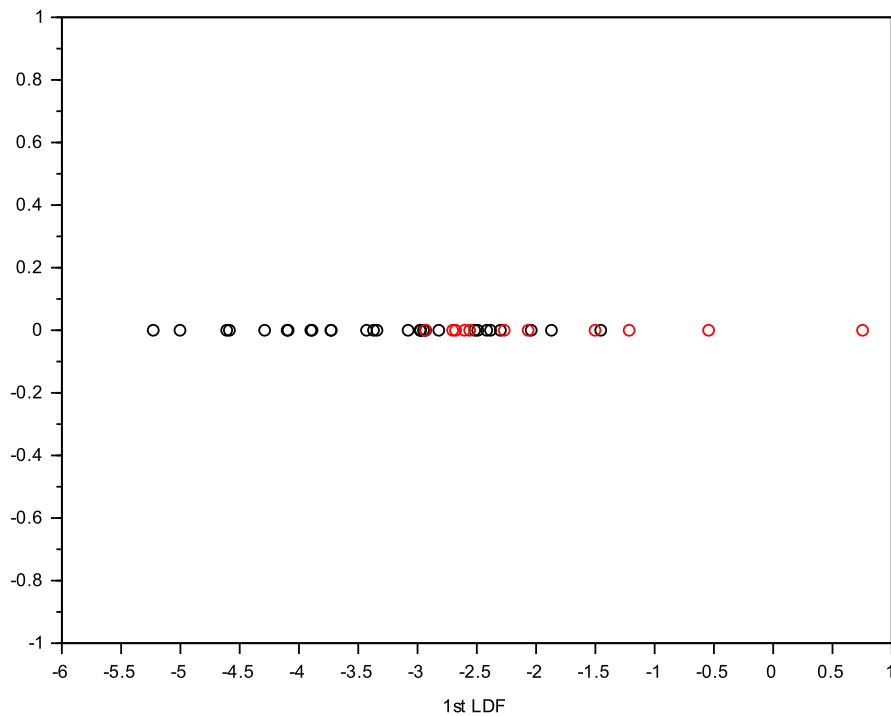


4.12 pav. Diferencinis spektras



4.13 pav. Filtruotas diferencinis spektras (mėlyna) ir jo Gauso aproksimacija (violetinė)

Naudodamiesi minėta procedūra apdorojome 40 spektrų, iš kurių apie 10 turi kofeino. Tada spektrams atlikome tiesinę diskriminantinę analizę: nagrinėjami duomenys – spektrų Gauso parametrai α , β ir γ . Šiuo atveju jų yra po 4 – iš viso 12 parametų. Klasės sužymėtos taip: 0 – nėra kofeino, 1 – yra. Duomenų projekcija pavaizduota 4.14 paveiksle.



4.14 pav. LDA duomenų projekcija. Juodos spalvos apskritimai reprezentuoja spektrus be kofeino, raudonos – spektrus, turinčius kofeino

Išvados

1. Iš skaičiavimų matome, kad centrų pradinis artinys turi būti pakankamai arti tikrosios reikšmės. Vidutinis centrų artinių atstumas nuo sprendinio yra 13,91; idealiu atveju, kai sprendinys pasiekiamas iš atskaitos taško, atstumas yra 48,22. Todėl procedūra komponentių centrų radimui yra būtina.
2. Globalus (SA) metodas duoda geresnius rezultatus nei lokalus (BFGS) metodas originaliu (atstumai: BFGS – 14,32; SA – 15,81) ir Gauso iškraipymo atveju (centrų artinių atstumai iki sprendinio yra: BFGS – 13,49; SA – 14,18). BFGS metodas geresnis sinusinio iškraipymo (atstumai: BFGS – 13,34; SA – 12,33) atveju. Tačiau SA metodas tinkamesnis, nes labiau tikėtina, kad jis suras tiksliausią aproksimaciją.
3. Deformacijos daro vienodą įtaką kiekvieno iš taikytų optimizavimo algoritmų rezultatams. Atstumų vidurkiai deformuotiems spektrams yra: BFGS – 13,42; SA – 13,25). T.y., šiuo atžvilgiu algoritmai yra lygiaverčiai, todėl deformacijos nenulemia metodo pasirinkimo.
4. Realių spektrų analizei reikalinga speciali procedūra, nes jie dažniausiai neturi matomų Gauso komponentių. Randamos paslėptos komponentės turėtų būti įvertintos spektroskopijos specialistų, kad būtų galima įsitikinti pasiūlyto metodo tinkamumu.
5. Pasiūlyta analizės procedūra patikrinta naudojantis LDA metodu. Iš 4.14 paveikslo matome, kad klasių atskyrimas nėra ypatingai geras: matomas žymus persiklojimas ties $x = -2,5$. Galimos persiklojimo priežastys: egzistuoja papildoma kofeino turinčių spektrų klasė arba apdorojimo metodas turi trūkumų.

Literatūra

- [1] P. U. Jepsen, D. G. Cooke, M. Koch, Terahertz spectroscopy and imaging—Modern techniques and applications, *Laser & Photonics Reviews* **5**(1), 124–166 (2011).
- [2] M. Panzner, U. Klotzbach, E. Beyer, F. Rutz, C. Jördens, M. Koch, Non-destructive investigation of paintings with THz-radiation, in *Proceedings of the 9th European Conference of Non-Destructive Testing* (2006).
- [3] N. Karpowicz, H. Zhong, C. Zhang, K.-I. Lin, J.-S. Hwang, J. Xu, X.-C. Zhang, Compact continuous-wave subterahertz system for inspection applications, *Applied Physics Letters* **86**(5), 054105 (2005).
- [4] K. Kawase, Y. Ogawa, Y. Watanabe, H. Inoue, Non-destructive terahertz imaging of illicit drugs using spectral fingerprints, *Opt. Express* **11**(20), 2549–2554 (2003).
- [5] C. Jördens, M. Koch, Detection of foreign bodies in chocolate with pulsed terahertz spectroscopy, *Optical Engineering* **47**, 47 – 47 – 5 (2008).
- [6] M. Yin, S. Tang, M. Tong, Identification of edible oils using terahertz spectroscopy combined with genetic algorithm and partial least squares discriminant analysis, *Analytical Methods* **8**(13), 2794–2798 (2016).
- [7] Y. Jiang, H. Ge, F. Lian, Y. Zhang, S. Xia, Early detection of germinated wheat grains using terahertz image and chemometrics, *Scientific reports* **6** (2016).
- [8] A. Hashimoto, H. Mori, M. Kanou, A. Yamanaka, T. Kameoka, Mid-infrared spectroscopic analysis on brewed coffee characteristics, in *Asian Pacific Confederation of Chemical Engineering congress program and abstracts* (The Society of Chemical Engineers, Japan, 2004), 791–791.
- [9] A. Savitzky, M. J. Golay, Smoothing and differentiation of data by simplified least squares procedures., *Analytical chemistry* **36**(8), 1627–1639 (1964).

- [10] G. Dzemyda, V. Šaltenis, V. Tiešis, *Optimizavimo metodai* (Mokslo Aidai, Vilnius, 2007).
- [11] S. Marsland, *Machine learning: an algorithmic perspective* (CRC Press/Taylor & Francis Group, Boca Raton, FL, 2015).
- [12] S. A. Martišius, V. Kėdaitis, *Statistika* (Vilniaus universiteto leidykla, Vilnius, 2013).
- [13] Scilab Enterprises, Home - Scilab, <http://www.scilab.org/>, [Internete; peržiūrėta 2016-05-27] (2016).
- [14] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (2011).