BEOGRADSKI UNIVERZITET

ELEKTROTEHNIČKI FAKULTET

Vladisav Jelisavčić

STRUKTUIRANO UČENJE NAD VELIKIM PODACIMA ZASNOVANO NA VEROVATNOSNIM GRAFOVSKIM MODELIMA

Doktorska Disertacija

Beograd, 2018.

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTRICAL ENGINEERING

Vladisav Jelisavčić

STRUCTURED LEARNING FROM BIG DATA BASED ON PROBABILISTIC GRAPHICAL MODELS

Doctoral Dissertation

Belgrade, 2018.

KOMISIJA ZA OCENU I ODBRANU DOKTORSKE DISERTACIJE

Mentor:

dr Veljko Milutinović, redovni profesor u penziji

Univerzitet u Beogradu - Elektrotehnički fakultet

Članovi komisije:

dr Boško Nikolić, redovni profesor

Univerzitet u Beogradu - Elektrotehnički fakultet

dr Nenad Mitić, vanredni profesor

Univerzitet u Beogradu - Matematički fakultet

Datum odbrane: _____

Prošireni rezime

Struktuirano učenje je oblast koja se bavi učenjem složenih objekata iz složenih (struktuiranih) podataka. Struktuirani podaci, u ovom kontekstu, predstavljaju podatke koji se sastoje od više međusobno zavisnih delova. Kod učenja nad ovakvim podacima potrebno je iskoristiti informaciju koja se sadrži i u samim delovima, kao i u njihovim međusobnih odnosima. Oblast struktuiranog učenja je važan segment mašinskog učenja, predmet je aktivnog istraživanja uz brojne aktuelne probleme i pokriva učenje iz teksta, signala (audio, slika, video), interakcija u socijalnim mrežama, biomedicinskih (genetskih, bolničkih) podataka, i brojne druge domene.

Struktuirana predikcija je diskriminativna podvrsta struktuiranog učenja i bavi se pronalaženjem veza između ulaza i složenih struktuiranih izlaza. Za razliku od regresije ili klasifikacije, kod kojih je cilj predvideti jedan izlaz, objekti o kojima se zaključuje struktuiranom regresijom, odnosno struktuiranom klasifikacijom, imaju složenu strukturu.

Jedan od izraženijih izazova struktuiranog učenja predstavlja skalabilnost. Sa povećanjem dimenzije problema, značajno se povećava vreme potrebno za obučavanje, usled prevelike složenosti algoritama za učenje. U opštem slučaju složenost algoritama za obučavanje raste eksponencijalno sa povećanjem dimenzije problema. Zbog toga modeli struktuiranog učenja koriste struktuiranu prirodu problema, čime se složenost svodi na polinomijalnu. Jedna od najčešćih pretpostavki kod struktuiranog učenja je tzv. Markovljeva pretpostavka kod koje se, na osnovu Hammersley-Clifford teoreme, struktura problema može predstaviti neusmerenim grafom.

U ovoj disertaciji, predmet istraživanja su Verovatnosni grafovski modeli, i to Markovljeva slučajna polja (Markov Random Field) i Uslovna slučajna polja (Conditional Random Field). Verovatnosni grafovski modeli su klasa algoritama za struktuirano učenje kod kojih je problem predstavljen pomoću slučajnih promenljivih. Markovljeva slučajna polja su verovatnosni grafovski modeli u kojem su promenljive predstavljene kao čvorovi, dok su interakcije između promenljivih predstavljene

kao grane neusmerenog grafa.

Cilj istraživanja u okviru doktorske disertacije je pronalaženje i analiza novih modela za struktuirano učenje i algoritama za njihovo obučavanje. Učenje se najčešće obavlja pronalaženjem maksimalne verodostojnosti (eng. maximum likelihood), pomoću metoda konveksne optimizacije. Efikasnost učenja, kao i zaključivanja na osnovu naučenog modela, u velikoj meri zavisi od osobina samog modela i pretpostavki nad kojima je zasnovan, kao i prilagođenosti modela problemu.

U sklopu doprinosa ove teze, najpre je osmišljena i realizovana generalizacija jedne postojeće GCRF formulacije. Zatim je izvršena eksperimentalna analiza predloženog modela na nekoliko sintetičkih i nekoliko stvarnih skupova podataka iz aktuelnih domena klimatologije i zdravstva, kao i teorijska analiza složenosti.

Zatim je uočen tip problema koji je veoma zastupljen a nije dovoljno obrađen u literaturi, i za koji je moguće uvesti dodatne pretpostavke za koje je zatim pokazano da mogu znatno olakšati problem učenja strukture iz mnogodimenzionalnih podataka. Predložena su i realizovana dva nova GMRF modela zasnovana na uvedenim pretpostavkama i L_1 regularizacione norme, i razvijeni su brzi i skalabilni algoritmi za njihovo učenje.

Pokazana je prednost predloženih modela nad postojećim rešenjima u vidu brzine i skalabilnosti. Predloženi algoritmi su eksperimentalno potvrđeni kroz poređenje sa najbržim postojećim rešenjima na nekoliko mnogodimenzionalnih sintetičkih problema kao i na pravim podacima iz oblasti ekspresije gena, DNK metilacije, i EEG signala. Zatim je odrađena teorijska analiza složenosti algoritama, i zatim eksperimentalno potvrđena sposobnost paralelizacije.

Za kraj su predložene još dve dodatne ekstenzije GCRF modela i postavljena je teorijska osnova za njihovo učenje.

Abstract

Structured learning is an area that deals with the learning of complex objects from complex (structured) data. Structured data, in this context, represent data that consists of several interdependent parts. Structured learning is an important segment of machine learning, and is an active research topic with numerous unsolved problems. Applications that heavily utilize this kind of learning include textual, signal processing (audio, image and video), social, biomedical (genetic, healthcare), and many other.

Structured prediction is a discriminant subtype of structured learning and is concerned with finding connections between inputs and complex structured outputs. Unlike regression or classifications, where the prediction output is a single variable, in structured regression, or structured classification, predictions have a complex structure.

One of the more prominent challenges in structured learning is scalability. With an increase of the dimension of the problem, the time required for learning is greatly increased, due to the excessive time complexity of algorithms required for learning. In general, the complexity of training algorithms is growing exponentially with the increase of problem dimension. Therefore, structured learning models utilize the structured nature of the problem, reducing the complexity down to polynomial. One of the most common assumptions in structured learning is the so-called Markov assumption in which, based on the Hammersley-Clifford theorem, the structure of the problem can be seen as undirected graph.

In this dissertation, main research subject are Markov Random Fields and Conditional Random Fields. The probabilistic graphical models are a class of algorithms for structured learning in which the problem is represented with random variables. Markov's random fields are probabilistic graph models in which variables are represented as nodes, while interactions among the variables are represented as the edges of an undirected graph.

The main research goal within this doctoral dissertation is to discover and analyze new models for structured learning and algorithms for their training. Learning is most often done by finding maximum likelihood from data, using convex optimization methods. The learning efficiency, as well as the predictive power of the learned model, largely depends on the characteristics of the model itself and the assumptions upon which it is based, as well as the applicability of the model to specific problem of interest.

As part of the contribution of this dissertation, first a generalization of an existing GCRF formulation was developed and realized. Then, the experimental analysis of the proposed model was carried out on several synthetic and several real data sets from the high-impact domains of climatology and health, as well as the theoretical complexity analysis.

Second, an additional assumption applicable to one common problem type is discussed, and then shown to greatly facilitate the problem of learning structure from high-dimensional data. Two new GMRF models were proposed and implemented based on the introduced assumption and L_1 regularization norms, and rapid and scalable algorithms for their learning were proposed and implemented.

The advantage of the proposed models over the state-of-the art is shown empirically while comparing speed and scalability. The proposed algorithms have been experimentally verified by comparing with the fastest existing solutions on several high-dimensional synthetic problems as well as on real data in the areas of gene expression, DNA methylation, and EEG signals. The theoretical complexity analysis of the algorithms was made, and then the experimentally confirmed together with parallelization capability.

Finally, two additional extensions of the GCRF model were proposed together with theoretical foundations for their learning.

Lista slika

Lista tabela

Sadržaj

1 Uvod

Čest problem u širokom spektru naučnih grana je modeliranje strukture zavisnosti između velikog broja varijabli u nekom složenom sistemu od interesa. Grafički modeli pružaju pogodan način predstavljanja i analize odnosa u takvim zadacima. Poželjan pristup je modeliranje zajedničke distribucije verovatnoće učenjem primera verovatnosnog grafičkog modela [38]. Neke od glavnih prednosti kod verovatnosnih grafičkih modela su garancije u vidu konvergencije algoritama učenja, konveksnosti kriterijumske funkcije, kao i modularnost i mogućnost uvođenja dodatnih pretpostavki u model, kao i jaka pozadina u matematičkim alatima, razvijenim u oblasti statistike i teorije verovatnoće. Uvođenjem dodatnih ograničenja ekspresivnosti probabilističkih grafičkih modela moguće je doći do modela sa vrlo atraktivnim osobinama.

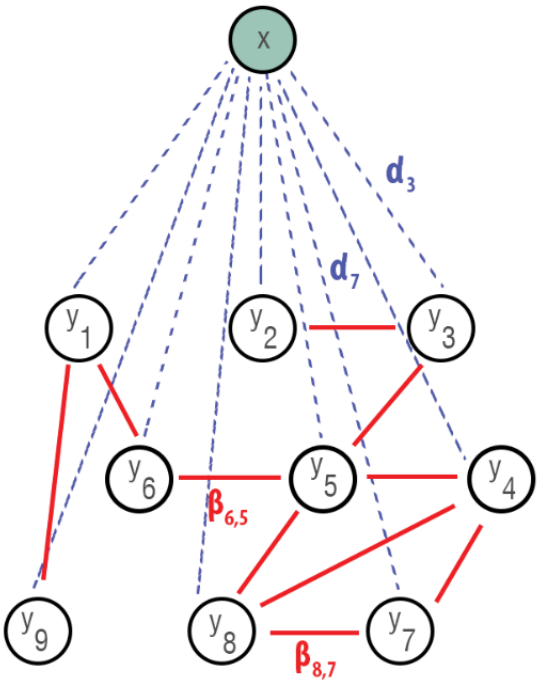
Jedan od izraženijih izazova struktuiranog učenja predstavlja skalabilnost. Sa povećanjem dimenzije problema, značajno se povećava vreme potrebno za obučavanje, usled prevelike složenosti algoritama za učenje. U opštem slučaju složenost algoritama za obučavanje raste eksponencijalno sa povećanjem dimenzije problema. Zbog toga modeli struktuiranog učenja koriste struktuiranu prirodu problema, čime se složenost svodi na polinomijalnu. Jedna od najčešćih pretpostavki kod struktuiranog učenja je tzv. Markovljeva pretpostavka kod koje se, na osnovu Hammersley-Clifford teoreme, struktura problema može predstaviti neusmerenim grafom.

Ova pretpostavka predstavlja osnov za čitavu klasu algoritama za struktuirano učenje poznatu kao Markovljeva slučajna polja, koja će biti predmet istraživanja u ovoj tezi.

U Markovljevim slučajnim poljima, odsustvo grane u grafu odgovara uslovnoj nezavisnosti između dve promenljive, dok prisustvo grane u grafu modeluje interakciju između susednih promenljivih. Uslovna slučajna polja su diskriminativna verzija Markovljevih slučajnih polja, koja služe za struktuiranu predikciju.

$$p(Y) = \frac{1}{Z} \prod f_i(y) \prod g_{ij}(y_i, y_j) \tag{1}$$

Na osnovu Hammersley-Clifford teoreme svako Markovljevo polje se može predstaviti kao proizvod potencijala, gde svaki potencijal odgovara kliki u grafu koji određuje uslovne nezavisnosti između promenljivih (jednačina 1). Slučajna promenljiva je uslovno nezavisna od ostatka grafa ukoliko su dati svi njeni susedi (koji čine takozvano "Markovljevo ćebe" za tu promenljivu, eng. Markov blanket). Na taj način struktura problema u vidu grafa međusobnih relacija i zavisnosti se preslikava u verovatnosnu distribuciju, pomoću koje je moguće vršiti učenje parametara i zaključivanje neopaženih promenljivih.



Slika 1: Grafička prezentacija primera jednog kondicionalnog slučajnog polja. Na grafičkom primeru se vidi 9 varijabli zavisnih od jedne zajedničke promenljive. Isprekidanom plavom linijom su obeležene kondicionalne zavisnosti, a punom crvenom linijom Markovljevo polje.

Na slici 1 prikazana je grafička reprezentacija jednog primera kondicionalnog slučajnog polja. Na slici se može videti dve vrste potencijala: asocijativni i interak-

cioni. Asocijativni potencijali odgovaraju pojedinačnim čvorovima u grafu (slučajnim promenljivama), dok interakcioni potencijali modeluju veze između slučajnih promenljivih (klike drugog reda u grafu). Kod Uslovnih slučajnih polja, obe vrste potencijala mogu takođe zavisiti od dodatnih promenljivih, čija se interakcija ne modeluje, već učestvuju u modelu samo kroz zavisnost izlaznih promenljivih od njih. Stoga, celokupan izraz se svodi na uslovnu umesto zajedničku verovatnosnu distribuciju (odakle i potiče naziv Uslovna slučajna polja). Ovakav pristup značajno smanjuje prostor pretrage, jer nije potrebno naučiti odnose izmedju opaženih varijabli, što je veoma primenjivo u problemima klasifikacije i regresije, gde imamo jasno diferenciranu ulogu ulaznih i izlaznih promenljivih.

Osnovni problemi kod struktuiranog učenja su: zaključivanje, dekodovanje, i učenje. Zaključivanje predstavlja proces određivanja najverovatnijih vrednosti promenljivih za date parametre modela. U opštem slučaju, egzaktno zaključivanje nije izvodljivo i neophodno je usvajanje dodatnih pretpostavki ili nalaženje aproksimativnog rešenja. Dekodovanje predstavlja pronalaženje verovatnoće za date vrednosti promenljivih. Učenje predstavlja pronalaženje optimalnih vrednosti parametara za dati model pomoću datih podataka uz pretpostavku da podaci odgovaraju modelu.

Vremenska kompleksnost algoritama za učenje zavisi i od tipa problema. Ukoliko je problem kontinualne prirode, najčešće su neophodne dodatne pretpostavke kako bi problem bio traktabilan. Najčešće pretpostavke koje se koriste se odnose na funkciju raspodele promenljivih. Na primer, kod Markovljevih slučajnih polja, korišćenje Gausovske pretpostavke omogućava analitičke izraze za zaključivanje, čime se u opštem slučaju dobija algoritam za učenje kubne složenosti. U slučaju mnogo-dimenzionalnih problema koji postaju sve aktuelniji, algoritmi sa kubnom složenošću postaju nedopustivo spori. Na primer, učenje retkog Markovljevog slučajnog polja sa oko deset hiljada promenljivih sa trenutno aktuelnim algoritmima na modernom hardveru traje oko sat vremena. Poređenja radi, pojedini problemi u

analizi socijalnih mreža, klimatskih promena, aktivnosti neurona u mozgu, genetskih podataka, mogu imati nekoliko redova veličine više promenljivih, kao i mnogo gušću strukturu.

Dodatan izazov pri struktuiranom učenju predstavlja veličina skupa za obučavanje. Sa povećanjem dimenzije problema, drastično se povećava broj parametara koje je neophodno optimizovati kako bi se algoritam uspešno obučio, čime se ujedno povećava i prostor pretrage parametara. Sa povećanjem prostora pretrage, neophodan je veći broj primera za obučavanje koji često nisu dostupni (npr. za analizu aktivnosti neurona u mozgu sa FMRI snimka pomoću 1000 voksela, bilo bi neophodno estimirati milion parametara, koristeći milion snimaka). Usled ovoga, modeli često unose dodatne pretpostavke, kao što su retkost (tzv. sparsity) strukture, ili određeni obrazac kao što je to slučaj u prostorno i temporalno uređenim problemima. Pronalaženje adekvatnih pretpostavki je predmet aktivnog istraživanja, čija efikasnost može zavisiti i od problema na koji se primenjuje.

Kako je zaključivanje u ovim modelima netraktabilno u opštem slučaju, uvodi se Gausova pretpostavka kako bi se omogućio efikasno zaključivanje. Ovakva pretpostavka dovodi do modela Gausovskog Markovljevog slučajnog polja (eng. Gaussian Markov Random Field, skraćeno GMRF). Mnogi multivarijatni procesi u prirodi se mogu efikasno aproksimirati pomoću multivarijatne normalne (MVN) distribucije. Pretpostavka MVN-a omogućava pouzdano GMRF učenje iz podataka, korišćenjem metode maksimalne verodostojnosti parametara. Učenje GMRF-a je ekvivalentno estimaciji retke inverzne kovarijansne matrice (nazvane još i matrica preciznosti).

U ovoj tezi će između ostalog biti predložene dve nove metode za skalabilno učenje Gausovskog slučajnog polja iz podataka.

2 Korišćenje strukture za poboljšanje predikcije

Struktuirano predviđanje je tehnika istovremenog predviđanja skupa povezanih varijabli na osnovu skupa ulaznih (tzv. obzervabilnih) varijabli [1]. Mnogi izazovni problemi u strukturiranom predviđanju nastaju usled eksponencijalne kompleksnosti prostora pretrage, i zahtevaju osmišljavanje složenih modela kako bi se opisao problem od interesa, što se odražava na izvodljivost i traktibilnost algoritama učenja i zaključivanja. Rezultujuće varijable koje se koriste za opisivanje predmeta od interesa često su međusobno povezane, a odnosi između varijabli formiraju strukturu. Ta struktura, iako potencijalno složena, obično se može iskoristiti u samom modelu kako bi problem postao traktabilan.

Struktuirana regresija, za razliku od struktuirane klasifikacije, zahteva da izlazne varijable uzimaju vrednosti iz kontinualnog domena. Jedan od najvećih izazova struktuirane regresije je efikasnost optimizacije, jer se optimizacija koristi u fazi učenja, a često čak i u fazi zaključivanja. Izbor modela kojim su izražene zajedničke ili uslovne raspodele izlaznih promenljivih zavisi od samog zadataka, dostupnosti primera za obuku i drugih ograničenja [49]. Diskriminativni modeli često imaju prednost u odnosu na generativne, usled relaksiranja pretpostavke o nezavisnosti [75]. Zbog toga se koriste Uslovna Slučajna Polja [40] (eng. Conditional Random Fields) koja pored izlaznih veličina koje se modeluju slučajnim promenljivima, uvode i tzv. observabilne promenljive čija raspodela nije uračunata modelom. Uslovna slučajna polja se često primenjuju u različitim domenima, uključujući oblasti kao što su računarska vizija (eng. Computer Vision) [53], obrada prirodnog jezika (eng. Natural Language Processing) [39] i bioinformatika [58].

2.1 Uvod u diskriminativne modele

Grafički modeli, kada se primenjuju na probleme gde je potrebno izvršiti predikciju više izlaza (eng. multi-target prediction), obično koriste uslove interakcije (u vidu

funkcija potencijala) da nametnu strukturu među izlaznim varijablama. Često se takva struktura zasniva na pretpostavci da "slične" izlazne promenljive moraju uzimati i slične vrednosti, gde je mera sličnosti između dve promenljive zadata nekom metrikom. Takva vrsta interakcije se ponaša kao glačajući filter i navodi izlazne vrednosti da budu međusobno bliže. U istraživanju predstavljenom u sledećem poglavlju relaksiramo tu pretpostavku i predlažemo funkciju koja se temelji na rastojanju i može se prilagoditi kako bi se osiguralo da varijable imaju manju ili veću razliku u vrednostima. Tekst i rezultati prikazani u ovom poglavlju su zasnovani na radu objavljenom u [70]. Koristili smo postojeći model Gausovskog Slučajnog Polja (eng. Gaussian Conditional Random Field, skraćeno GCRF), koji smo unapredili i predložili proširenje potencijala interakcije dodatnim članom kojim se uračunava i udaljenost pored sličnosti. Prošireni model je zatim upoređen sa polaznim u nekoliko različitih problema struktuirane regresije. Povećanje prediktivne preciznosti zabeleženo je i na sintetički generisanim primerima kao i kroz aplikacije na realnim podacima, uključujući i izazovne probleme iz domena klimatologije, (climate) i zdravstva (healthcare).

2.2 Pregled postojećih rešenja

Kontinualna uslovna slučajna polja su tip *Uslovnih slučajnih polja* koja modeluju probleme kontinualne prirode. Iako predstavljaju moćan alat, primena kontinualnih uslovnih slučajnih polja na problem struktuirane regresije je donekle otežana visokom računskom složenošću koja je posledica izračunavanja normalizacione funkcije (eng. Partition function). Takođe, zaključivanje kod ovih modela često zahteva komputacijski skupe metode odabiranja. Da bi ublažili neke od tih problema i učinili metode primenljivijim, jedno od rešenja je korišćenje pretpostavke Gausovske distribucije. Nedavno je predloženo nekoliko formulacija CRF modela nazvanih Gaussian Conditional Random Fields (GCRF) [77, 55, 66, 84]. Pristupi poput [84, 66] koriste tehnike regularizacije, uglavnom zasnovane na L_1 normi, kako bi naučili

graf u vidu retke precision matrice. Drugi pristupi, poput [77], imaju specijalizovanije formulacije usko povezane sa problemom koji modeluju. U radu [55] predložen je model za predviđanje optičke vidljivosti (eng. Aerosol Optical Depth) koji se oslanja na poznatu strukturu koja poboljšava učenje prediktivnog modela. U ovom poglavlju ćemo se usredsrediti na ovu formulaciju GCRF-a.

Gausovska formulacija asocijativnih i interakcionih potencijala GCRF obezbeđuje da celokupan model ima oblik multivarijatne Gausove distribucije. Ova osobina GCRF-a omogućava efikasno zaključivanje, kao posledicu toga što je moguće naći algebarski izraz za normalizacionu funkciju. Štaviše, problem GCRF učenja (pronalaženja optimalnih parametara) je konveksan, što omogućava korišćenje pouzdanih algoritama konveksne optimizacije i garantuje visokokvalitetno rešenje. Sve ove karakteristike čine GCRF obećavajućim alatom za širok spektar aplikacija uključujući energetiku [16], internet oglašavanje [85, 67], klimatologiju [55] i zdravstvo [54, 72]. Postoji nekoliko unapređenja i modifikacija ovog modela, kao na primer: za primenu na nekompletnim podacima (sa nedostajućim vrednostima) [68]; sa simultanim obučavanjem neuralnih mreža [56]; sa propagacijom greške predikcije [24, 25]; za ansambl struktuiranih modela [52]; za usmerene grafove [81]; i za višeslojne mreže [23].

Jedan od izazova struktuirane regresije je pronalaženje adekvatne strukture koja odgovara problemu koji se modeluje. U zavisnosti od postavke, postoji nekoliko pristupa ovom problemu pomoću Gausovskih uslovnih slučajnih polja.

Jedan pristup se zasniva na učenju cele strukture u vidu kovarijansne matrice, koristeći regularizaciju koja indukuje njenu inverznu matricu (tj. matricu parcijalnih kovarijacija, eng. Precision matrix) da bude retka [84], [66]. Ovakakv pristup ne koristi deljenje parametara već se efektivno svodi na učenje cele strukture (u vidu retke matrice nezavisnih parametara).

Još jedan pristup se zasniva na korišćenju dodatne informacije u vidu predefinisanog težinskog grafa [55]. U praksi, ovakva vrsta dodatne informacije dolazi iz

domenskog znanja ili toplogije samog problema. Ovakav pristup znatno umanjuje prostor pretrage optimizacionog problema, jer se smanjuje broj parametara koje je potrebno naučiti. Sa smanjenjem broja parametara, smanjuje se i rizik od preobučavanja, jer se znatno smanjuje broj instanci neophodnih za obučavanje modela. Ovakav benefit ne dolazi bez cene, jer sa smanjenjem prostora pretrage smanjuje se i prediktivna moć modela. Predefinisani težinski graf može se posmatrati kao ograničenje u prostoru optimizacione pretrage, i ukoliko taj predefinisani težinski graf ne opisuje dobro problem ni naučeni model neće dobro generalizovati.

Treći pristup je hibrid između učenja cele matrice parametara i korišćenja predefinisanog težinskog grafa. U ovakvom pristupu, grane grafa su predefinisane ali se težinski koeficijenti svake grane uče kao parametri modela. Na ovaj način je moguće u izvesnoj meri napraviti kompromis između prethodna dva pristupa kao što će i biti pokazano u nastavku rada.

U nastavku će biti opisan model za struktuiranu regresiju predstavljen u [55]. Ovaj model je definisan kao oteženjeni proizvod asocijativnih i interakcionih potencijala:

$$P(Y|X) \quad = \quad \frac{1}{Z}\exp(-\sum_{k}\alpha^{k}\sum_{i}(y_i - R_i^k(X))^2 \;-\; \sum_{l}\beta^{l}\sum_{(i,j)}S_{ij}^{l}(y_i - y_j)^2) \quad (2)$$

Asocijativni potencijal u ovom modelu definiše se kao kvadratna razlika između promenljive čvora y_i i neke nezavisno estimirane vrednosti R_i dobijene pomoću nekog klasičnog (nestruktuiranog) algoritma mašinskog učenja. Model omogućuje pridruživanje jednog ili više asocijativnih potencijala svakom čvoru, čime se omogućuje korišćenje više različitih prediktora.

Potencijal interakcije definisan je kao ponderisana kvadratna razlika između promenljivih dodeljenim susednim čvorovima u grafu, pri čemu je težina S_{ij} proporcionalna odgovarajućoj grani grafa. Kao i u slučaju aktivacionih potencijala,

model dozvoljava pridruživanje jednog ili više grafa, od kojih svaki doprinosi modelu svojim interakcionim potencijalom, što je realizovano sumama po k i l u (2). Faktori ponderisanja α i β koji odgovaraju asocijativnim i interakcionim potencijalima respektivno, su parametri modela i mogu se naučiti optimizovanjem makismalne verodostojnosti.

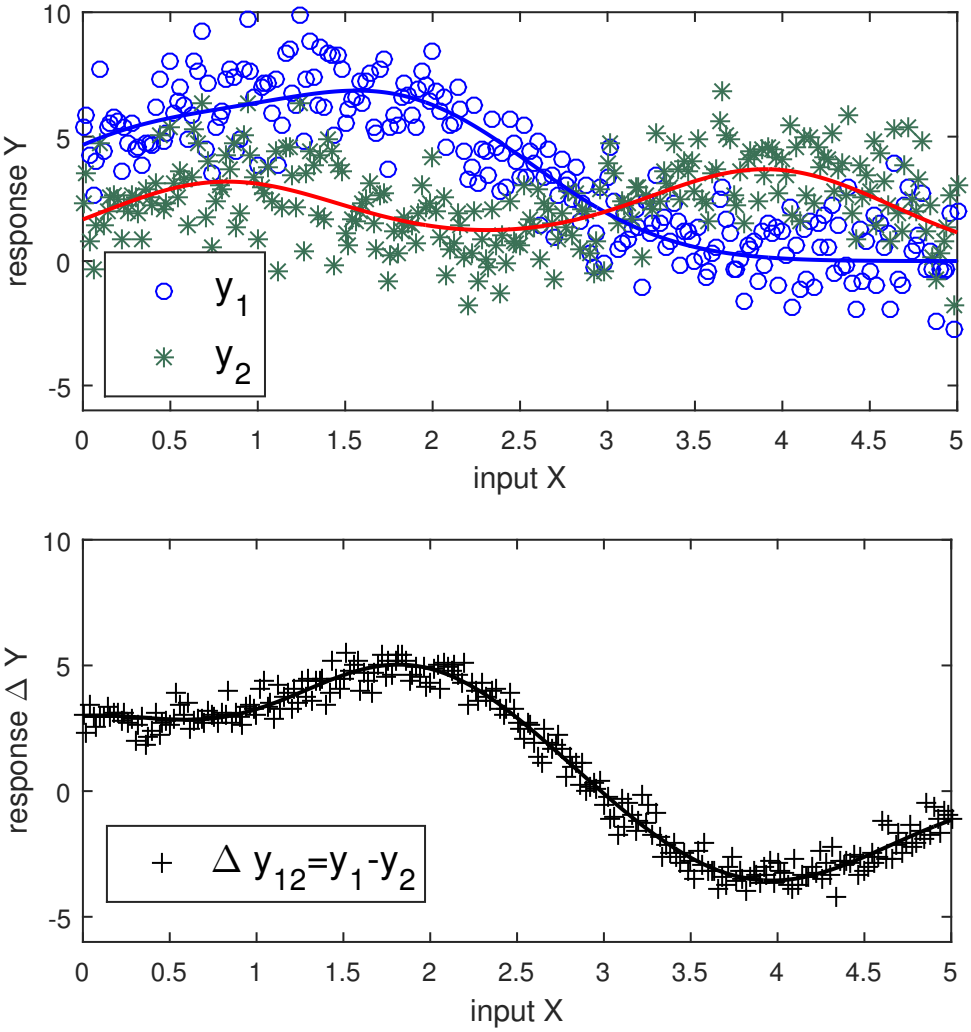
Ovakav model se oslanja na unapred definisanu strukturu datu u obliku neusmerenog težinskog grafika S. Za neke probleme ovaj graf se pojavljuje prirodno, npr. u problemima u kojima se mogu lako identifikovati korelacioni obrasci, kao u prostornim ili vremenski zavisnim problemima. Kada je graf predefinisan (grane grafa su unapred zadate i nisu deo optimizacionog problema), prostor pretrage parametara je znatno smanjen, što omogućava efikasno učenje. Međutim, ako izabrani graf nije odgovarajući za problem, prostor parametara je ograničen na podprostor prvobitnog problema koji možda ne sadrži željeno rešenje, i može doći do tzv. underfitting-a.

Funkcije potencijala interakcije definisane u (2) kažnjavaju razliku između predviđanja na susednim čvorovima, i na taj način deluju kao glačajući filter. Intenzitet ovog efekta je srazmeran težini grana u grafu.

Ovako postavljen model omogućuje efikasno zaključivanje jer izraz (2) zapravo može biti predstavljen kao multivarijatna Gausova uslovna raspodela, usled čega normalizaciona funkcija ima analitičko rešenje, a modus ovakvog modela, tj rezultat predikcije se jednostavno dobija kao očekivana vrednost multivarijatne normalne raspodele. Funkcija verovatnoće je konveksna funkcija parametara α i β, što omogućava korišćenje pouzdanih i efikasnih metoda optimizacije.

2.3 DGCRF

Pretpostavka data u modelu (2) je da što su dve varijable povezanije, njihove vrednosti bi trebale biti sličnije. Interakcioni potencijal je formiran tako da penalizuje razliku između povezanih izlaznih varijabli, čineći ih sličnijim i stoga delujući kao



Slika 2: Dve slučajne varijable povezane visoko korelisanim aditivnim šumom. Na levoj slici je prikazano 250 opservacija zašumljenih signala y_1 i y_2, kao i polazni deterministički signali. Na desnoj slici predstavljena je njihova razlika $\Delta y_{12} = y_1 - y_2$. Usled visoke korelacije, oduzimanjem merenja moguće je filtrirati značajan deo šuma. Otkrivanje pravog signala razlike Δy_{12} je očigledno mnogo lakši zadatak u poređenju sa pronalaženjem originalnog signala iz šumovitih opservacija pojedinačnih slučajnih promenljivih y_1 i y_2. U ovoj situaciji, procenjena razlika je pouzdanija i može se koristiti za poboljšanje predviđanja ciljnih varijabli. Takav korelisani šum je čest u biološkim podacima, na primer u studijama ekspresije gena [19].

glačajući filter. Ova pretpostavka je prisutna i u drugim grafovskim modelima koji penalizuju razliku u povezanim varijablama, kao što je to slučaj pronalaženje obeležja pomoću grafovske LASSO [21]. Međutim, ponekad je ovo nepoželjna osobina, a jedno predloženo rešenje za taj problem je uvođenje znaka prilikom oduzi-

manja vrednosti dva izlaza, nazvano Graphical Fused LASSO penal [37]. Međutim, ovakav trik nije uvek primenjiv, jer uvodi dodatne probleme kada znak nije dobro odabran, kao i zbog diskontinuiteta kriterijumske funkcije koji zahteva složenije metode optimizacije [86].

U nastavku, biće predložena jedna generalizacija najčešće korišćenog tipa interakcionog potencijala, odnosno onog koji se bazira na razlici parova susednih promenljivih. Ova generalizacija se zasniva na uvođenju dodatnog člana kojim modelujemo distancu izmedju dva čvora, koji proširuje obim njegove primene. U ovakvoj postavci, model se više ne ponaša isključivo tako da uglačava susedne promenljive, već može učiniti da susedne varijable budu više različite (udaljene). Sa tako definisanim potencijalom interakcije, model teži da primora razliku između susednih varijabli tako da budu jednake navedenom članu koji predstavlja udaljenost izmedju dva čvora.

U radu koji je jedan od naučnih doprinosa ove teze, istražujemo ponašanje takvih funkcija interakcije na modelu Gausovskih uslovnih slučajnih polja (GCRF). U predloženom pristupu, karakteristike zaključivanja i učenja originalnog modela ostaju očuvane, a opseg primene samog modela je proširen, što rezultira povećanom tačnošću struktuirane regresije. Originalni model (2) je zapravo poseban slučaj predloženog modela kada su zadata rastojanja između susednih varijabli postavljena na nulu. Ovakav model nazvali smo Daljinska Gausova uslovna slučajna polja (eng. Distance-based Gaussian Conditional Random Field).

2.3.1 Metod

Polazni model opisan u (2) može se posmatrati kao proizvod ponderisanih potencijala. Sada ćemo predstaviti drugačiji pogled fokusirajući se na Gausovsku prirodu modela, i važne osobine Gausove distribucije. Ovo stanovište će olakšati ukazivanje na neke važne karakteristike razvijenog modela.

U opštem slučaju, MRF (i CRF) sa unarnim i parnim potencijalima na skupu od

n varijabli $Y = [y_1 ... y_n]^T$, mogu se faktorisati kao:

$$p(Y) = \frac{1}{Z} \prod f_i(y_i) \prod g_{ij}(y_i, y_j) \tag{3}$$

Za razliku od MRF modela kod kojih slučajne promenljive Y i X zajedno kovariraju, kod CRF modeluje se samo uslovna verovatnoća Y u zavisnosti od X, promenljive X su uvek opservirane. Jedna od prednosti Uslovnih slučajnih polja je to što se izostavljanjem strukture između samih promenljivih X može značajno smanjiti prostor pretrage parametara. Ovakva postavka pogoduje problemu regresije, tj pronalaženje vrednosti izlaznih promeljivih na osnovu skupa opserviranih ulaznih varijabli.

Ako usvojimo Gausove funkcije za aktivacione potencijale (f_i) i potencijale interakcije (g_{ij}) dobijamo Gausovsko uslovno slučajno polje (GCRF). Ako sada preformulišemo aktivacione potencijale iz modela (2), oni će imati oblik Gausove funkcije sa parametrima (R, σ):

$$f_i(y_i) = \exp(-\frac{(y_i - R_i(X))^2}{2\sigma_i^2}) \tag{4}$$

gde je $R_i(X)$ srednja vrednost estimirana na osnovu podataka pomoću neke od postojećih metoda regresije, i koja zavisi od opserviranih vrednosti promenljivih X. Član $\sigma_i^2 = \frac{1}{2}\alpha^{-1}$ predstavlja varijansu kao funkciju deljenog parametra α koji je potrebno naučiti.

Interakcioni potencijali takođe imaju oblik Gausove funkcije sa parametrima $(0, \sigma_{ij})$:

$$g_{ij}(y_i, y_j) = g_{ij}(\Delta y_{ij}) = \exp(-\frac{\Delta y_{ij}^2}{2\sigma_{ij}^2}) \tag{5}$$

gde je: $\Delta y_{ij} = y_i - y_j$, očikvana vrednost je nula i varijansa je: $\sigma_{ij}^2 = (2\beta S_{ij})^{-1}$, β je deljeni parametar koji je potrebno naučiti.

Interpretacija (4) and (5) je sledeća:

1. Gausovski asocijativni potencijali (4) primoravaju vrednost svakog čvora y_i

da bude blizu odgovarajuće vrednosti $R_i(X)$, sa parametrom preciznosti $\sqrt{\alpha}$ koji je potrebno naučiti iz podataka.

2. Gausovski interakcioni potencijali (5) primoravaju vrednost Δy_{ij} da bude blizu nule, sa preciznošću proporcionalnom $\sqrt{\beta S_{ij}}$. Prema tome, S_{ij} se može tumačiti kao mera sličnosti (veće vrednosti S_{ij} dovode do manje razlike Δy_{ij}). Faktor proporcionalnosti $\sqrt{\beta}$ za interakcione potencijale se takođe uči kao parametar modela.

Ovaj oblik interakcije u kojem je očekivana srednja vrednost nula teži da učini vrednosti povezanih variabli sličnijima. Ta karakteristika bi mogla biti poželjna pod pretpostavkom da su povezane varijable uvek slične, ali to ne mora biti slučaj. Zato proširujemo model potencijala interakcije (5) uvođenjem dodatnog termina $D_{ij}(X)$, koji predstavlja meru udaljenosti koja zavisi isključivo od X:

$$g_{ij}(y_i, y_j) = g_{ij}(\Delta y_{ij}) = \exp\left(-\frac{(\Delta y_{ij} - D_{ij}(X))^2}{2\beta_{ij}^2}\right) \tag{6}$$

Novi GCRF model sada možemo posmatrati kao:

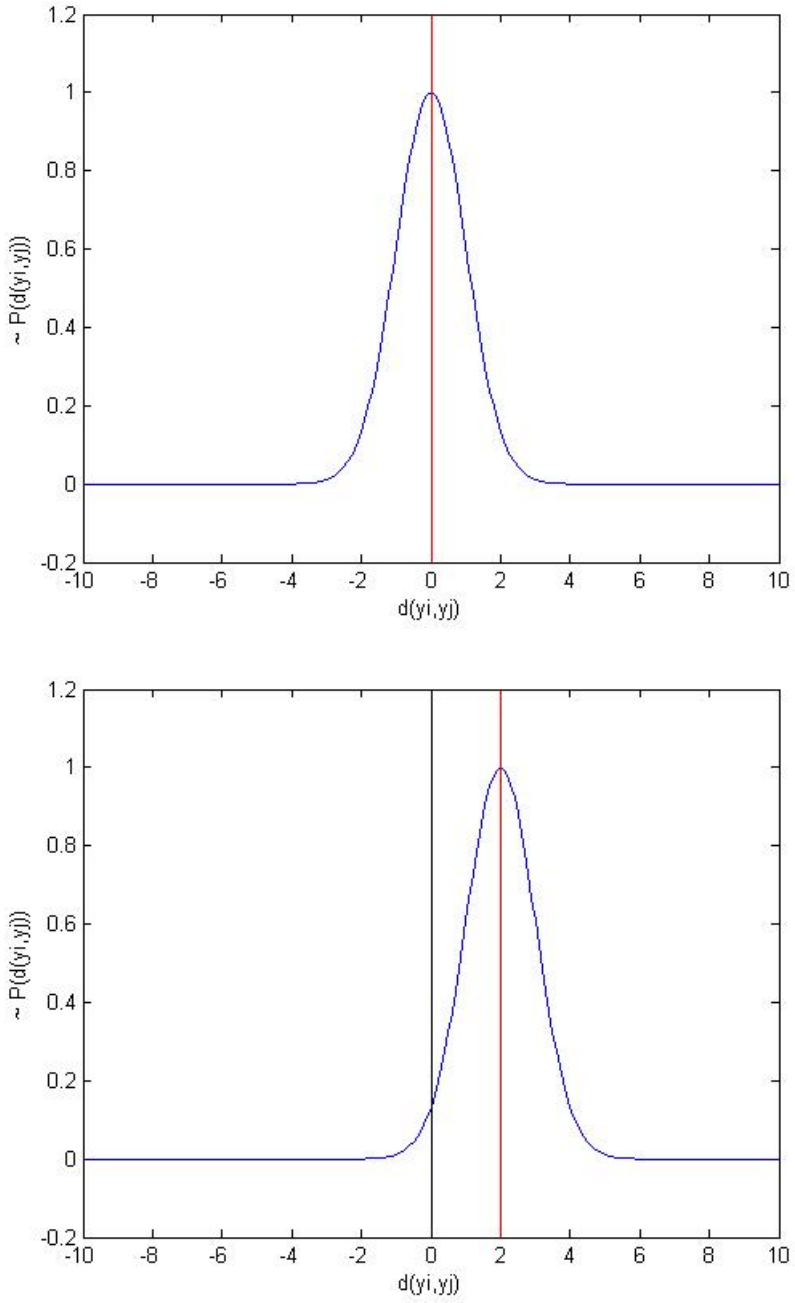
$$P(Y|X) = \frac{1}{Z}\exp((-\sum_k \sum_i \alpha_i^k (y_i - R_i^k(X))^2 - \sum_l \sum_{(i,j)} \beta_{ij}^l (y_i - D_{ij}^l(X) - y_j)^2)) \tag{7}$$

Učenje i zaključivanje Predloženi model nasledjuje sva poželjna svojstva osnovnog modela koja omogućavaju efikasno učenje i zaključivanje.

Prvo ćemo analizirati zaključivanje. Modelovana uslovna raspodela je oblika:

$$P(Y|X) = \frac{1}{Z}\exp\left(-E\right) \tag{8}$$

Ako izjednačimo exponent modela u (8) kao sumu oteženjenih kvadratnih potencijala (9) i eksponent multivarijatne Normalne raspodele (10):



Slika 3: Interakcioni potencijali. Na gornjoj slici je predstavljen interakcioni potencijal zero mean GCRF, a na donjoj slici je predstavljen interakcioni potencijal DGCRF koji ima parametar udaljenosti srednje vrednosti od nule.

$$E \quad = \quad \sum_k \sum_i \alpha_i^k (y_i - R_i^k(X))^2 \quad + \quad \sum_l \sum_{(i,j)} \beta_{ij}^l (y_i - D_{ij}^l(X) - y_j)^2) \quad (9)$$

$$E = \frac{1}{2}(Y - \mu)^T \Sigma^{-1} (Y - \mu) \quad (10)$$

Dobijena matrica preciznosti $Q = \Sigma^{-1}$ je sastavljena od $Q1$ (asocijativnog potencijala) and $Q2$ (interakcionog potencijala), što je rezultat već pokazan u [55]:

$$Q1_{i,j} = \begin{cases} \sum_k \alpha_i^k, & i = j \\ 0, & i \neq j \end{cases} \quad (11)$$

$$Q2_{i,j} = \begin{cases} \sum_l \sum_{(i,j)} \beta_{ik}^l, & i = j \\ -\sum_l \beta_{ij}^l, & i \neq j \end{cases} \quad (12)$$

Takođe dobijamo:

$$\mu = \Sigma b \quad (13)$$

dok je b moguće predstaviti sledećim izrazom:

$$b_i = 2 \sum_k \alpha_i^k R_i^k(X) + 2 \sum_l \beta_{ij}^l D_{ij}^l(X) \quad (14)$$

Pošto je razlika anti-simetrična, $D_{ij}(X) = -D_{ji}(X)$ može se koristiti da se trenira samo jedan smer, dok je drugi smer moguće estimirati kao negaciju prvog (14).

Usled toga što je $\Sigma^{-1} = 2(Q1 + Q2)$, koristeći jednačine (11) do (14) moguće je odrediti očekivane vrednosti za svaku od izlaznih promenljivih.

Iz jednačina je uočljivo da se predloženi model razlikuje u odnosu na originalni samo u izrazu za b, u dodatnom članu koji modeluje distancu (14).

Slobodni parametri α i β koji određuju odnos uticaja strukture i nestruktuiranih prediktora je moguće naučiti na osnovu podataka, i to metodom maximalne verodostojnosti.

U nastavku su predstavljeni izrazi za ažuriranje parametara modela prilikom učenja i njihovo izvođenje. Polazna tačka u izvođenju je izraz za diferencijal log-likelihood modela $d \log P$, predstavljen u originalnom modelu [55]:

$$d \log P = -\frac{1}{2}(Y-\mu)^T d\Sigma^{-1}(Y-\mu) + (db^T - \mu^T d\Sigma^{-1})(Y-\mu) + \frac{1}{2}Tr(d\Sigma^{-1}\Sigma) \quad (15)$$

Na osnovu (15) moguće je dobiti funkcionalne izraze za izvod svakog od parametara:

$$\frac{d \log P}{d\alpha_i} = -(Y-\mu)^T I^{(i)}(Y-\mu) + (2V^{(i)T} - \mu^T I^{(i)})(Y-\mu) + Tr(I^{(i)}\Sigma) \quad (16)$$

$$\frac{d \log P}{d\beta_{ij}} = -(Y-\mu)^T I^{(i,j)}(Y-\mu) + (2V^{(i,j)T} - \mu^T I^{(i,j)})(Y-\mu) + Tr(I^{(i,j)}\Sigma) \quad (17)$$

gde matrice $I^{(i)}$, $I^{(i,j)}$ predstavljaju parcijalne izvode matrice preciznosti od α_i i β_{ij} parametara respektivno, a vektori $V^{(i)}$, $V^{(i,j)}$ predstavljaju parcijalne izvode vektora b pomoćnih promenljivih.

$$I^{(i)}_{a,b} = \begin{cases} 1, & \text{for a = b = i.} \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

$$I_{a,b}^{(i,j)} = \begin{cases} \sum_j 1, & \text{for } a = b = i \text{ , } (i,j) \text{ is edge in graph.} \\ -1, & \text{if } (a,b) \text{ or } (b,a) \text{ are in graph.} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

$$V_a^{(i)} = \begin{cases} \sum_b 1, & \text{for } a = i. \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

$$V_a^{(i,j)} = \begin{cases} \sum_j 1, & \text{for } a = i \text{ where } (i,j) \text{ in graph.} \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

2.3.2 Interpretacija modela

Interpretacija modela definisanog u jednačini (7), kojeg nazivamo i Distance GCRF ili DGCRF, je sledeća:

1. Gausovski asocijativni potencijali su istog oblika kao i u originalnom GCRF modelu [55], uz razliku da ne postoji deljenje parametara između čvorova.

2. Gausovski interakcioni potencijali (6) sada teže da postave razliku između i-tog i j-tog čvora da budu tačno $D_{ij}(X)$. Preciznost ("jačina") interakcije svakog para varijabli (y_i, y_j) je proporcionalna $\sqrt{\beta_{ij}}$.

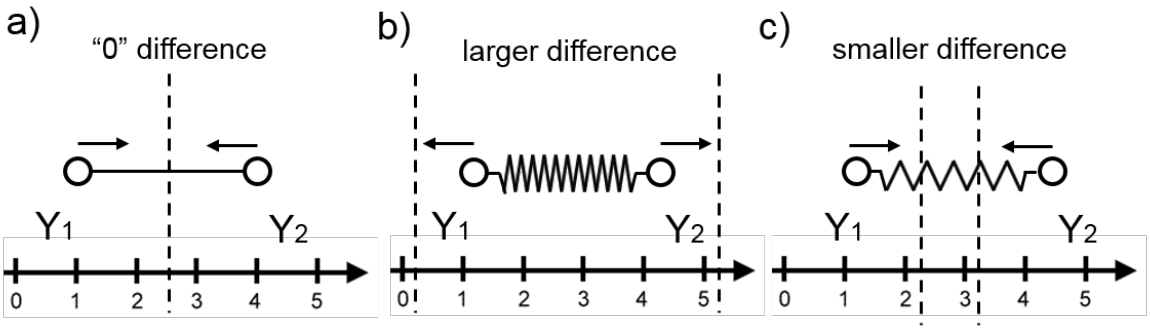
Takođe, moguće je primetiti da u novoj formulaciji interakcionog potencijala ne postoji član koji odgovara unapred predefinisanoj sličnosti S_{ij}. U novom modelu težinski koeficijenti grana grafa se uče na osnovu podataka (ali ne i grane koje su predefinisane) i ne postoji deljenje parametara između različitih grana grafa.

Vrednosti matrice udaljenosti D direktno utiču na kvalitet predikcije. Ukoliko ova matrica ne odgovara problemu, predikcije će biti lošije (u smislu srednje kvadratne greške). Jedan od načina da se estimiraju ovi parametri je korišćenjem regresionih metoda kao i u slučaju učenja nestruktuiranih prediktora pojedinačnih izlaznih varijabli (parametri $R_i(X)$ modela), uz bitnu razliku što je sada potrebno

naučiti prediktore razlike $\Delta y_{i,j}$ između varijabli y_i i y_j na osnovu opservabli X. Logično pitanje koje se nameće je da li je potrebno učiti razliku dva modela, kada smo već naučili pojedinačne modele? Odgovor na ovo pitanje je da to ne mora nužno biti slučaj, što je ilustrovano primerom na slici 2 u kojem dve varijable imaju visoko korelirani šum. Prilikom estimacije pojedinačnih promenljivih, šum u značajnoj meri utiče na estimirane vrednosti (Slika 2 levo). Međutim, kada se estimira razlika ove dve promenljive (Slika 2 desno), može doći do poništavanja šuma usled korelisanosti i omogućiti bolju estimaciju nego u slučaju originalnih izlaza.

Kako bi se bolje objasnio model, moguće je napraviti analogiju između ponašanja prvobitno uvedenog potencijala interakcije (5) i gumene trake. Koeficijent ponderisanja u potencijalu interakcije slično se ponaša kao koeficijent elastičnosti gume; što je koeficijent veći, to je veća sila koja pokušava da smanji rastojanje između njenih krajeva. Što je još važnije, elastična traka može samo smanjiti razliku, ne može je povećati, jer gumena traka ne može potisnuti svoje krajeve u suprotnim smerovima. U nekim slučajevima, estimirane vrednosti dve promenljive nije potrebno približiti već udaljiti. U grafičkim modelima sa diskretnim vrednovanima varijablama, takvi parovi se zovu "odbojni čvorovi" (eng. repulsive nodes).

Korigovani potencijal interakcije (6) ne primorava promenljive da imaju tačno nultu razliku, već umesto toga, uvodi parametar rastojanja. Kada je taj parametar rastojanja fiksiran, potencijali interakcije teže da postave susedne vrednosti na tu razliku. Ukoliko su vrednosti dve susedne promenljive suviše udaljene (tako da je njihova razlika veća od parametra), ovako definisan potencijal interakcije će težiti da ih približiti, tj. da smanji njihovu razliku, slično kao kod originalnog GCRF-a. Ako su vrednosti dve susedne promenljive suviše blizu (tj. njihova razlika je manja od odgovarajućeg parametra udaljenosti), potencijal interakcije će težiti da ih razdvoji ka željenom rastojanju (tj. da poveća njihovu razliku). Ovo ponašanje se razlikuje od modela gumene trake i može se posmatrati kao model opruge, gde opruge imaju neku nominalno dužinu koju će zadržati u mirovanju, ali će se u slučaju istezanja



Slika 4: Interpretacija ponašanja DGCRF modela. Običan GCRF se ponaša kao elastična traka, tj može samo da učini povezane varijable više ili manje sličnim (a), u zavisnosti od "koeficijenta elastičnosti" beta. DGCRF se ponaša kao opruga, tj može po potrebi da približi (c) ili odalji (b) povezane promenljive, u zavisnosti od "nominalne dužine opruge" $D(X)$ i "koeficijenta elastičnosti" beta.

ili kompresije odupreti promeni, težeći da vrati sistem u nominalno stanje. (Slika 4)

2.3.3 Empirijska evaluacija

U ovom poglavlju prikazana je empirijska evaluacija razvijenog modela. Kako bismo okarakterisali unapređenje ovog modela u odnosu na početni, sprovedeno je nekoliko experimenata opisanih u nastavku.

Sintetički podaci

GCRF model zasnovan je na pretpostavci da su očekivane vrednosti i kovarijansa procesa koji generiše uzorke deterministička funkcija (R i Σ) koja zavisi od ulaznih promenljivih (X) problema. Pretpostavka je da postoji više izlaznih promenljivih (Y) koje treba predvideti, dok prostor ulaznih promenljivih (X) može ali i ne mora biti multidimenzionalan. Generisano je nekoliko sintetičkih primera, tako što su najpre generisane vrednosti pomoću parametrizovanih polinoma, na koje je dodat korelisan šum kako bi se uključio efekat strukturne povezanosti između izlaznih

promenljivih, opisan u prethodnim poglavljima. Kao strukturni obrazac između varijabli postavili smo dva prototipska primera, linearni lanac i regularnu dvodimenzionalnu mrežu (grid). Takve strukture su jednostavne i retke (imaju mali broj konekcija), ali iako veštački generisane mogu opisati različite probleme koji imaju vremensku komponentu, ili mogu se naći u aplikacijama sa sekvencijalnom ili prostornom zavisnošću.

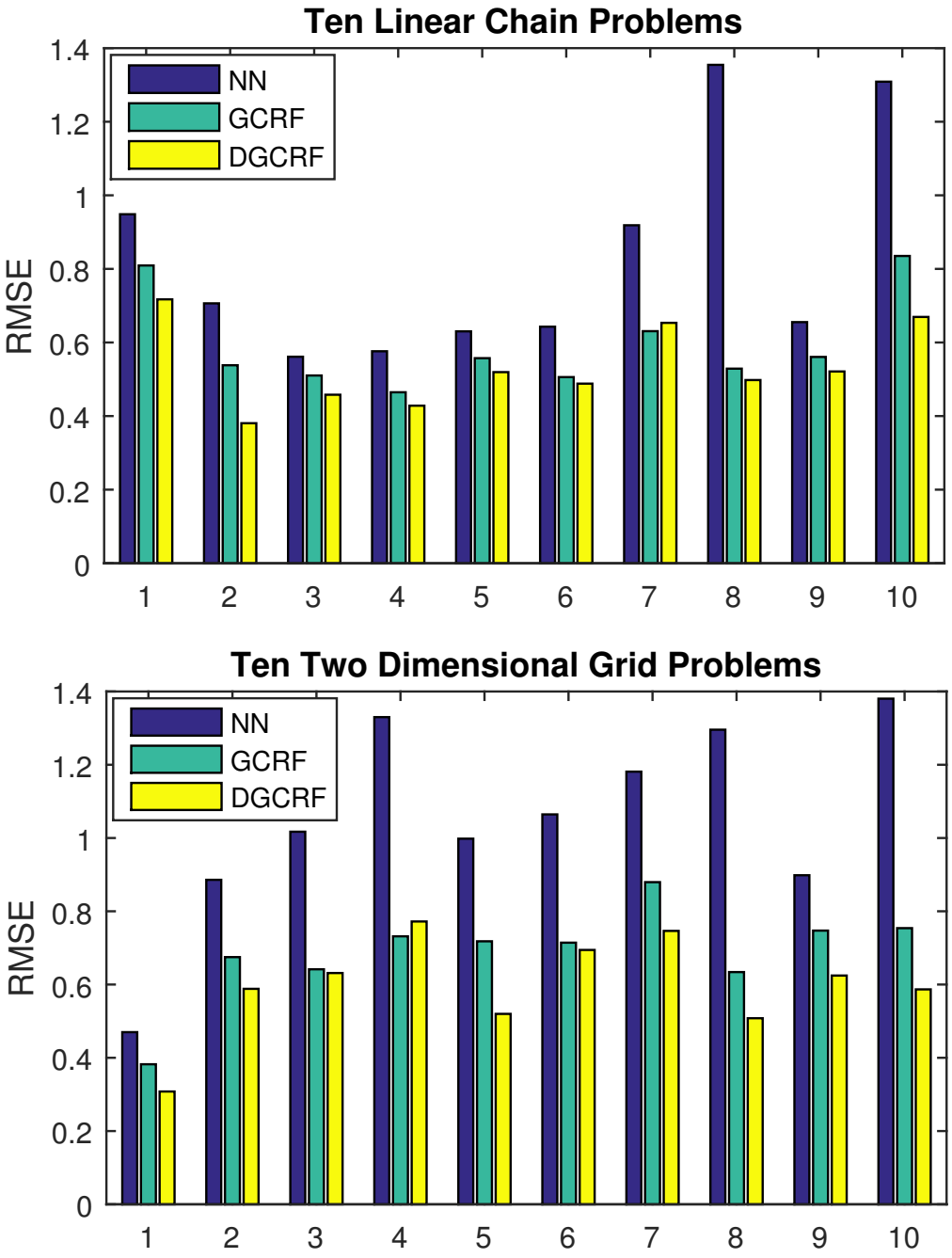
U prvom tipu sintetičkih eksperimenata razmotreno je 10 primera linearne lančane veze od 10 varijabli. Veza između promenljivih u lancu predstavljena je kao odgovarajuća retka matrica. Slično tome, u drugoj vrsti sintetičkih primera, mrežna struktura koja sadrži 9 čvorova (3 x 3), stvorena je deset puta uz slučajnu inicijalizaciju očekivanih vrednosti prediktora i šuma.

U svim sintetičkim eksperimentima uzorkovali smo 250 primeraka iz prethodno opisanog procesa i na njima obučili neuronsku mrežu (NN), koja je izabrana kao nestrukturirani prediktor. Još 150 primera iskorišćeno je kako bi pomoću naučenih neuralnih mreža generisala predviđanja na kojima smo zatim obučavali GCRF i DGCRF modele. Oba pristupa su se oslanjali na nestruktuirane predikcije i testirani su na još 600 testnih primera. Performanse predviđanja merene su kao koren srednje kvadratne greške (RMSE). Svaki eksperiment je ponovljen deset puta kako bi se dobila procena pouzdanosti modela. Rezultati prikazani u tabeli 10 (takođe prikazani na slici 5) su prosečne performanse na deset različitih sintetičkih skupova podataka.

Tabela 1: Greške predikcije (RMSE) tri različite metode na dva različita tipa grafa. Najbolje performanse za svaki tip je označena (p-vrednosti, 0.011 i 0.0043).

Metod	Ulančana struktura	2D rešetka
Nestruktuiran-NN	0.830 ± 0.295	1.052 ± 0.270
Struktuiran-GCRF	0.594 ± 0.128	0.688 ± 0.127
Struktuiran-DGCRF	$\mathbf{0.533 \pm 0.111}$	$\mathbf{0.598 \pm 0.134}$

Predloženi DGCRF model, sa modifikovanim potencijalom interakcije (6) je u



Slika 5: Rezultati na sintetičkim podacima. Greška predikcije tri različite metode na dva problema sa različitom vrste strukture. Na levoj slici je prikazan koren srednje kvadratne greške (RMSE) na deset različitih sintetičkih problema sa strukturom linearnog ulančanog grafa, dok se na desnoj slici mogu videti deset različitih problema na dvodimenzionalnoj rešetki.

odnosu na polazni GCRF povećao kvalitet prediktora (uz statističku značajnost ispod granice od 0.05).

Aplikacija na pravim podacima

Predloženi metod je dalje okarakterisan kroz primenu na pravim podacima iz dve izazovne aplikacije koje su opisane u ovom odeljku, i eksperimentalno potvrđen kroz uporednu analizu sa alternativama.

Predikcija broja primljenih pacijenata u bolnicama u Kaliforniji (HCUP dataset)

Prva aplikacija novog modela je procena mesečne stope hospitalizacije obolelih od sepse u Američkoj saveznoj državi Kalifornija [63]. Podaci za ovaj eksperiment pružili su Projekat zdravstvene zaštite i upotrebe (eng. Health Care and Utilization Project, skraćeno HCUP) i Državne baze podataka hospitalizovanih pacijenata (eng. State Inpatient Databases, skraćeno SID). Za analizu su odabrane informacije o prijemu i hospitalizaciji pacijenata kojima je dijagnostikovana sepsa iz 231 bolnice tokom 108 meseci. Sepsa je dijagnoza sa jednom od najviših stopa smrtnosti u SAD [69]. Predikcijom ukupnog broja pacijenata sa dijagnozom sepse moguće je pomoći u optimizaciji bolničkih resursa i smanjiti troškove vezane za njih. U našim eksperimentima razmatrane su vremenske serije koje se sastoje od zdravstvenih evidencija koje se odnose na sepsu iz 231 bolnice. Stopa prijema u narednom mesecu predviđena je iz prethodnih tromesečnih stopa prijema za svaku bolnicu. Na svakom čvoru u grafu koji sadrži 231 bolnicu u Kaliforniji obučavali smo nestruktuirane prediktore koristeći podatke iz početnih 75 meseci. Koristeći postupak desetostruke kros-validacije (10 K-fold) na tih 75 meseci naučeni su i struktuirani prediktori. Preostalih 33 meseca korišćeni su kao test. Graf sličnosti između bolnica estimiran je iz statističkih bolničkih podataka (vezanih ne samo za

sepsu) kako bi se obezbedila struktura neophodna za struktuirane modele. Zatim je sličnost između bolnica dobijena kao Jensen-Shannon-ova sličnost raspodele mortaliteta, a zadržani su samo veliki koeficijenti kako bi se dobio redak graf visoko povezanih bolnica (slika 6, donji panel).

Predloženi metod se dalje karakterizuje i upoređuje sa alternativama na dve izazovne aplikacije iz stvarnog sveta koje su ukratko opisane u ovom odeljku.

Procena količine padavina u kontinentalnom delu Sjedinjenih Američkih Država (RAIN dataset)

Druga aplikacija razvijenog modela je na podacima pribavljenim iz Nacionalnog centra za klimatske podatke (eng. National Climatic Data Center) [48]. Podaci o količini padavina su sakupljeni sa meteoroloških stanica širom SAD-a. U eksperimentalnoj analizi posmatrali smo mesečnu količinu padavina tokom 708 meseci u 1132 lokacije. Pored padavina, koristili smo 6 varijabli stečenih iz projekta NCEP/NCAR Reanalysis 1 [35]: Lagranžovska tendencija vazdušnog pritiska (omega), atmosferska voda, relativna vlažnost, temperatura, zonalne i meridionalne komponente vetra, koje se obično koriste za predviđanje klimatskih parametara[1]. Ove varijable su korišćene kako bi se procenio nivo padavina. Nekoliko nestruktuiranih prediktivnih modela naučeno je na slučajno odabranih 250 od početnih 400 meseci, dok su strukturirani modeli naučeni za preostalih 150 meseci, a učinak je procenjen na narednih 308 meseci. Zatim je kreirana struktura interakcija formiranjem grafa prostornog susedstva na osnovu tri najbliže lokacije (slika 6, gornji panel). Motiv za formiranje ovakvog grafa je zasnovan na ideji da će obližnja mesta imati slične klime. Ceo eksperiment je ponovljen 10 puta.

U oba eksperimenta na pravim aplikacijama, DGCRF model je bio precizniji u poređenju sa alternativnim modelima (Tabela 2). Nestrukturirani model koji se ko-

[1]podaci dostupni na veb stranici Nacionalne Okeanografske i Atmosferske Administracije (eng. National Oceanic and Atmospheric Administration, skraćeno NOAA): http://vvv.esrl.noaa.gov/psd/

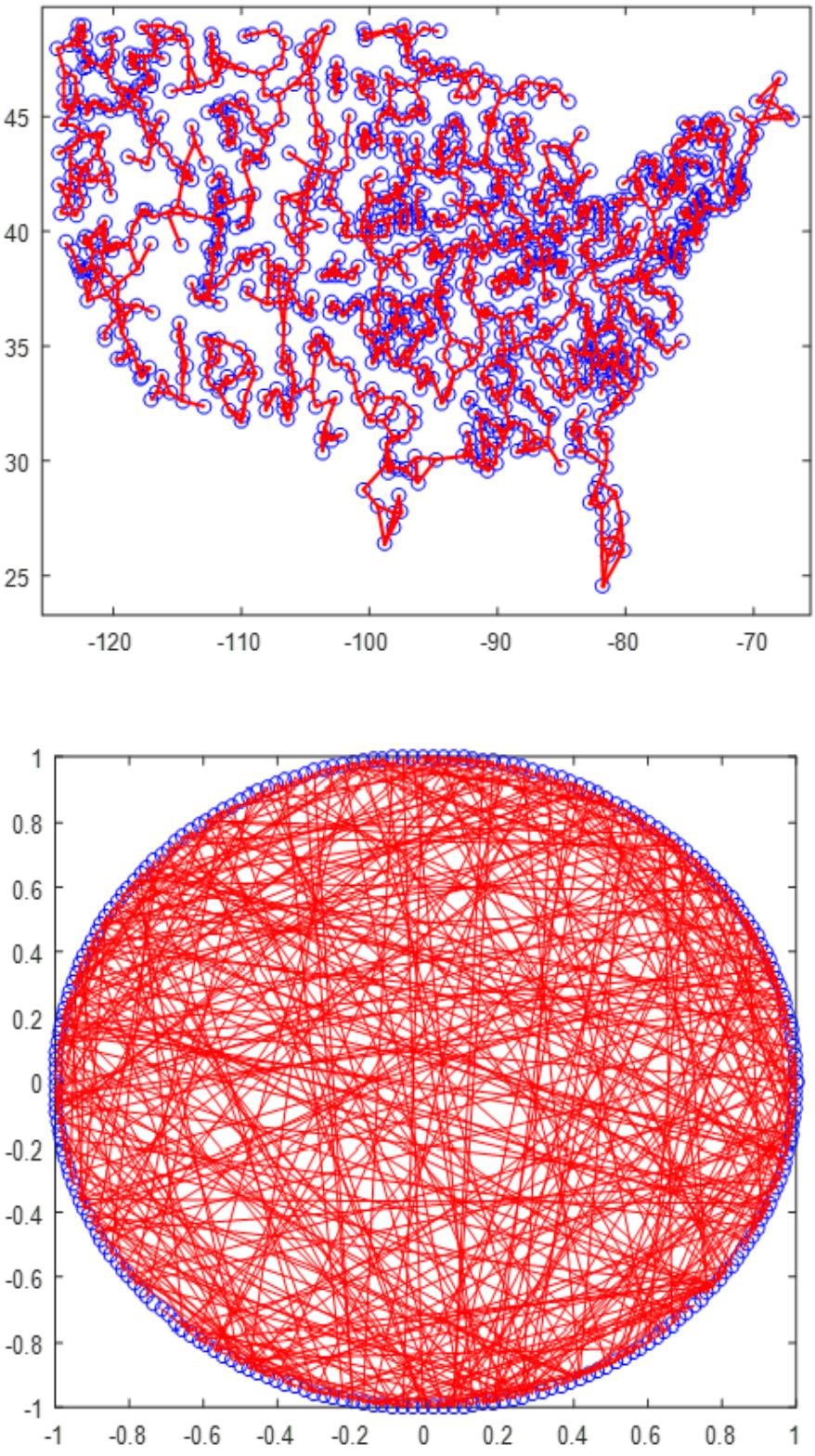
risti za predikciju oba strukturirana pristupa je Gaussian Process Regression (GPR) sa Gausovim jezgrom, gde su hiperparametri naučeni metodom maksimalne verodostojnosti. Poboljšanje tačnosti DGCRF-a, u poređenju sa osnovnim GCRF-om, u dvostranom t-testu je statistički značajno (p-vrednost 2.5e-9 i 7.8e-8 na RAIN i SEPSIS datasetu respektivno). Ukupni rezultati sugerišu prednost u korišćenju potencijala interakcije DGCRF u odnosu na polazni model GCRF.

Tabela 2: Greška predikcije (RMSE) na dve realne aplikacije. Najbolje performanse su istaknute u tabeli.

Metod	RAIN	SEPSIS
Nestruktuirani-GPR	1.799 ± 0.010	1.272 ± 0.020
Struktuirani-GCRF	1.790 ± 0.007	1.265 ± 0.020
Struktuirani-DGCRF	$\mathbf{1.767 \pm 0.004}$	$\mathbf{1.238 \pm 0.018}$

2.3.4 Diskusija i ideje za budući rad

U ovom poglavlju prikazali smo jedno unapređenje Gausovskih slučajnih uslovnih polja. Predložili smo modeliranje interakcija zasnovanih na daljini između izlaznih varijabli u mreži i pružili dokaze da je ovakav pristup tačniji od alternativa zasnovanih na sličnosti korišćenih u objavljenim metodama. To je postignuto reformulacijom modela GCRF-a i njegovom generalizacijom kako bi se omogućio nenulti parametar srednjeg rastojanja u potencijalu interakcije. Dodatni stepen slobode uveden kod DGCRF modela rezultirao je povećanjem tačnosti predviđanja u odnosu na najčešće korišćeni osnovni model, što je eksperimentalno pokazano na više sintetičkih problema, i dve realne aplikacije. Dodatno, predložena generalizacija nije ugrozila atraktivna svojstva brzog zaključivanja i konveksnosti optimizacije parametara u struktuiranoj regresiji. Iako je predstavljen model interakcije zasnovan na daljini posebno pogodan za GCRF, on nije ograničen samo na korišćenje u isključivo Gausovskim modelima. Predstavljen model interakcije se lako može primeniti i na druge verovatnosne grafovske modele.



Slika 6: Grafovske strukture. Na gornjoj slici je graf susednosti korišćen za klimatološku aplikaciju, a na donjoj slici je prikazan graf koji opisuje sličnost između bolnica po distribuciji stope smrtnosti.

2.3.5 Analiza vremenske kompleksnosti

Polazni model GCRF, kao specijalna varijanta razvijenog modela ne koristi parametre udaljenosti u potencijalu interakcije, iz čega sledi da je novi model zahtevniji utoliko što je neophodno naučiti regresivni model za svaki od njih. Broj nestruktuiranih modela koje je potrebno naučiti (s obzirom da ne postoji deljenje parametara) je jednak broju grana u grafu koji odgovara problemu. U najnepovoljnijem slučaju, kada je struktura potpun graf, broj pojedinačnih modela raste kvadratno sa brojem izlaznih promenljivih. Međutim, ovakav scenario se najčešće ne sreće u praksi. Struktura u problemima od interesa je često proređena, kao što je pokazano u sekciji sa rezultatima. Mnogi složeni sistemi poput bioloških, društvenih, itd. imaju tendenciju da imaju tzv. scale-free svojstvo, koje karakteriše eksponencijalna raspodela stepena čvorova grafa, što rezultira u praktično linearnoj zavisnosti broja veza od broja čvorova. Više o problemima koji imaju ovo svojstvo, kao i o metodama koji ovo svojstvo eksploatišu će biti u poglavlju koje se bavi učenjem strukture na osnovu podataka.

Što se tiče povećanja broja slobodnih parametara modela, koje je nastalo relaksiranjem pretpostavke deljenih parametara između grana grafa, povećana je i komputaciona zahtevnost. To povećanje se ogleda u činjenici da je potrebno više iteracija kako bi algoritam optimizacije konvergirao u prostoru sa više dimenzija. Međutim, asimptotska kompleksnost ostaje ista, jer glavno opterećenje i dalje dolazi od matrične inverzije.

Metod predložen u ovom poglavlju ima dodatni računski trošak u poređenju sa polaznim modelom, ali je primenljiv svuda gde je primenjiv i polazni model, jer imaju ista ograničenja u pogledu veličine problema.

3 Učenje strukture iz podataka

U ovom poglavlju fokus će biti na nesupervizovanoj estimaciji strukture zavisnosti slučajnih promenjivih iz obzerviranih podataka. Tekst, metodologija i rezultati prikazani u ovom poglavlju su zasnovani na radovima objavljenim u [71], i u [34].

3.1 Uvod

Uprkos tome što Gausova pretpostavka omogućava efikasno zaključivanje i učenje parametara modela, efikasna upotreba ovakvih modela zavisi i od osobina strukture problema. Smanjenje broja parametara, tj retkost u matrici parametara modela je u direktnoj vezi sa jednostavnošću problema i dobro je usklađena sa principom Okamove oštrice. Proređenost smanjuje tendenciju preobučavanja (overfit-a) i povećava interpretabilnost modela otkrivanjem ne tako očiglednih odnosa između varijabli. Stoga, forsiranje retkosti parametara pri učenju GMRF-a predstavlja atraktivan način otkrivanja osnovne strukture podataka.

Sve ovo čini GMRF veoma korisnim alatom za statističko učenje, koji je primenjiv u raznim domenima gde su prisutni visoko-dimenzionalni podaci, uključujući biomedicinu [15, 65], epidemiologiju [60], geo-spacijalne probleme [4], ekologiju [33], analizu slike [17], računarsku viziju [43], i čak i društvene nauke [20].

Može se pokazati da je učenje GMRF modela ekvivalentno rekonstrukciji inverzne matrice kovarijanse (matrice preciznosti) iz podataka. Inverzna kovarijansna matrica nosi informacije o uslovnoj zavisnosti između promenljivih, ili ekvivalentno graf povezivanosti. Postoji nekoliko metoda razvijenih za rešavanje estimacije problema retke inverzne matrice kovarijanse. U toj oblasti, posebna pažnja posvećena je računskoj efikasnosti metoda. Sa sve većom količinom i dimenzionalnošću dostupnih podataka, postoji potreba za još bržim i efikasnijim algoritmima [47, 87, 12, 2, 21, 18, 59, 57, 30, 31, 79],

Savremene aplikacije nameću potrebu za obradom sve većih višedimenzionalnih

datasetova, pa su i efikasnost i skalabilnost od presudnog značaja. Jedan od opštih način postizanja poboljšanja komputacione zahtevnosti je eksploatacija (očiglednih) strukturnih regularnosti u problemu. Najčešće korišćena pretpostavka je retkost, što rezultira velikim brojem poželjnih osobina: jednostavnost, povećana generalizibilnost, smanjeni zahtevi za izračunavanje i reprezentaciju. Predložene su razne metode za rešavanje problema estimacije inverzne kovarijansne matrice [18, 21, 30, 47, 57, 59, 79]. Većina ovih pristupa zasnovana je na optimizaciji regularizovane maksimalne verodostojnosti, gde se L_1 penal direktno primenjuje na elemente inverzne kovarijansne matrice.

U nastavku ovog odeljka biće predstavljene dve nove metode za brzo učenje retkih GMRF-ova iz visokodimenzionalnih podataka, koji se oslanjaju na Cholesky dekompoziciju.

Prvo će biti predložen model u kojem predstavljamo regularizacioni član, zasnovan na L_1 normi, koji penalizuje aproksimirani proizvod Cholesky faktora inverzne kovarijansne matrice i biće definisane pretpostavke modela i izvođenje algoritma za učenje zasnovanog na koordinatnom spustu. U nastavku će biti pokazano da novi model ima poželjne osobine konveksnosti i garanciju konvergencije algoritma za učenje. Takođe, učenje ovog modela koristi tzv. Active-Set pristup sličan onom koji je predložen u [31], kako bi se postiglo značajno ubrzanje. Takođe, ovaj pristup pripada klasi tzv. "Embarrassingly parallel" algoritama, i može se lako skalirati na multi-threaded arhitekturama.

Zatim, usvojićemo drugačiju pretpostavku strukture grafa i definisaćemo optimizacionu funkciju koja direktno penalizuje elemente Cholesky faktora umesto elemente inverzne kovarijansne matrice. U nastavku će biti pokazano da je ova pretpostavka posebno prilagođena tzv. Scale-Free grafovima. U odeljku Metoda prikazane su pogodne osobine takve formulacije, na osnovu čega predlažemo SNETCH, efikasan algoritam optimizacije zasnovan na koordinatnom spustu. Pokazujemo da ovaj pristup takođe pripada klasi "Embarrassingly parallel" problema, i uz ko-

rišćenje active-set pristupa može postići značajno ubrzanje u poređenju sa postojećim metodama.

U oba modela koja predlažemo, jedna od glavnih prednosti je da, dok optimizujemo direktno Cholesky faktore, inverzna kovarijansna matrica parametra će u svakom trenutku biti pozitivno definitna (PD), dok većina drugih pristupa mora redovno da proverava taj uslov, i koriguje trenutni optimizacioni korak (korišćenjem Cholesky dekompozicije, Armijo backtrack pretrage, Schur komplementa itd.), što značajno usporava optimizaciju i čini je neskalabilnijom.

3.2 Pregled postojećih rešenja

Problem estimacije nepoznate kovariansne matrice Σ (ili inverzne kovarijanse matrice Σ^{-1}) je problem u kojem je potrebno rekonstruisati (ili aproksimirati) matricu na osnovu uzoraka iz multivarijatne distribucije. Empirijska kovarijansna matrica $\frac{1}{N-1}\sum(y-\hat{y})(y-\hat{y})^T$ je nepristrasan estimator kovarijansne matrice. Međutim, empirijska kovarijansna matrica i njena inverzna matrica nisu stabilne za probleme sa velikim brojem dimenzija.

Ukoliko se usvoje dodatne pretpostavke o distribuciji modela (npr. da je distribucija uzoraka multivarijatna Normalna raspodela), moguće je primeniti metodu maksimalne verodostojnosti. Ovo je preferirani pristup učenju kovarijanske matrice (i njene inverzije) u literaturi.

3.2.1 Estimacija maksimalne verodostojnosti

Slučajni vektor $x \in \mathbb{R}^{p\times 1}$ (p-dimenzionalni kolona-vektor slučajnih promenljivih) ima ne-degenerisanu multivarijatnu normalnu distribuciju sa kovarijansnom matricom Σ ako i samo je matrica $\Sigma \in \mathbb{R}^{p\times p}$ pozitivno definitna i funkcija gustine verovatnoće x je

$$d(x) = (2\pi)^{-p/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{22}$$

gde je $\mu \in \mathbb{R}^{p\times 1}$ očekivana vrednost x. Ako pretpostavimo n nezavisnih i identično raspodeljenih (IID) opservacija: $\{x_1, x_2, ..., x_n\} \in \mathbb{R}^{p\times 1}$, ukupna verodostojnost je proizvod pojedinačnih verodostojnosti.

S obzirom da je estimator maksimalne verodostojnosti vektora srednje vrednosti μ srednja vrednost odabiraka, tj. vektor $\overline{x} = (x_1 + \cdots + x_n)/n$, centriramo podatke i usvajamo matričnu notaciju: $X = [x_1 - \overline{x}, x_2 - \overline{x}, ..., x_n - \overline{x}] \in \mathbb{R}^{p\times n}$. Nakon što odbacimo konstantan faktor skaliranja $(2\pi)^{-np/2}$ verodostojnost se može izraziti kao:

$$\mathcal{L}(\Sigma^{-1}) \propto \det(\Sigma^{-1})^{n/2} \exp\left(-\frac{1}{2} tr\left(X^T \Sigma^{-1} X\right)\right) \tag{23}$$

U praksi, lakše je optimizovati logaritam verodostojnosti (log-likelihood) i svesti problem na minimizaciju negativne logaritmovane verodostojnosti:

$$\hat{\Sigma}^{-1} = \underset{\Sigma^{-1}}{\operatorname{argmin}} \frac{1}{2} tr\left(X^T \Sigma^{-1} X\right) - \frac{n}{2} log|\Sigma^{-1}| \tag{24}$$

sa PD uslovom $x^T \Sigma^{-1} x > 0$, $\forall x \in \mathbb{R}^{p\times 1}$.

Lako je uočiti da u slučaju kada je slučajna promenljiva normalno raspodeljena, kovarijansna matrica sempla je raspodeljena po Wishart raspodeli, i njena estimacija maksimalne verodostojnosti je $S = XX^T/n$. Međutim, ako je $n < p$, kao što je često slučaj kod visokodimenzionih problema, empirijska kovarijansna matrica nema pun rang (singularna je) i stoga nije moguće odrediti inverznu kovarijansnu matricu.

3.2.2 Penalizovana estimacija maksimalne verodostojnosti

Zbog toga se uvode dodatne pretpostavke, obično kroz apriori vrednosti parametara, što je samo još jedan pogled na regularizaciju. Inverzna kovarijansna matrica kodira strukturu nezavisnosti, i ona treba da bude retka.

L_0 norma na inverznoj kovarijansnoj matrici Regularizacija koja forsira retkost matrice je prvobitno bila primenjena u [14] optimizacijom verodostojnosti sa minimalnim brojem nenultih elemenata u inverznoj kovarijansnoj matrici: $min\|\Sigma^{-1}\|_0$. Međutim, ovaj kombinatorno-optimizacioni problem brzo postaje neizračunljiv (netraktabilan) sa povećanjem broja varijabli. Prvobitni pristupi za rešavanje ovog problema bili su zasnovani na greedy search metodama u prostoru mogućih grafova [42], koji se nisu dobro skalirali zbog neophodnosti izračunavanja maksimalne verodostojnosti u svakoj iteraciji i zato što ih nije moguće primeniti u slučajevima kada je dimenzionalnost problema veća od broja uzoraka.

Često je broj uzoraka manji od broja varijabli koje je potrebno estimirati, što je uslovilo drugačiji pristup učenja retke matrice inverzne kovarijanse iz podataka [3]. Problem kombinatorne optimizacije L_0 norme aproksimiran je pogodnom konveksnom formulacijom u vidu L_1 norme, koja takođe uslovljava izvestan stepen retkosti parametara. Uvedena L_1 norma je pogodna jer je konveksna, ali ipak vodi do retkog rešenja. Prva metoda koja je koristila L_1 normu bila je Neighborhood selection [47], koja bira povezane varijable rešavanjem problema nekoliko razdvojenih LASSO [78] problema. Iako pristup nije direktno rešavanje problema L_1 regularizovane maksimalne verodostojnosti, može se posmatrati kao jednostavna aproksimacija tog problema, po jedna za svaku varijablu za koju je potrebno obučiti regresiju u odnosu na druge.

Nakon [47], usledilo je nekoliko pristupa [87, 12, 2] koji su rešavali direktno problem maksimalne verodostojnosti L_1 pomoću alata konveksne optimizacije [10]. Međutim, svi su bili računski zahtevni i pogodni samo za male probleme. Posle toga, predložene su druge metode kao što su Grafički LASSO [21] koji je poboljšao računsku efikasnost učenja matrice retke preciznosti predlaganjem brže optimizacije zasnovane na algoritmu koordinatnog spusta. U [18] predložen je brz projected gradient metod, koji može rešiti i opštije vrste problema sa blok-retkosti. Pomenute metode optimizuju dualni problem, dok SINCO algoritam [59] rešava direktni prob-

lem koristeći greedy koordinatni spust. GISTA pristup [57], zasnovan na metodi proksimalnog gradijenta, ima atraktivne teorijske garancije u vezi sa graničnim greškama i brzinom linearne konvergencije.

Svi pomenuti algoritmi koriste samo informacije o gradijentu prvog reda, što u najboljem slučaju ograničava njihovu konvergenciju na linearnu. Među prvim pristupima koji koriste informacije o gradijentu drugog reda su QUIC [30], koji koristi koordinatni spust zasnovan na Njutnovoj metodi sa Armijo backtrack pretragom. Pravac spuštanja se prvo pronalazi korišćenjem koordinatnog spusta zasnovanog na active-setu, nakon čega sledi neefikasni line search da bi se pronašao korak koji dovoljno smanjuje vrednost optimizcione funkcije na osnovu Armijo pravila. Parametri se drže unutar pozitivno definisanog konusa u svakoj iteraciji, izvršavanjem Cholesky dekompozicije, i pronalaženjem veličine koraka koja osigurava da inverzna kovarijansna matrica parametara bude pozitivno definitna. Poboljšanje QUIC pristupa za velike višedimenzionalne podatke [31] smanjuje računsku zahtevnost tako što traži pravac spuštanja pametnom upotrebom conjugate gradient metode. Rad [31] takođe preporučuje izbegavanje koraka u kojem se koristi Cholesky dekompozicija, koji sa povećanjem veličine problema postaje usko grlo optimizacije. Kao što ćemo pokazati u nastavku, uzrok tome je odabir takvih grafova za koje popunjavanje faktora Cholesky matrice raste nelinearno sa veličinom problema. Iz toga razloga metod definisan u [31] osigurava pozitivnu definitinost matrice parametara metodom Schur komplementa. Pristup BCDIC [79] navodi dalje napretke u ovom pravcu tako što optimizuje nekoliko varijabli istovremeno, koristeći optimizacionu metodu blok-koordinatnog spusta. Nadogradnja ovog pristupa nazvana ML-BCDCIC ide korak dalje koristeći hijerarhiju na više nivoa [80]. Sva tri pristupa estimacije velike inverzne kovarijansne matrice [31, 79, 80] koriste METIS [36] algoritam za brzo particionisanje grafova, kako bi razlikovali submatrice sa većim i manjim gustinama popunjenosti. Ovaj pristup prirodno odgovara grafovima generisanim kroz stohastički blok-model (tzv. community networks) i Erdos-Renyi proces (slučajne

binarne mreže).

L_1 penal na elemente Cholesky faktora Još jedna formulacija problema, CSEPSNL, je predložena u [32]. U tom radu, L_2 i L_1 penal je primenjen na elemente Cholesky faktora L (gde je $\Sigma^{-1} = LL^T$), umesto direktno na elemente inverzne kovarinansne matrice Σ^{-1}.

Drugi pristupi estimacije inverzne kovarijansne matrice koji uvode drugačije pretpostavke kao što su struktura niskog ranga [27] i tzv. ridge penal [29] nisu u fokusu ovog istraživanja.

3.3 Metod

U ovom odeljku izvešćemo dva različita pristupa učenja GMRF, odnosno estimacije inverzne kovarijansne matrice, SCHL koji je pogodniji za probleme čija struktura odgovara tipu Erdos-Renyi, i SNETCH koji je pogodniji za probleme čija struktura poseduje scale-free svojstvo.

Kako bismo izveli SCHL i SNETCH modele, i procedure za njihovo obučavanje potrebno je najpre poći od standardnog pristupa u kojem se kriterijumska funkcija raščlani na diferencijabilni $g(L)$ (glavni) i nediferencijabilni deo $h(L)$ (koji uzrokuje retkost rešenja):

$$l(L) = g(L) + h(L) \tag{25}$$

Najpre ćemo posmatrati diferencijabilni deo, i iskoristiti činjenicu da je trag matrice primenjen na proizvod matrica invarijantan na ciklične permutacije tog proizvoda, kao i to da je determinanta matričnog proizvoda jednaka proizvodu determinanti:

$$g(L) = \frac{1}{2} tr(\Sigma^{-1} X X^T) - \frac{n}{2} log(det(\Sigma^{-1})) \tag{26}$$

pod uslovom $x^T\Sigma^{-1}x > 0, \forall x \in \mathbb{R}^{p\times 1}$.

Usko grlo u pronalaženju optimalnog rešenja problema (26) je izračunavanje člana $log(det(\Sigma^{-1})$, za koje je potrebno $O(p^3)$ koraka u naivnoj implementaciji. Takođe, potrebno je obezbediti ispunjenje uslova pozitivne definitnosti: $x^T\Sigma^{-1}x > 0$, $\forall x \in \mathbb{R}^{p\times 1}$, što takođe doprinosi računskoj kompleksnosti. Iz tog razloga u nastavku je predložena drugačija parametrizacija koja dozvoljava efikasnije računanje $log(det(\Sigma^{-1}))$ člana.

Iskoristićemo Cholesky reparametrizaciju problema (26). S obzirom da je inverzna kovarijansna matrica pozitivno definitna, može se dekomponovati na Cholesky faktore:

$\Sigma^{-1} = LL^T$ gde je L donje-trougaona matrica sa pozitivnih dijagonalnim elementima. Lako se može uočiti da je u ovakvoj parametrizaciji problema (26) uslov da matrica Σ^{-1} mora biti pozitivno definitna pretvoren u znatno jednostavniji uslov $L_{ii} > 0$. Takođe, moguće je uočiti još nekoliko drugih prednosti koje će biti eskploatisane u našem modelu:

1. Ukoliko je Cholesky matrica L retka matrica i ima odgovarajući obrazac popunjenosti, LL^T će takođe biti retka matrica;

2. Cholesky faktori se mogu efikasno iskoristiti prilikom računanja inverzne matrice. Ukoliko su Cholesky faktori dostupni tokom cele optimizacije bez troškova dodatnog računanja, inače skupe operacije koje uključuju matričnu inverziju se mogu obaviti efikasno

Pored ovih prednosti, korišćenje Cholesky parametrizacije kriterijumske funkcije (26) takođe ima nekoliko loših strana, na koje ćemo posebno obratiti pažnju. Jedna od njih je problem regularizacije inverzne kovarijansne matrice koju je znatno teže optimizovati u novoj formulaciji.

Diferencijabilni deo kriterijumske funkcije sada postaje:

$$g(L) = \frac{1}{2} tr(LL^T XX^T) - \frac{n}{2} log(|L||L^T|) \tag{27}$$

uz ograničenje: $L_{ii} > 0, \forall i$ koje ćemo izostaviti u daljem tekstu zbog čitljivosti, i smatrati da se podrazumeva.

Izraz XX^T u prvom članu moguće je zameniti empirijskom kovarijansnom matricom, i uočiti da je determinanta Cholesky faktora u drugom članu jednaka proizvodu dijagonalnih elemenata $det(LL^T) = \prod_{i=1}^{p} L_{ii}^2$, čime diferencijabilni deo kriterijumske funkcije postaje:

$$g(L) = \frac{n}{2} tr(LL^T S) - \frac{n}{2} log(\prod_{i=1}^{p} L_{ii}^2) = \frac{n}{2} tr(LL^T S) - n \sum_{i=1}^{p} log(L_{ii})$$

U novoj reparametrizaciji kriterijumske funkcije, log det član se može efikasno izračunati u $O(p)$ vremenu. Dalje, deo kriterijumske funkcije koji zavisi od podataka se može razdeliti u nekoliko nezavisnih podciljeva, što je jedan od preduslova za efikasno skaliranje algoritma na visokodimenzionalne podatke. Ovo se može uočiti daljim dekomponovanjem matrice L, koju možemo posmatrati kao sumu p različitih $p \times p$ matrica ranga 1 (koje ćemo obeležiti sa L'_j), gde svaka matrica sadrži samo kolonu j matrice L, dok su ostali elementi postavljeni na nula.

Nakon ovog zapažanja, diferencijabilni deo kriterijumske funkcije dobija oblik (uz odbacivanje konstante proporcionalnosti n):

$$g(L) = \frac{1}{2} \sum_{j=1}^{p} tr(L'_j {L'_j}^T S) - \sum_{j=1}^{p} log(L_{jj}) \tag{28}$$

Na osnovu ovoga je moguće izvesti izraze za parcijalne izvode diferencijabilne funkcije $g(L)$ po parametrima, koji su nam neophodni za algoritam optimizacije. Uočavamo dve različite grupe promenljivih: dijagonalne (when $i = j$), i vandijagonalne (when $i \neq j$), i dve grupe jednačina za ažuriranje parametara pri optimizaciji. Dalje je moguće uprostiti izraz korišćenjem standardnog matričnog računa:

$$\frac{\partial tr(L'_j {L'_j}^T S)}{\partial L_{ij}} = \frac{1}{2} tr((\frac{\partial L'_j}{\partial L_{ij}} {L'_j}^T + L'_j \frac{\partial {L'_j}^T}{\partial L_{ij}})S) = tr(\frac{\partial L'_j}{\partial L_{ij}} {L'_j}^T S) = L_j^T S_i \tag{29}$$

U ovom izrazu vektor S_i je red i matrice S, dok L_j označava kolonu j matrice L. Prilikom izvođenja izraza (29) korišćena je činjenica da je $\frac{\partial L_j}{\partial L_{ij}}$ zapravo nula matrica sa jednim jedinim nenultim elementom vrednosti 1 u i-tom redu and j-toj koloni, kao i ciklično svojstvo trag operatora za simetrične matrice:

$$Tr(\frac{\partial L_j}{\partial L_{ij}} L_j^T S) = Tr(L_j \frac{\partial L_j^T}{\partial L_{ij}} S) \tag{30}$$

Konačno, dobijeni su parcijalni izvodi po parametrima diferencijabilnog dela kriterijumske funkcije:

$$\nabla g(L)_{ij} = \begin{cases} \sum_{k=1}^{p} L_{kj} S_{ik}, i \neq j \\ \sum_{k=1}^{p} L_{kj} S_{ik} - \frac{1}{L_{ii}}, i = j \end{cases} \tag{31}$$

Kao što se može primetiti, diferencijabilni deo kriterijumske funkcije $g(L)$ može se rastaviti na p nezavisnih problema, po jedan za svaku kolonu L matrice, čime su se stekli preduslovi za razvijanje algoritma koji ima visok stepen paralelizma, s obzirom da p koordinata može biti ažurirano u isto vreme. Ovo je veoma korisno svojstvo kada je reč o skalabilnosti jer omogućuje nezavisno i paralelno računanje delova objektivne funkcije, i to sa nivoom paralelizma srazmernom dimenziji problema.

Sada kada imamo izraze za izvode po parametrima, diferencijalni deo kriterijumske funkcije je moguće optimizovati pomoću koordinatnog spusta, i to tako što iterativno rešavamo uslove optimalnosti prvog reda, tj. nalazimo nule izraza 31:

$$L_{ij} = -\frac{\sum_{k \neq i} L_{kj} S_{ik}}{S_{ii}} \tag{32}$$

$$L_{ii} = \frac{-\sum_{k \neq i} L_{ki} S_{ik} + \sqrt{(\sum_{k \neq i} L_{ki} S_{ik})^2 + 4 S_{ii}}}{2 S_{ii}} \tag{33}$$

Drugo rešenje kvadratne jednačine 33 se uvek odbacuje, čime je zadovoljen uslov da L_{ii} mora biti pozitivno da bi matrica Σ^{-1} bila pozitivno definitna. Takođe, može se izvući jos jedan važan zaključak iz jednačina (32) i (33): ukoliko je kolona j matrice L retka sa prostornom kompleksnošću $O(S_{active})$, ažuriranje bilo koje promenljive L_{ij} će imati $O(S_{active})$ vremensku kompleksnost.

U nastavku će biti predstavljena optimizacija nediferencijabilnog dela kriterijumske funkcije, čija je uloga da penalizuje gustinu inverzne kovarijansne matrice. Kao što će biti pokazano u nastavku, proređenost inverzne kovarijansne matrice se može postići na više načina.

Prvo će biti prikazan novi SCHL metod za brzo učenje proređenih Gausovskih Markovljevih slučajnih polja iz kontinualnih podataka. Proređenost inverzne kovarijansne matrice je postignuta njenim dekomponovanjem na Cholesky faktore i dodavanjem L_1 regularizacije na jednu aproksimaciju njihovog proizvoda. Prikazani metod ima sličnosti sa CSEPNL metodom predstavljenim u [32] ali se utoliko razlikuje od njega jer penalizuje aproksimaciju proizvoda Cholesky faktora umesto što direktno penalizuje same elemente.

Zatim će takođe biti predstavljen novi SNETCH metod, u kojem je L_1 penal postavljen na elemente Cholesky faktora (kao u [32]) umesto na elemente inverzne kovarijansne matrice. Biće pokazano da takva kriterijumska funkcija poseduje atraktivna svojstva koja dozvoljava da se razvije veoma brz optimizacioni algoritam zasnovan na koordinatnom spustu i koji je posebno pogodan za probleme sa Scale-free strukturom.

3.4 SCHL

Da bismo naučili strukturu iz podataka, oslanjamo se na regularizaciju L_1 normom. Postojeći pristupi [87, 2, 21] se oslanjaju na primenu L_1 penala na elemente inverzne kovarijansne matrice. Međutim, pošto se naš metod oslanja na Cholesky parametrizaciju, a u takvoj parametrizaciji inverzna kovarijansna matrica je složena funkcija parametara, direktna primena L_1 norme nije jednostavna. Da bismo izveli odgovarajuće jednačine, počećemo od izraza za standardni L_1 penal na elementima inverzne kovarijansne matrice:

$$h(L) = \lambda ||LL^T||_1 \tag{34}$$

gde je $||LL^T||_1$ suma apsolutnih vrednosti elemenata inverzne kovarijansne matrice:

$$|LL^T|_{ij} = |\sum_{k=1}^{p} L_{ik}L_{kj}^T| = |\sum_{k=1}^{p} L_{ik}L_{jk}| \tag{35}$$

$$h(L) = \lambda \sum_{i=1}^{p}\sum_{j=1}^{i} |\sum_{k=1}^{j} L_{ik}L_{jk}| \tag{36}$$

Nijedan član L_1 penala u jednačini (36) nije konveksan, što čini optimizaciju globalnog cilja (25) teškim, jer nije moguće koristiti standardne alate konveksne optimizacije. Kako bismo zaobišli ovaj problem, predložena je relaksacija regularizujućeg člana:

$$h(L) = \sum_{i=1}^{p}\sum_{j=1}^{i} \lambda_{ij} \sum_{k=1}^{j} |L_{ik}L_{jk}| \tag{37}$$

Ovako relaksiran penal se ponaša nešto drugačije od originalnog L_1 penala na inverznoj kovarijansnoj matrici. Međutim, za dovoljno retke probleme, i ovaj penal će uzrokovati proređenost u inverznoj kovarijansnoj matrici. U nastavku 3.4.5 je dokazana konveksnost ovako relaksiranog penala, za odgovarajuće izabrane parametre λ_{ij}. Bez gubitka generalizacije, pretpostavimo da su vrednosti λ_{ij} identične u

daljem tekstu.

Nova kriterijumska funkcija, sa penalom definisanim u (37), može biti optimizovana pomoću metode koordinatnog spusta. Da bismo definisali jednačine za ažuriranje parametara pri optimizaciji, potrebno je krenuti od izraza za sub-gradijent nediferencijabilnog dela kriterijumske funkcije $h(L)$, i njegove (m, n) komponente:

$$\nabla h(L)_{mn} = \begin{cases} A_{mn}, L_{mn} > 0 \\ -A_{mn}, L_{mn} < 0 \\ \in [-A_{mn}, A_{mn}], L_{mn} = 0 \end{cases} \tag{38}$$

gde je A_{mn} vrednost izvoda po promenljivama L_{mn}, ukoliko postoji.

Izraz A_{mn} se dobija nakon primene standarne algebre:

$$\begin{aligned} A_{mn} &= 2\lambda \sum_{j=1}^{m-1} \frac{\partial \sum_{k=1}^{j} |L_{mk} L_{jk}|}{\partial L_{mn}} + \\ 2\lambda \sum_{i=m+1}^{p} \sum_{j=1}^{i-1} \frac{\partial \sum_{k=1}^{j} |L_{ik} L_{jk}|}{\partial L_{mn}} &+ \lambda \frac{\partial \sum_{k=1}^{p} |L_{mk}^2|}{\partial L_{mn}} \\ = 2\lambda \sum_{j=1}^{m-1} \frac{\partial |L_{mn} L_{jn}|}{\partial L_{mn}} &+ 2\lambda \sum_{i=m+1}^{p} \frac{\partial |L_{in} L_{mn}|}{\partial L_{mn}} + 2\lambda L_{mn} \end{aligned} \tag{39}$$

Uprošćavanjem gornjeg izraza dobija se:

$$A_{mn} = 2\lambda \sum_{i=1}^{p} sgn(L_{mn}) |L_{in}| = 2\lambda sgn(L_{mn}) \sum_{i=1}^{p} |L_{in}|$$

Konačno, izraz za (m, n) komponentu subgradijenta $h(L)$ je:

$$\nabla h(L)_{ij} = \begin{cases} 2\lambda \sum_{k=1}^{p} |L_{kj}|, L_{ij} > 0 \\ -2\lambda \sum_{k=1}^{p} |L_{kj}|, L_{ij} < 0 \\ \in [-2\lambda \sum_{k=1}^{p} |L_{kj}|, 2\lambda \sum_{k=1}^{p} |L_{kj}|], L_{ij} = 0 \end{cases}$$

Subgradijent cele optimizacione funkcije (25) je:

$$\nabla l(L)_{ij} = \begin{cases} \sum_k L_{kj}(S_{ki} + 2sgn(L_{kj})\lambda), L_{ij} > 0 \\ \sum_k L_{kj}(S_{ki} - 2sgn(L_{kj})\lambda), L_{ij} < 0 \\ sgn(Z)max(0, |Z| - 2\lambda \sum_k |L_{kj}|), L_{ij} = 0 \end{cases} \tag{40}$$

gde je $Z = \sum_k L_{kj} S_{ki}$ i $i \neq j$.

Na osnovu ovog moguće je razviti uslove za active-set metod:

$$|\sum_k L_{kj} S_{ki}| < 2\lambda \sum_k |L_{kj}| \tag{41}$$

Ukoliko je $L_{ij} = 0$ i zadovoljen je uslov (41) vrednost parametra L_{ij} će ostati nula. Za subgradijent dijagonalnih elemenata $\nabla l(L)_{ii}$ važi sledeća jednakost (oslanjajući se na činjenicu da L_{ii} mora biti pozitivno):

$$\nabla l(L)_{ii} = \sum_{k=1}^{p} L_{ki}(S_{ki} + 2sgn(L_{ki})\lambda) - \frac{1}{L_{ii}} \tag{42}$$

Uočićemo par veoma korisnih osobina: prvo, da bi ažurirali vrednost parametra L_{ij} potrebno je pristupiti samo elementima iz i-te kolone empirijske kovarijansne matrice; drugo, L_{ij} zavisi samo od vrednosti elemenata j-te kolone matrice L. Stoga, ažuriranja parametara koji pripadaju različitim kolonama L se efektivno mogu odvijati u paraleli.

Za optmizaciju ovako definisane kriterijumske moguće je koristiti koordinatni spust, tako što ćemo rešavati uslov optimalnosti prvog reda, za svaku koordinatu iterativno. Ukoliko je ispunjen uslov (41) i $L_{ij} = 0$, nije potrebno ažurirati parametar L_{ij}, iz čega sledi da se active set pristup prirodno nadovezuje na ovu parametrizaciju i omogućava ubrzanje optimizacije kriterijumske funkcije. U svim ostalim slučajevima, izjednačavanjem (40) i (42) sa nulom moguće je naći jednačine ažuriranja za svaku od koordinata:

$$L_{ij} = -\frac{\sum_{k\neq i} L_{kj}S_{ik} + sgn(L_{ij})2\lambda\sum_{k\neq i}|L_{kj}|}{S_{ii}} \tag{43}$$

$$L_{ii} = \frac{1}{2S_{ii}}\left(-\sum_{k\neq i} L_{ki}S_{ki} - 2\lambda\sum_{k\neq i}|L_{ki}| + \right. \\ \left. + \sqrt{(\sum_{k\neq i} L_{ki}S_{ki} + 2\lambda\sum_{k\neq i}|L_{ki}|)^2 + 4S_{ii}}\right) \tag{44}$$

Kako bismo dodatno unapredili vremensku kompleksnost algoritma za učenje, iskorišćen je active-set metod za odabir parametara koji će se optimizovati, po uzoru na [31]. Polazna ideja ovog metoda je da ukoliko možemo da izdvojimo parametre čiji subgradijent ima vrednostu nula (tj. one za koje važi uslov (41)), njih možemo da preskočimo u trenutnoj iteraciji koordinatnog spusta (jer su te koordinate već u optimalnim tačkama pa dodatna optimizacija ne bi imala nikakvog efekta). Stoga, možemo da particionišemo sve optimizacione promenljive u dva skupa: skup aktivnih i skup fiksnih parametara. Puna iteracija koordinatnog spusta po svim parametrima se svodi na optimizaciju fiksnih parametara za kojom sledi optimizacija aktivnih parametara, pritom ispunjavajući sve neophodne uslove za konvergenciju koordinatnog spusta kao što je već pokazano u [30].

3.4.1 Empirijska evaluacija

Predloženi SCHL metod je upoređen sa savremenim algoritmima za pronalaženje retke inverzne kovarijansne matrice QUIC [30], BCDIC [79] i CSEPNL [32]. U tu svrhu sprovedeno je nekoliko računarskih eksperimenata na sintetičkim i stvarnim skupovima podataka koji su opisani u naredna dva pododeljka. U eksperimentima smo koristili kodove QUIC[2] i BCDIC[3] koje nude njihovi autori. Kod za CSEPNL nije bio dostupan, pa smo implementirali pristup zasnovan na detaljima datim

[2]http://www.cs.utexas.edu/ sustik/QUIC/
[3]https://github.com/erantreister/Multilevel-BCDIC.m

Tabela 3: Poređenje vremena izvršavanja SCHL, QUIC, BCDIC, i CSEPNL methode za problem učenja sintetičke proređene inverzne kovarijansne matrice. Sintetički retki grafovi koji su korišćeni u eksperimentima imaju između 1,000 i 20,000 čvorova, dok je broj odbiraka za učenje bio 1,000 .

problem		SCHL		QUIC	
size	nnz	time	nnz	time	nnz
1,000	35,234	1.5	35,372	2.9	35,324
5,000	178,570	24.1	173,542	116.6	175,000
10,000	358,174	100.8	338,274	526.1	387,102
15,000	733,368	450.2	745,792	1,699.5	688,936
20,000	979,534	1,237.0	1,021,962	4,600.4	1,053,550

problem		BCDIC		CSEPNL	
size	nnz	time	nnz	time	nnz
1,000	35,234	7.0	35,324	14.1	35,480
5,000	178,570	179.4	187,478	567.1	169,774
10,000	358,174	920.2	386,898	3,311.9	352,526
15,000	733,368	3,381.7	734,914	9,540.2	718,432
20,000	979,534	10,390.8	1,031,074	25,561.3	1,116,954

u [32]. Ne upoređujemo se sa Big&QUIC (koji je unapređenje QUIC modela) jer je taj model pogodniji za velike probleme, dok su za probleme koji mogu stati u memoriju (kakvi su i problemi u sprovedenim eksperimentima) autori samog modela preporučili QUIC.

3.4.2 Sintetički podaci

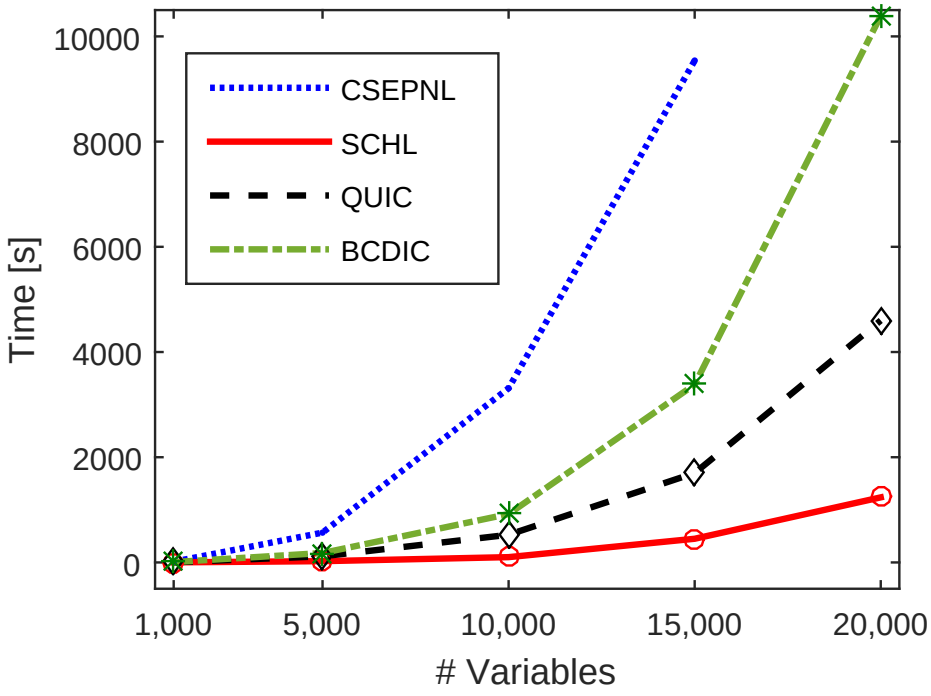
Ponovili smo eksperimentalnu postavku korišćenu u [30] kako bismo napravili procenu komputacione efikasnosti. Konstruisani su grafovi susednosti, tj proređene simetrične matrice sa nasumično dodeljenim elementima izvan glavne dijagonale. Pozitivna definitnost matrice je obezbeđena tako što je dodata dijagonalna matrica čime je obezbeđeno da konačna matrica bude dijagonalno dominantna. Zatim je ta matrica korišćena za generisanje odbiraka iz multivarijatne normalne raspodele. Generisano je nekoliko problema različitih veličina koristeći više različitih struktura slučajnog grafa. Generisani grafovi su imali od 1000 do 20,000 čvorova (varijabli) kako bi se uporedili efikasnost i skalabilnost algoritama za učenje inverzne

Tabela 4: Poređenje kvaliteta naučene strukture pomoću SCHL i tri druge metode za učenje proređene inverzne kovarijansne matrice. Korišćenje pet različitih metrika (Jaccard, R^2, precision, recall, accuracy) potvrđuje da su sve četiri metode naučile strukturu uporedivog kvaliteta.

algorithm	SCHL	QUIC	BCDIC	CSEPNL
Jaccard Index	0.80784	0.76987	0.76933	0.77594
R^2	0.55633	0.6611	0.66104	0.70376
Precision	0.89523	0.87309	0.87238	0.88162
Recall	0.89219	0.86689	0.86689	0.86619
Accuracy	0.99998	0.99997	0.99997	0.99997
Lambda	0.4	0.4125	0.4125	0.347
nnz - true	29998	29998	29998	29998
nnz - est	29930	29856	29872	29648

kovarijansne matrice. Gustina grafa je odabrana tako da srednji stepen čvorova bude u intervalu 15-20. U svakom eksperimentu stvorena je empirijska matrica kovarijanse od 1000 uzoraka generisanih iz multivarijatne normalne distribucije sa nultim očekivanjem i kovarijansom koja odgovara (inverznoj) generisanoj matrici susedstva. Svi eksperimenti su sprovedeni u single-threaded okruženu na Intel (R) Core (TM) i7-4770 CPU @ 3.40GHz procesoru sa 32 GB RAM-a. Rezultati se mogu videti na slici 7, a dodatni detalji su navedeni u tabeli 3. Kako bismo evaluirali algoritme, podesili smo hiperparametar kojim se podešava jačina penala λ za svaki algoritam posebno tako da svaki pristup nauči strukturu sa sličnim brojem nenultih elemenata ("nnz-est") kao i originalan sintetički graf ("nnz-true"). Rezultati pružaju dokaze da je naš SCHL algoritam najbrži, i da se bolje skalira sa povećanjem broja varijabli. Prema literaturi [79], BCDIC se očekuje da bude brži od QUIC-a, međutim, u našim eksperimentima je sporiji, verovatno zato što su naši problemi veće gustine.

S obzirom da poznajemo generativni proces koji generiše sintetičke opservacije, moguće je uporediti koliko dobro algoritmi otkrivaju pravu strukturu. Kvalitet dobijenih rešenja smo procenili upoređivanjem naučene matrice sa matricom susedstva na osnovu koje su generisani podaci. Korišćene su metrike: Jaccard index, precision, recall, accuracy za procenu preklapanja samih veza u grafu, tj preklapanja nenultih



Slika 7: Poređenje vremena izvršavanja SCHL, QUIC, BCDIC i CSEPNL metode učenja retke inverzne kovarijansne matrice na sintetičkim podacima. Sintetički generisani grafovi pomoću kojih je generisano 1000 odbiraka imaju od 1000 do 20,000 čvorova.

Tabela 5: Poređenje potrebnog vremena za učenja grafa na podacima ekspresije gena.

	SCHL	QUIC	BCDIC	CSEPNL
time	5,038.6	10,011.0	15,929.1	26,665.7
nnz	366,064	346,536	346,516	357,812

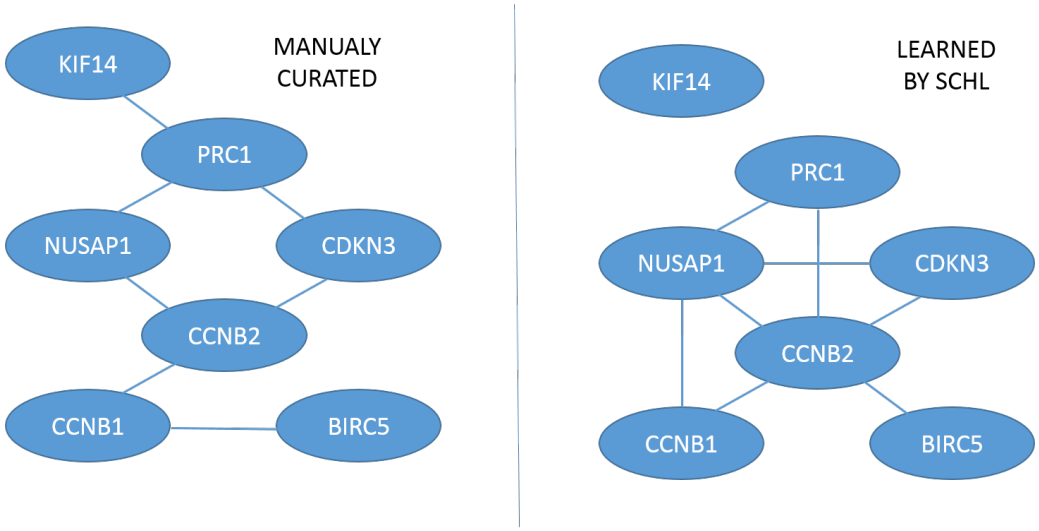
elemenata u matricama, i R^2 za procenu koliko se magnitude veza preklapaju. U ovom eksperimentu generisan je jedan problem veličine 10,000 sa osnovnom strukturom uvezanog linearnog lanca, iz koje smo uzorkovali 1000 odbiraka na kojima smo estimirali inverznu kovarijansnu matricu za svaki od algoritama. Rezultati su navedeni u Tabeli 4 i može se videti da su svi pristupi proizveli rešenja uporedivog kvaliteta. Proređenosti rezultata svake od metoda (nnz-est) su nezavisno prilagođene tako da približno odgovaraju stvarnoj vrednosti (nnz - true).

3.4.3 Aplikacije na realnim podacima

Napredak u tehnologiji merenja doveo je do izobilja količine podataka, kao i do povećanja same dimenzionalnosti podataka, povećavajući potrebu za automatskim i nenadziranim otkrivanjem zavisnosti između varijabli. Ovo podstiče primenu pristupa učenja kao što je SCHL na podacima iz različitih domena kao što su biomedicina, obrada slike, ekologija, i drugi, radi otkrivanja prethodno nepoznatih veza i regularnosti. U nastavku predstavljamo primenu naše SCHL metode za otkrivanje strukture povezanosti na skupu realnih podataka iz biomedicinskog domena.

Ekspresija gena kod pacijenata sa dijagnostikovanom sepsom

U ovom eksperimentu koristili smo skup realnih podataka koji sadrži profile ekspresije gena koji su sakupljeni od 53 subjekta na dnevnom nivou u vremenskom intervalu od 5 dana [51]. Podaci se sastoje od ukupno 163 uzoraka od pacijenata koji su primljeni sa dijagnozom sepse u odeljenja intenzivne nege u bolnicama u Sidneju, Australija. Merenja se sastoje od nivoa ekspresije 24,840 gena, a dobijeni nivoi ekspresije gena su kvantilno normalizovani i logaritmovani. Primenili smo naš SCHL pristup na svih 163 uzorka kako bismo otkrili strukturu zavisnosti između



Slika 8: Poređenje ručno pronađene mreže gena koji utiču na komplikacije sepse, i mreža dobijena učenjem GMRF iz ekspresije gena pomoću SCHL algoritma.

nivoa ekspresije gena pacijenata obolelih od sepse, i SCHL je najbrže konvergirao od sva četiri korišćena algoritama (Tabela 5). U sprovedenom eksperimentu smo postavili hiperparametar regularizacije na adekvatno visok nivo i dobili proređen graf koekspresija sa 170,612 veza (od mogućih 308,500,380). Da bismo se uverili u verodostojnost dobijene mreže ko-ekspresija, pogledali smo objavljenu literaturu na primeru mreže gena relevantnih za stanje sepsa. Uzeli smo mrežu od 7 gena povezanih sa 7 linkova [82], koji su poznati u literaturi i koji su bili povezani sa sindromom akutnog respiratornog stresa prouzrokovanog sepsom. U mreži koja je izvedena pomoću našeg SCHL pristupa, podmreža ovih 7 gena sadrži 8 ivica. Ove mreže su prikazane na slici 8, i može se videti da se 4 veze preklapaju, što je značajan rezultat uzimajući u obzir visoku proređenost mreže ko-ekspresija. Ovi rezultati pružaju dokaze da se SCHL može koristiti za izgradnju ko-ekspresionih mreža, koje su od velikog biološkog interesa jer ko-izraženi geni su funkcionalno povezani, i često uključeni u iste puteve [74] ili kontrolisani istim regulatornim procesima [69].

3.4.4 Diskusija i ideje za budući rad

U ovom poglavlju, predstavljen je SCHL, novi pristup učenju retkih GMRF modela. SCHL pristup koristi reparametrizaciju zasnovanu na Cholesky dekompoziciji. Takođe smo uveli novi član za selekciju mreže, zasnovan na L_1 normi, koji penalizuje aproksimiran proizvod Cholesky faktora. Ovako definisan model je konveksan i može biti efikasno optimizovan koristeći koordinatni spust. Dodatno ubrzanje je postignuto korišćenjem active-set pristupa, koji smanjuje broj potrebnih računskih operacija tako što se fokusira samo na podskup aktivnih parametara. Pokazano je i da algoritam optimizacije novog modela pripada klasi tzv. "embarrassingly parallel" algoritama, i da se može lako skalirati na multi-threaded arhitekturama. U nekoliko experimenata SCHL se pokazao brži od drugih od sada najboljih metoda (QUIC, BCDIC, i CSEPNL) za estimaciju retke inverzne kovarijansne matrice. Takođe je demonstrirana primenljivost modela u značajnoj aplikaciji koja uključuje ekspresije gena u pacijentima obolelim od sepse, i njegova sposobnost za otkrivanje strukturalnih veza iz podataka u ovakvim aplikacijama.

3.4.5 Analiza vremenske kompleksnosti

Algorithm 1 CoordinateDescent

```
1:  procedure SparseChol
2:      A_S ← ∀ (i,j), |S_ij| < λ
3:      A_J ← 1..N
4:      main loop:
5:      for j ∈ A_J do
6:          for i ∈ A_s(J) do
7:              L_ij ← min(f(l, A_s, λ))          ▷ Eqs. (43) & (44)
8:              if |∇L_*j| < ϵ then
9:                  A_J ← A_J \ {j}
10:     if A_j = {∅} then return
11:     else
12:         goto main loop.
```

Vremenska kompleksnost algoritma za obučavanje se sastoji iz dva dela: pretprocesorski korak tokom kojeg se računa empirijska kovarijansna matrica i njeno

zaokruživanje malih elemenata na nulu čime se vrši inicijalni odabir aktivnog skupa, drugi deo koji čini optimizacija korišćenjem koordinatnog spusta. Odabir inicijalnog aktivnog skupa košta $O(p^2)$ jer je neophodno izračunati celu kovarijansnu matricu. Optimizacija se sastoji iz tri ugnježdene petlje (algoritam 1). Spoljašnja petlja iterira sve dok se ne postigne konvergencija, i stoga ima kompleksnost $O(T)$, gde je T broj iteracija neophodnih da se zadovolji uslov konvergencije. Kao što je već napomenuto, iz jednačina (43) i (44), koordinatni spust se radi nezavisno po kolonama matrice L, odakle sledi da je T zapravo supremum broja iteracija svake od grupa promenljivih potrebnih da se postigne konvergencija.

Središnja petlja iterira kroz sve aktivne kolone, što je inicijalno ceo set, pa ima kompleksnost $O(P)$. Ova petlja se može efikasno paralelizovati, jer se promenljive svake kolone računaju nezavisno. Unutrašnja petlja iterira kroz aktivni skup za kolonu j, i ima vremensku kompleksnost $O(A_s)$, gde A_s odgovara gustini kolone j, tj broju nenula. Ukupna kompleksnost je $O(N^2 + NA_sT)$. Pseudokod celog algoritma koordinatnog spusta sa aktivnim setom je prikazan algoritmom 1.

U nastavku ćemo pružiti dokaz da je kriterijumska funkcija sa penalom definisanim u (37) konveksna.

Teorema 1. *Sledeća funkcija je konveksna:*

$$h(L) = \sum_i \sum_j \lambda_{ij} \sum_k |L_{ik} L_{jk}| = tr(\Lambda |L| |L|^T) \tag{45}$$

gde je $|L| = [\ |L_{ij}|\ ]$ *matrica čiji elementi su apsolutne vrednosti elemenata matrice* L.

Dokaz. Iako pojedinačni članovi stvarno nisu konveksni, dokazaćemo da je penal (45) u celini konveksan za pozitivnu semidefinitnu matricu hiperparametara Λ.

Najpre, posmatrajmo diferencijabilne regione funkcije $h(L)$ gde važi:

$$L_{ij} \neq 0, \forall (i,j) \tag{46}$$

Pošto je Λ matrica pozitivno semidefinitna, može se dekomponovati na Cholesky faktore: $\Lambda = VV^T$, i zatim se korišćenjem standarde algebre dobija:

$$h(L) = tr(VV^T|L||L^T|) = tr(V^T|L||L^T|V) = ||X||^2 \tag{47}$$

što predstavlja kvadriranu Frobenius normu matrice $V^T|L|$ i, stoga, funkcija je konveksna unutar svakog diferencijabilnog regiona (jer se znak ne menja u okviru regiona).

Zatim, dva susedna diferencijabilna regiona su razdvojena nediferencijabilnim regionom definisanim kada je jedan od parametara L_{ij} nula. Lako se može pokazati da je levi izvod $\lim_{L_{ij}\to -0} \frac{\partial h(L_{ij})}{\partial L_{ij}}$ uvek manji ili jednak desnom izvodu $\lim_{L_{ij}\to +0} \frac{\partial h(L_{ij})}{\partial L_{ij}}$, usled čega je ceo izraz (45) konveksan.

□

U eksperimentalnim primerima, korišćena je pozitivna skalarna veličina lambda (koja je eqvivalentna skaliranoj matrici jedinica koja je PSD).

3.5 SNETCH

Jedna od interesantnih strukturnih osobina koja se često pojavljuje u kompleksnim sistemima je exponencijalna distribucija stepena čvorova mreže, svojstvo koje imaju tzv. scale-free mreže. Scale-free grafovi poseduju nekoliko interesantnih osobina, jedno od njih je ta da njihovi Cholesky faktori mogu da se naprave proređenim (uz pravilno odabran redosled čvorova u matrici susednosti). U sekciji empirijske evaluacije SNETCH modela, pokazaćemo da je pronalaženje grafa pomoću L_1 penala na Cholesky faktoru pogodno za primenu na probleme sa Scale-Free strukturom. Koristeći sintetički primer, pokazaćemo da naš metod pouzdanije pronalazi strukturu u odnosu na do sada najbolje pristupe QUIC [30] i BCDIC [79], koji optimizuju stan-

dardnu kriterijumsku funkciju sa penalom na elementima inverzne kovarijansne matrice $\|\Sigma^{-1}\|_1$. Takođe, empirijski će biti potvrđena brzina i skalabilnost predloženog rešenja na velikim scale-free problemima od preko 200,000 promenljivih. Vreme izvršavanja sugeriše da je naš pristup brži za više nego red veličine u odnosu na dva do sada najbrža metoda primenjiva na velike probleme, Big & QUIC [31] i ML-BCDIC [80]. I na kraju, pokazaćemo efikasnost i ubrzanje koje se dobija paralelizacijom predloženog algoritma.

U nastavku biće prikazan novi optimizacioni algoritam za objektiv predložen u CSEPNL [32]. Takođe, biće pažljivo analizirana postavka modela i biće pokazano da je model pogodan za koordinatni spust, kao i da omogućava korišćenje active-set pristupa, što dalje omogućava jako efikasno učenje. Takođe, biće pokazano kako razvijeni optimizacioni algoritam pripada klasi "Embarrassingly parallel" algoritama, jer je optimizaciju moguće raščlaniti po kolonama Cholesky faktora i rešavati potpuno nezavisno. Kod grafova Erdos-Renyi tipa, broj nenula u Choelsky faktoru raste nelinearno sa povećanjem dimenzije grafa, što onemogućava primenu našeg metoda na ogromnim retkim grafovima. Međutim, kao što će se pokazati, efekat popunjenosti Cholesky faktora kod visokodimenzionalnih problema nije isti kod svih tipova grafova. Poznato je da scale-free grafovi, uz odgovarajuće sortiranje čvorova, imaju retke Cholesky faktore. Ovo je jedan od glavnih motiva za razvijanje našeg SNETCH pristupa, utoliko što reparametrizacija problema estimacije inverzne kovarijansne matrice kroz dekompoziciju u Cholesky faktore poseduje nekoliko korisnih osobina, kao što je predstavljeno u prethodnim poglavljima (jedna od njih je brzo računanje log det člana u $O(p)$ operacija, kao što je i brza provera pozitivne definitnosti kovarijansne matrice). Na kraju, biće pokazano kroz nekoliko primera da čak i kada struktura problema nije scale-free, metod je i dalje primenljiv i rezultati su uporedivi sa ostalim algoritmima.

Scale-free Postoji nekoliko radova koji se bave učenjem strukture scale-free grafova na osnovu kontinualnih podataka, uglavnom kroz primenu penala koji forsira eksponencijalnu raspodelu stepeni čvorova na naučenom grafu [13]. U modelu predloženom u [61] postavljen je prior u vidu scale-free strukture i razvijen je metod odabiranja baziran na Metropolis-Hastings algoritmu za učenje parametara. U modelu [76] predložena je i metoda zasnovana na rangiranju u kojem se dinamički procenjuje stepen čvora, uz novu optimizaciju zasnovanu na alternating direction method of multipliers [9]. Takođe, prikazan je još jedan pristup zasnovan na regularizaciji stepena čvorova sa algoritmom minorize-maximize [44]. Iako svi pristupi [13, 44, 61, 76] otkrivaju scale-free strukturu, metode koje primenjuju regularizaciju stepena čvorova u grafu, i forsiraju power-law distribuciju još uvek nisu dovoljno efikasne i ne skaliraju se dovoljno dobro kao metode za opštu estimaciju retke inverzne kovarijansne matrice. Ovde ne tvrdimo da naša formulacija u svom rešenju forsira scale-free strukturu, već tvrdimo da su problemi sa scale-free strukturom vrlo pogodni za naš pristup. Zbog toga, pristupi koji nameću scale-free strukturu koristeći različite tehnike regularizacije su van okvira ovog rada.

Usvojili smo parametrizaciju zasnovanu na Choleksy dekompoziciji L slično kao i za metodu SCHL, i zamenili smo skalarnu vrednost λ ne-negativnom matricom Λ (kao što je to urađeno u [8]), where $*$ is Hadamard product:

$$f(L) = \frac{1}{2} tr\left(X^T L L^T X\right) - \frac{n}{2} log|LL^T| + \|\Lambda * L\|_1 \tag{48}$$

Diferencijabilni deo kriterijumske funkcije je već analiziran u prethodnim poglavljima (27). Sada razmatramo regularizacioni član. Kao i za SCHL model, koordinatni spust sa active-set pristupom se prirodno nameće kao atraktivan metod, jer je regularizacioni član (kao i diferencijabilni) separabilan.

Da bismo izveli izraze za ažuriranje parametara, najpre posmatramo sub-diferencijal $h(L)$, a zatim i njegovu (i, j) komponentu.

$$\nabla h(L)_{ij} = \begin{cases} \lambda_{ij}, L_{ij} > 0 \\ -\lambda_{ij}, L_{ij} < 0 \\ \in [-\lambda_{ij}, \lambda_{ij}], L_{ij} = 0 \end{cases} \tag{49}$$

Subgradijent cele optimizacije (48) je:

$$\nabla f(L)_{ij} = \begin{cases} \sum_k L_{kj} S_{ki} + sgn(L_{ij})\lambda_{ij}, L_{ij} \neq 0, i \neq j \\ \sum_k L_{kj} S_{kj} - \frac{1}{L_{ii}} + \lambda_{ij}, i = j \\ \mathrm{S}(\sum_k L_{kj} S_{ki}, \lambda_{ij}), L_{ij} = 0, i \neq j \end{cases} \tag{50}$$

gde je S(z,r) soft thresholding funkcija:

$$\mathrm{S}(z, r) = sign(z) max(|z| - r, 0) \tag{51}$$

Sada kada imamo izraze za sub-gradijent moguće je optimizovati cilj (48) koristeći koordinatni spust. Izrazi za ažuriranje pojedinačnih parametara mogu se dobiti pomoću kriterijuma optimalnosti (50). Opet, najpre ćemo izvesti izraz za ažuriranje jedne promenljive koje se nalazi izvan glavne dijagonale (51) za koju dobijamo da je prosto rešenje linearne jednačine sa soft-threshold funkcijom.

$$L_{ij}^{t+1} = \begin{cases} -\frac{\sum_{k \neq i} L_{kj}^t S_{ik} + sign(L_{ij}^t)\lambda_{ij}}{S_{jj}}, L_{ij}^t \neq 0 \\ -\frac{\sum_{k \neq i} L_{kj}^t S_{ik} - sign(M)\lambda_{ij}}{S_{jj}}, L_{ij}^t = 0, M \geq \lambda_{ij} \\ 0, L_{ij}^t = 0, |M| < \lambda_{ij} \end{cases} \tag{52}$$

gde je $M = \sum_{k \neq i} L_{kj}^t S_{ik}$.

Ako je L dijagonalna matrica (što je slučaj u prvoj iteraciji, nakon što inicijalizujemo algoritam pomoću identitet matrice), ažuriranje svakog od parametara je moguće izvršiti u $O(1)$ vreme, rezultat analogan onome iz [30]):

$$L_{ij}^1 = \begin{cases} -\frac{S_{ij} - sign(S_{ij})\lambda_{ij}}{S_{jj}}, |S_{ij}| > \lambda_{ij} \\ 0, |S_{ij}| < \lambda_{ij} \end{cases} \tag{53}$$

Početni skup nenultih vrednosti parametara može biti konstruisan na osnovu pravila za ažuriranje parametara (53). Ovaj skup čini početni aktivni set, i njegovo računanje zahteva $O(p^2)$ operacija, i ujedno predstavlja najskuplji deo algoritma.

Izrazi za ažuriranje dijagonalnih parametara (parametara na dijagonali L matrice) su:

$$L_{ii}^{t+1} = \frac{-(\sum_{k \neq i} L_{ki}^t S_{ik} + \lambda_{ii}) + \sqrt{(\sum_{k \neq i} L_{ki}^t S_{ik} + \lambda_{ii})^2 + 4S_{ii}}}{2S_{ii}} \tag{54}$$

Drugo rešenje kvadratne jednačine se uvek odbacuje, jer L_{ii} mora biti pozitivno kako bi matrica Σ^{-1} bila pozitivno definitna. Posmatrajući jednačine za ažuriranje (52) i (54) uočavamo da ako je L_j retka matrica sa prostornom kompleksnošću $O(s_j)$, ažuriranje svakog parametra će imati $O(s_j)$ vremensku kompleksnost.

Algorithm 2 SNETCH algorithm

```
procedure optimizeL(L, Y, λ, ϵ)
    for each column of L do
        S_active ← initActiveSet(Y, λ)
        while |∇L_j| < ϵ do
            for i ∈ S_active do
                L_ij ← argmin(f_ij(L_j, λ))          ▷ Defined in (52), (54)
            t ← λbound(L_j)
            if t < λ then
                S_active ← initActiveSet(Y, t)
    return L
```

Izrazi za ažuriranje (52) i (54) predstavljaju korake koordinatnog spusta, za Scale-free Networks Estimation Through Cholesky factorization (SNETCH) algoritam opisan u pseudokodu 2. Nakon inicijalizacije identitet matricom, koordinate se iterativno ažuriraju sve dok se ne zadovolji kriterijum konvergencije, tj. promena magnitude gradijenta padne ispod prage ϵ. Konačan graf, tj. inverzna kovarijansna matrica se dobija proizvodom optimalnih Cholesky faktora $\Sigma_*^{-1} = L_*^T L_*$.

3.5.1 Interpretacija modela

Pronalaženje retke Cholesky matrice ne vodi nužno do retke inverzne kovarijansne matrice, stoga je neophodno dodatno pojašnjenje kako bi se opravdao cilj (48) kao i analiza širokog skupa problema kako bi se identifikovalo na kojima je SNETCH metod primenjiv. Najpre ćemo uočiti dve poznate činjenice o Čolesky faktorizaciji. Prvo, nalaženje obrasca popunjenosti Cholesky faktora odgovara nalaženju Chordal completion inverzne kovarijansne matrice. Drugo, retkost Cholesky faktora u velikoj meri zavisi od redosleda obilaska čvorova u grafu. S obzirom da penal kriterijumske funkcije (48) pokušava da nađe proređen chordal graf koji upotpunjava matricu inverzne kovarijanse (za izabrano sortiranje čvorova), ova vrsta optimizacije u velikoj meri zavisi od tipa strukture skrivene u podacima. Sugerisano je još ranije [26, 64] da se scale-free grafovi prirodno uklapaju u ovakav pristup jer ih je lako trijangulirati. Učenje scale-free GMRF-ova predstavlja važan, izazovan, i još uvek nedovoljno istražen problem, koji je centralan deo istraživanja, prikazan u ovom poglavlju.

Jedan od najvećih izazova u ovakvom pristupu je pronalaženje najboljeg redosleda prilikom numeracije čvorova (tj pronalaženje permutacije redova matrice susednosti koja daje najređi Cholesky faktor). Poznato je da je ovo NP-kompletan problem [22]. Nekoliko heuristika postoji za rešavanje ovog problema. Prilikom eksperimentalne evaluacije, koristili smo Approximate Minimum Degree algoritam kao međukorak prilikom optimizacije, što je uslovilo dobre rezultate.

3.5.2 Empirijska evaluacija

Da bi se empirijski testirale karakteristike predloženog pristupa i njegovog rešenja, izvršili smo obimne procene sintetički generisanih problema od interesa. Na datim primerima, upoređujemo naš SNETCH sa dva najsavremenijih pristupa za učenje proredjenih matrica srednje veličine (tj. kakda svi podaci mogu da stanu u memoriju) QUIC [30] i BCDIC [79].

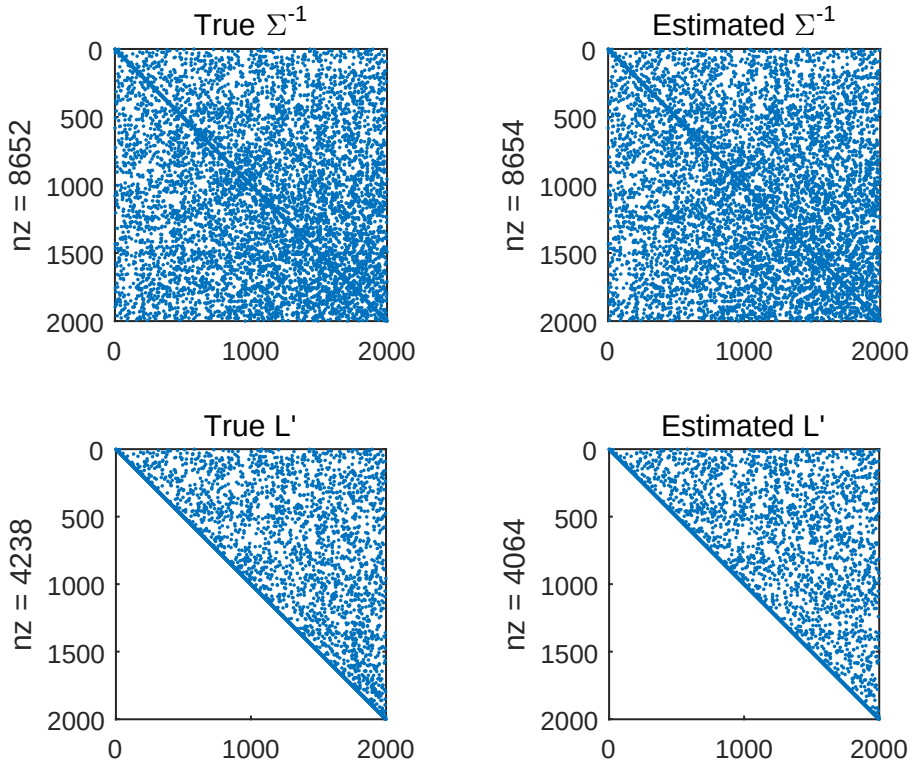
Uzorci iz željene distribucije su generisani prvo projektovanjem odgovarajuće

retke inverzne kovarijansne matrice, u kojoj je kodirana struktura od interesa. Nakon toga, uzorci su izvučeni iz MVN distribucije sa nultom srednjom vrednošću i dizajniranom strukturom povezanosti. U našim eksperimentima simulirano je nekoliko tipova grafova, od kojih svaki predstavlja drugačiju vrstu izazova. Kao i u pristupima evaluacije opisanim u [30] i [79], napravili smo linearni lanac i slučajni binarni graf. Takođe, analizirali smo još dva sintetička problema od interesa, grafik sa retkim Cholesky faktorima ("idealna" pretpostavka za naš model) i scale-free grafik, generisan korišćenjem procesa preferencijalnog povezivanja.

Počećemo sa generisanjem sintetičkog primera zasnovanog na našoj pretpostavci za modeliranje, tj. Choleski faktor inverzne kovarijansne matrice je proredjen. Slučajna proredjena Choleski struktura grafa je dobijena slučajnim dodeljivanjem odredjenog broja ne-nultih pozicija elemenata u Choleski matrici, a potom je dobijena retka inverzna kovarijansna matrica kao proizvod retkih Choleski faktora. Proredjenost se kvantifikuje sa brojem ne-nultih (“ nz ” u Slikama i Tabelama) elemenata matrice.

Na slici 9, sa leve strane vidimo formu originalne (prave) inverzne kovarijanse, a desno, matricu procenjenu našim pristupom. Na dnu slike su odgovarajući, dizajnirani i estimirani, Choleski faktori. U Tabeli 6 prikazani su rezultati sva tri konkurentna algoritma i može se videti da naš SNETCH pristup postiže superiorne performanse u smislu brzine i kvaliteta dobijenog rešenja. Retkost procenjenih rešenja podešena je pomoću hiperparametra λ da bi se dobio broj nz elemenata blizu originalnog - pravog nz-a, ali i što je još važnije koje je približno konkurentskim pristupima, radi fer upoređivanja.

Kvalitet rešenja smo merili pomoću Jaccard indeksa, metrika precision i recall, i sve tri sugerišu da je naš pristup otkrio prave nulte elemente mnogo bolje od konkurencije (Tabela 6). Međutim, u drugim radovima, pristupi su obično testirani pod različitim strukturnim pretpostavkama. Stoga smo upoređivali performanse na najčešćim sintetičkim primerima: linearno ulančani graf i slučajno raspodeljena



Slika 9: Sintetički primeri dobijeni pomoću slučajno semplovanog proređenog Cholesky faktora

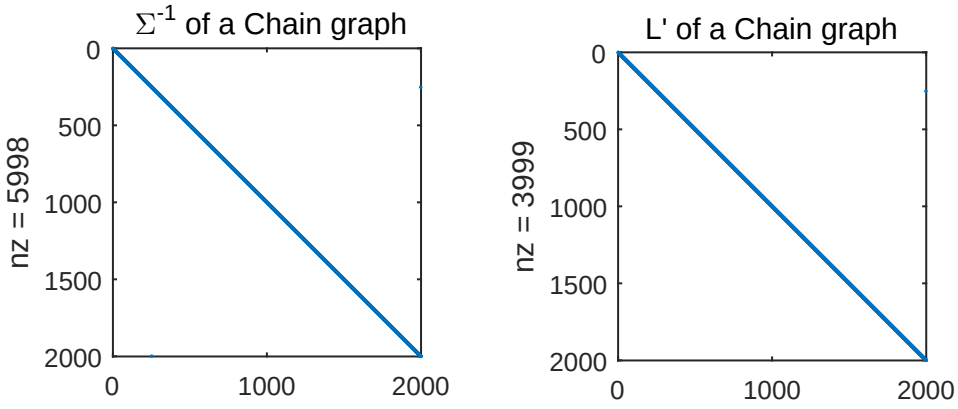
Tabela 6: Poređenje performansi učenja grafa na problemu proređenog Cholesky faktora.

Random Cholesky Factor (p=2,000; n=2,000; true nz = 8,652)						
Model	Time [s]	Jaccard	Prec.	Recall	λ	nz
SNETCH	1.21	0.745	0.854	0.854	0.0818	8,654
BCDIC	9.04	0.500	0.667	0.667	0.0878	8,654
QUIC	20.2	0.501	0.668	0.668	0.0878	8,654

retka matrica.

Lanac graf je kreiran određivanjem odgovarajućih obrazaca veza u matrici susednosti. Matrica susednosti je kasnije pretvorena u Laplacian i dijagonalna dominantnost je obezbeđena dodavanjem male vrednosti na dijagonalu. Dobijena matrica je uzeta za originalnu inverznu kovarijansnu matricu za "bojenje" uzoraka iz obične MVN distribucije.

Na slici 10 pokazan je primer (linearnog) lanca koji ima ekstremno retku matricu i odgovarajući retki Choleski faktor, a u tabeli 7 rezultati pokazuju da su svi pristupi savršeno rekonstruisali originalnu strukturu, ali SNETCH je to učinio mnogo brže.



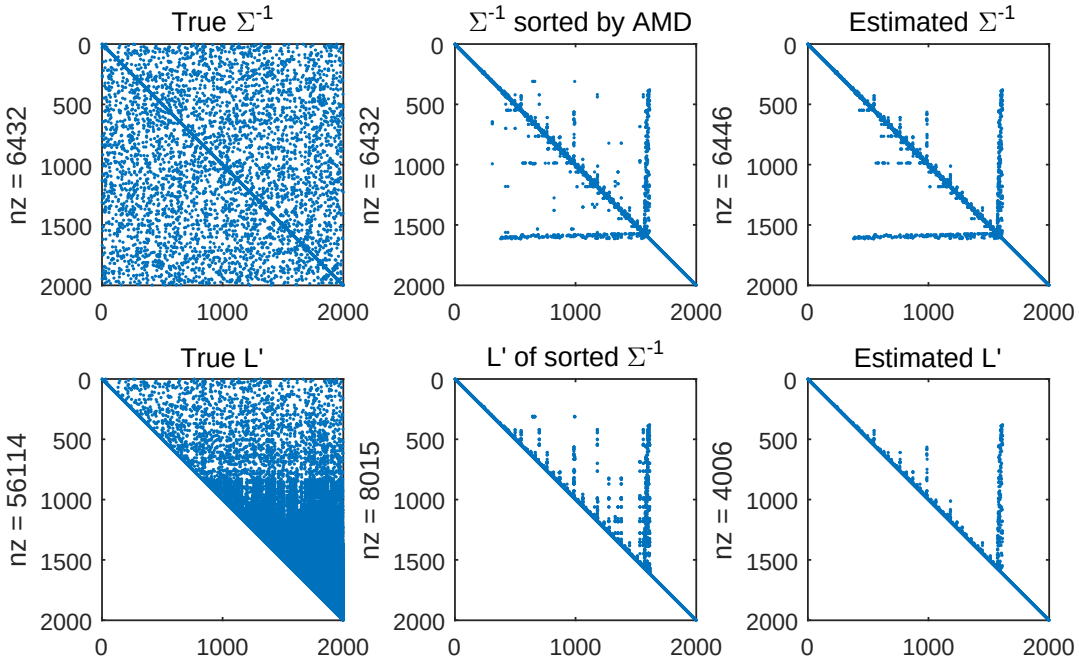
Slika 10: Sintetički primer koji ilustruje proređenu strukturu linearnog lanca.

Tabela 7: Poređenje performansi učenja grafa na problemu linearnog lanca.

Linear Chain Structure (p=2,000; n=2,000; true nz = 5,998)						
Model	Time [s]	Jaccard	Prec.	Recall	λ	nz
SNETCH	0.69	1	1	1	0.2	5,998
BCDIC	5.45	1	1	1	0.2	5,998
QUIC	3.39	1	1	1	0.2	5,998

Slučajna inverzna kovarijansna matrica je dobijena dodeljivanjem određenog broja ne-nultih elemenata u matrici. Matrica je učinjena simetričnom dodavanjem sopstvenog transponovanja, a pozitivna definitnost je osigurana dodavanjem male

pozitivne konstante na dijagonalu. Ovo je najčešći način generisanja sintetičkih primera, a primer je prikazan na Slici 11.



Slika 11: Sintetički primer dobije generisanjem proređene slučajne inverzne kovarijansne matrice.

Međutim, Cholesky faktor zavisi od redosleda varijabli (redova i kolona), i može se napraviti retkim (do određenog stepena) odgovarajućom izmenom redosleda. Jedan poznati heuristički pristup je sortiranje Približno minimalnim stepenom (AMD), što ima za cilj da dovede do retkih Choleski faktora. Takvo svojstvo je vizuelizovano na srednjim delovima Slike 11, gde je originalna matrica sortirana pomoću AMD. Nakon preuređivanja, naš pristup je mogao rekonstruisati preciznost brže ali nešto manje tačno (za datu veličinu uzorka) od konkurentnih pristupa kao što je i prikazano u Tabeli 8.

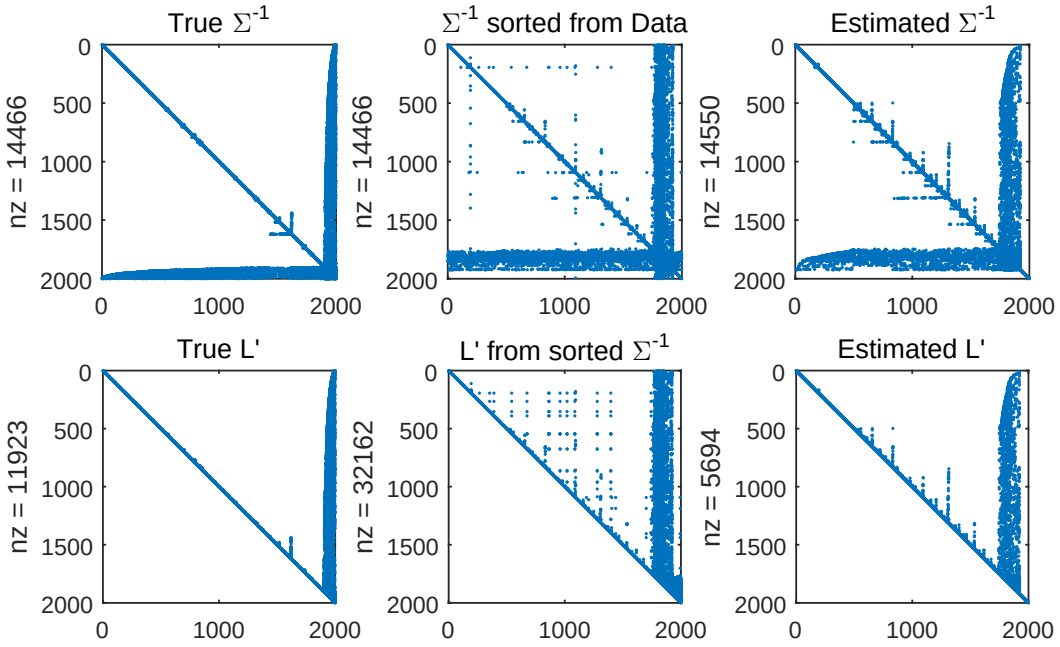
Ipak, u praksi, najinteresantnije strukture se nalaze negde između dva ekstrema slučajnog (bez strukture) i linearnog lanca (totalnog reda). Stoga ćemo se nadalje usredsrediti na takve relevantne strukture.

Tabela 8: Poređenje performansi učenja grafa na problemu slučajne matrice.

Random Precision Matrix (p=2,000; n=2,000; true nz = 6,432)						
Model	Time [s]	Jaccard	Prec.	Recall	λ	nz
SNETCH	0.62	0.840	0.911	0.914	0.162	6,446
BCDIC	7.78	0.986	0.992	0.994	0.11	6,440
QUIC	3.24	0.986	0.991	0.994	0.11	6,444

Scale-free mreže Scale-free mreže imaju vrlo intrigantne karakteristike [6, 73], a mogu se naći u brojnim realnim fenomenima, od mreža za citiranje radova do internet mreže i mreža koje regulišu procese u živim ćelijama [5]. Stoga smo ocenili prikladnost našeg pristupa za otkrivanje scale-free struktura.

Generisali smo scale-free probleme procesom preferencijalnog vezivanja sa Poasonovim rastom [62]. Graf se započinje sa nekoliko (unapred određenih) grupa gusto povezanih čvorova (klika), a svaki novi dolazni čvor je dodat grafu koji se povezuje sa k postojećih, gde je k izvučeno iz Poissonove distribucije (na osnovu verovatnoće proporcionalne stepenu čvora).



Slika 12: Sintetički primer koji ilustruje strukturu koja poseduje Scale-Free svojstvo

Primer inverzne kovarijansne matrice sa scale-free osobinama je prikazan na

slici 12. Na levoj strani je originalna matrica, a njen Choleski faktor je proređen. Tipično, u aplikacijama varijable nisu sortirane, i srednji delovi slike pokazuju kako se mogu adekvatno sortirati sortiranjem početnog aktivnog skupa dobijenog iz posmatranih podataka. S desne strane je struktura procenjena našim pristupom, a rezultati u Tabeli 9 ponovo ukazuju na to da naš pristup rekonstruiše osnovne zavisnosti bolje i brže od alternativa.

Tabela 9: Poređenje performansi učenja grafa na problemu sa Scale-Free strukturom.

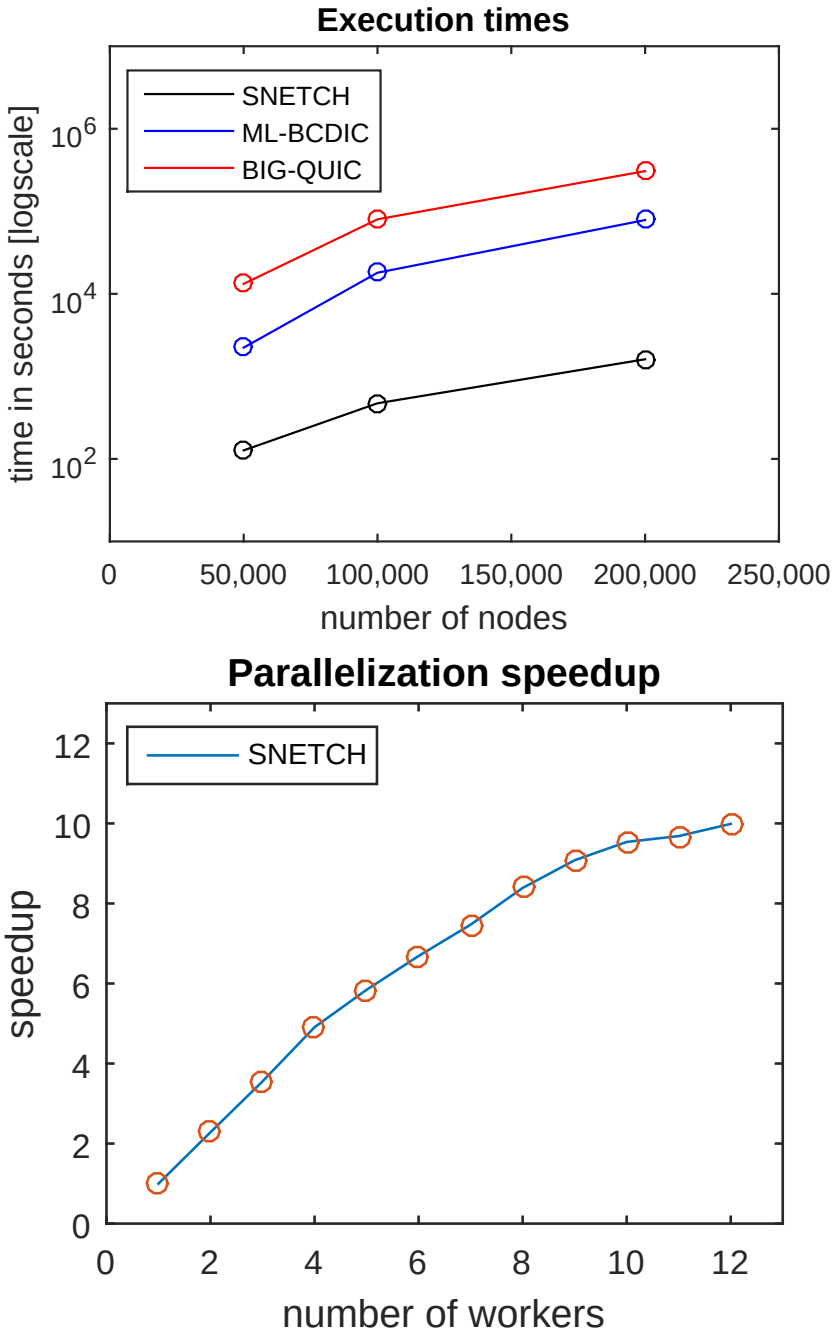
Scale-Free Structure (p=2,000; n=2,000; true nz = 14,466)						
Model	Time [s]	Jaccard	Prec.	Recall	λ	nz
SNETCH	0.78	0.407	0.577	0.580	0.0775	14,550
BCDIC	5.89	0.364	0.532	0.536	0.0717	14,548
QUIC	4.25	0.364	0.532	0.536	0.0717	14,550

Skalabilnost U svim tim navedenim primerima, naš algoritam je bio najbrži. Sledeći experiment ima za cilj da uporedi SNETCH protiv skalabilnih alternativa BIG & QUIC [31] i ML-BCDIC [80] u mnogo većim problemima gde su podaci veći od memorije računara. U tom cilju, stvorili smo sintetičke probleme sa scale-free strukturom različitih veličina: 50,000; 100,000 i 200,000 varijabli.

U levom panelu Slike 13, za logaritamsku y-osu, može se videti da naš pristup postiže vreme izvršavanja koje je ponekad dva reda magnitude brže od najsavremenijih konkurenata. Tabela 10 daje iste rezultate u više detalja, i treba napomenuti da u svim eksperimentima SNETCH [4], BIG & QUIC [5] i ML-BCDIC [6] su izvršavani na jedno-nitnom procesu.

Paralelizacija Štaviše, pošto naš pristup ima visok potencijal za paralelizovanje, prikazujemo na najvećem primeru koja brzina se može postići. Desna tabla na Slici 13 prikazuje kako se optimizacioni deo može ubrzati do 12 procesa. Pošto se

[4]https://github.com/vladisav/SNETCH
[5]http://bigdata.ices.utexas.edu/software/1035/
[6]https://github.com/erantreister/Multilevel-BCDIC.m



Slika 13: Skalabilnost SNETCH metode. Na gornjoj slici je predstavljeno poređenje vremena izvršavanja tri različite metode estimacije proređene inverzne kovarijansne matrice nad scale-free problemima do 200,000 promenljivih. Sve metode su pokrenute na jednoj niti. Na donjoj slici je prikazano ubrzanje dobijeno paralelizacijom SNETCH pristupa na scale-free problemu od 200,000 čvorova.

Tabela 10: Poređenje vremena izvršavanja [sec] SNETCH vs BIG & QUIC vs ML-BCDIC metoda na sintetičkim scale-free podacima.

scale-free		SNETCH	ML-BCDIC	BIG & QUIC
p= 50,000	time	126.9	2,230	13,221
249,970	nz	281,260	284,260	286,696
p=100,000	time	539.4	18,031	74,425
499,930	nz	765,990	752,250	757,194
p=200,000	time	1,611.8	78,349	307,081
2,199,434	nz	2,061,680	2,070,664	2,070,740

svaki od pod-ciljeva može paralelno optimizovati, očekuje se linearno ubrzanje. Eksperimentalni rezultati su u saglasnosti sa očekivanim stepenom paralelizma i čak pokazuju super-linearno ubrzanje za neke konfiguracije, što se može objasniti boljom iskorišćenošću keša memorije i manjim troškovima komunikacije, jer je potreban samo mali deo podataka kako bi se optimizovala svaka kolona. Paralelizacija je postignuta korišćenjem Matlab Parallel Computing Toolbox-a.

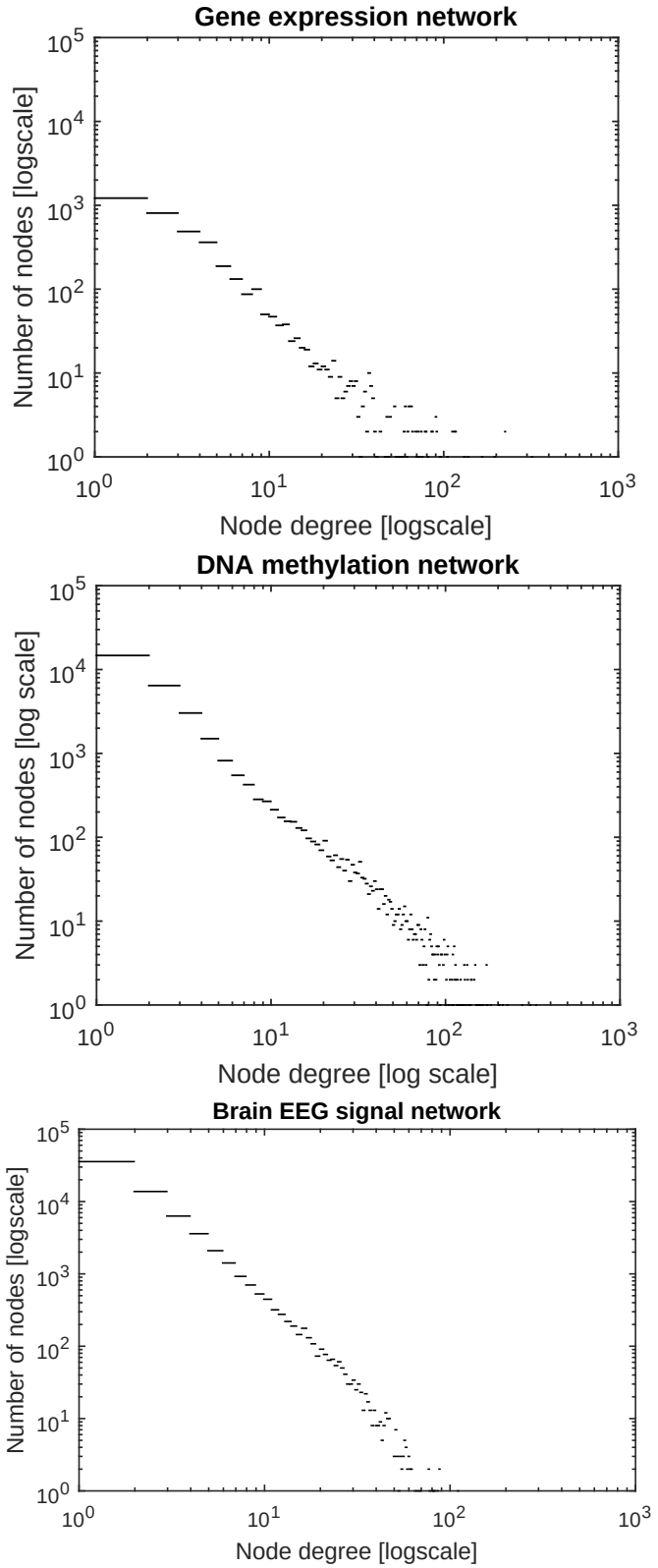
3.5.3 Aplikacije

Otkrivanje obrazca povezanosti izmedju velikog broja varijabli je uobičajena tema u velikom broju domena, a učenje GMRF-ovima je korišteno za rešavanje raznih problema: od biološkog funkcionisanja ćelije, do obrasca povezivanja mozga, i interakcija na društvenim mrežama.

Da bi se okarakterisale strukturne osobine nekih realnih skupova podataka, primenili smo naš SNETCH pristup u tri veoma bitne biomedicinske primene: ekspresiju ljudskog gena tokom respiratorne infekcije [45], DNA metilaciju kod zdravih ljudi [28] i elektroencefalogram (EEG) moždanih signala [83].

Podaci o genetskoj ekspresiji su dobijeni sa GEO[7], a sastoje se od 12,023 obeležja (gena) merenih u 2,886 uzoraka krvi uzetih od ljudskih subjekata zaraženih respiratornim virusima (grip tipa H3N2, Rinovirus i respiratorni sincitijalni virus) [45]. Podaci su standardizovani tako da imaju nultu srednju vrednost i normalizovani

[7]GSE73072, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73072



Slika 14: Raspodela stepeni naučene mreže povezanosti ko-ekspresije gena, DNK metilacione mreže, i EEG mreže signala u mozgu. Za sve tri mreže uočlivo je izraženo scale-free svojstvo.

Tabela 11: Vreme učenja strukture grafa za probleme genetske ekspresije, DNK metilacije, i EEG signala.

dataset	size	SNETCH	ML-BCDIC	BIG & QUIC
Gene Expr	time	130.5	342.1	1,442.9
p=12,023	nz	35,123	33,677	35,921
DNA Meth	time	12,667	400,000+	400,000+
p=473,034	nz	613,868	-	-
Brain EEG	time	32,803	400,000+	400,000+
p=904,739	nz	1,090,607	-	-

su da imaju jedinične vrednosti na dijagonali kovarijansne matrice. Otkrivanje mreže za koregulaciju gena je od visokog biološkog interesa jer ko-izraženi geni su funkcionalno povezani, često uključeni u iste putanje ili kontrolisani istim procesom [74].

Podaci o DNA metilaciji su dobijeni sa GEO[8] i ima nivoe metilacije izmerene na 473,034 genomske lokacije, za 656 zdravih ljudi (uzoraka) uključenih u istraživanje o korisnosti metilacionih markera kao prediktora biološke starosti [28]. Podaci su transformisani logit funkcijom [11], a zatim standardizovani i normalizovani. Izvlačenje mreže korelacije DNK metilacije je važno jer je metilacija jedan od najstabilnijih epigenomičnih mehanizama kroz koje životna sredina utiče na funkcionisanje organizma [7].

EEG podaci su dobijeni iz repozitorijuma[9]. Podaci se sastoje od 910,476 signala izmerenih u dve sekunde od strane Emotive EPOC uredjaja. Posle uklanjanja 5,737 vremenskih serija koje su imale oštećena očitavanja (konstantni signal), podaci su standardizovani i normalizovani. Zavisnost između zabeleženih događaja može pružiti važan uvid u to kako mozak reaguje na različite situacije [83].

Za određene nivoe proređenosti dobijamo mrežu za ko-regulaciju gena od 35,123 nenulta elementa, mrežu korelacije DNK metilacije koja se sastoji od 613,868 nenultih elemenata i EEG korelacionih mreža mozga od 1,090,607 elemenata. Okarakter-

[8]GSE40279, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279
[9]http://www.mindbigdata.com/opendb

isali smo njihove šablone stepena povezanosti čvorova, koje su prikazane na Slici 14. Može se videti da svi imaju veoma karakterističan obrazac scale-free zakona, koji se na logaritamskoj skali pojavljuje kao prava linija. Ovi realni primeri, kao i niz drugih, ukazuju na potencijal za široku primenljivost predloženog SNETCH pristupa.

Tabela 11 pokazuje da je naš pristup estimirao strukturu za najkraće vremene, za svaki skup podataka. Dok na dva visokodimenzionalna problema konkurentni pristupi nisu završili optimizaciju za 400,000 sekundi (Big & QUIC je prijavio završavanje druge iteracije na 104,596 sekundi na DNK podacima i prve iteracije na 140,013 sekundi na EEG podacima).

3.5.4 Diskusija i ideje za dalji rad

Predstavili smo novu metodu za učenje velikih Gausovskih Markovljevih slučajnih polja i pokazali da je posebno pogodna za probleme sa scale-free strukturom. U poređenju sa drugim kompetitivnim GMRF modelima, naš model se pokazao efikasnijim na problemima kojima je Cholesky faktor grafa redak, što je osobina koja odgovara scale free strukturama. Takođe smo pokazali da se isti metod može primeniti uz dobre performanse i na nekoliko drugih retkih problema različite strukture. Empirijski rezultati pokazuju značajno ubrzanje na nekoliko sintetičkih skupova podataka. Takođe primenili smo model na probleme pronalaženja strukture pri ekspresiji gena, DNK metilacijama i EEG signalima, i pokazali da svaki od ovih problem ispoljava scale-free osobine. Pošto retkost Cholesky faktora u velikoj meri zavisi od pronalaženja najboljeg redosleda čvorova (za koji je poznato da je NP-kompletan problem), predložili smo korišćenje Approximate Minimum Degree algoritma, kao podkorak u optimizaciji za ublažavanje ovog problema, što se pokazalo kao dobro rešenje u našim eksperimentima. Potrebno je dalje istražiti načine kako bi se model poboljšao sa naprednijom strategijom numerisanja čvorova. Takođe, u našim eksperimentima, korišćena je optimizacija prvog reda, dok bi op-

timizacija predstavljenog pristupa kvadratnim algoritmom mogla obezbediti bolje stope konvergencije. Predloženi pristup pronalaženja relacija bi se mogao proširiti i na učenje diskriminativnog modela, tj. Gausovskog uslovnog slučajnog polja, i primeniti na strukturnim problemima regresije [70] kao što je predloženo u nastavku teze.

Dokaz konveksnosti Predloženi objektiv (48) je konveksan. S obzirom da je $g(L)$ separabilno po kolonama, dovoljno je pokazati da je optimizacijska funkcija za jednu kolonu konveksna, što se može učiniti napomenom da je Hessian matrica glatkog dela jednaka sumi podmatrice empirijske kovarijansne matrice i dijagonalne matrice sa jednim pozitivnim elementom, koji je uvek pozitivno definitan:

$$H_{g(L_j)} = S_{jj} + D_{jj} \succeq 0 \tag{55}$$

Penalizacioni član je takodje konveksan, pošto je to standardna L_1 norma nad optimizacionim varijablama, otežana ne-negativnim faktorima [41].

4 Učenje strukture za diskriminativne modele

4.1 Uvod i relevantna literatura

U radu [84] autori su se bavili problemom učenja uslovnih modela baziranih na grafu sa proređenom strukturom jer ta strukturna predpostavka pogoduje generalizacionim mogućnostima modela, kao i skalabilnosti metode na velike probleme. U novijem istraživanju [46] je predložen blok-koordinatni pristup za ubrzavanje prethodno pomenutog algoritma. U nastavku, oslanjajući se na pretpostavke uvedene prilikom razvijanja prethodnih modela, predstavićemo teorijski novi model Gausovskih slučajnih polja pogodan za učenje strukture i predikciju na scale-free problemima. Predstavljen je algoritam za obučavanje zasnovan na koordinatnom spustu, jednačine za ažuriranje parametara, i preliminarni rezultati evaluacije ovakvog pristupa.

4.2 SFGCRF

4.2.1 Metod

Ciljni vektor $y \in \mathbb{R}^{p\times 1}$ (p-dimenzioni vektor kolona), je uslovljen (zavistan) od vektora slucajnih promenljivih $x \in \mathbb{R}^{q\times 1}$ (q-dimenziona kolona vektor slučajnih promenljivih). Funkcija uslovne gustine verovatnoće y je:

$$p(Y|X) = (2\pi)^{-p/2}\,\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(y-\mu(X))^{\mathrm{T}}\Sigma^{-1}(y-\mu(X))\right) \tag{56}$$

Vektor μ predstavimo kao funkciju parametara (kao što autori [84] rade):

$$\mu(x) = \Sigma\Theta^{T}x \tag{57}$$

Nakon smene i množenja u gornjem izrazu dobija se:

$$p(Y|X) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y - 2x^T \Theta y)\right) \tag{58}$$

gde je normalizaciona funkcija Z izražena kao:

$$Z(X) = (2\pi)^{-p/2} \ \det(\Sigma)^{-1/2} \exp\left(\frac{1}{2}(\Sigma\Theta^T x)^T \Sigma^{-1}(\Sigma\Theta^T x)\right) \tag{59}$$

Sada se vrši negativno logaritmovanje izraza i uvodi se trag skalara:

$$l(L) = \frac{1}{2} tr\left(YY^T\Sigma^{-1}\right) - tr(\Theta Y X^T) - \frac{1}{2} log|\Sigma^{-1}| + \frac{1}{2} tr(XX^T\Theta\Sigma\Theta^T) \tag{60}$$

Uvodimo dodatne smene:

$YY^T = S^{yy}$, što je $p \times p$ matrica

$YX^T = S^{yx}$, što je $p \times m$ matrica

$XX^T = S^{xx}$, što je $m \times m$ matrica

Dok je matrica Y dimenzija $p \times n$, gde je n broj semplova, a matrica X je dimenzija $m \times n$, gde je m broj feature-a. Matrica Θ je dimenzija $m \times p$, gde je p broj čvorova.

$$l(L) = \frac{1}{2} tr\left(S^{yy}\Sigma^{-1}\right) - tr(\Theta S^{yx}) - \frac{1}{2} log|\Sigma^{-1}| + \frac{1}{2} tr(S^{xx}\Theta\Sigma\Theta^T) \tag{61}$$

Sada se parametrizacija menja tako što se Choleski faktorizuje inverzna kovarijansna matrica:

$$l(L) = \frac{1}{2} tr\left(S^{yy}LL^T\right) - tr(\Theta S^{yx}) - \frac{1}{2} log|LL^T| + \frac{1}{2} tr(S^{xx}\Theta {L^{-1}}^T L^{-1}\Theta^T) \tag{62}$$

U ovom izrazu je nezgodno što postoji član L^{-1}, zato ubacujemo još jednu smenu:

$$L^{-1}\Theta^T = W^T \tag{63}$$

W je matrica $p \times m$ parametara koje ćemo takodje da optimizujemo. Odatle dobijamo:

$$\Theta = WL^T \tag{64}$$

Sada jednačine poprimaju sledeći oblik:

$$l(L) = \frac{1}{2}tr\left(LL^T S^{yy}\right) - tr(WL^T S^{yx}) - \frac{1}{2}log|LL^T| + \frac{1}{2}tr(WW^T S^{xx}) \tag{65}$$

Prvi i treći član su isti kao i u GMRF modelu prethodno razmatranom; to su empirijski član i logaritmovana determinanta inverzne koverijanse. Četvrti član je uslovni član, a treći član je funkcija obeležja i izlaza.

Dosadašnja derivacija je samo reparametrizacija originalnog problema; ako bi koristili istu regularizaciju kao i u [84] (tj L_1 penalizovali originalne parametre predstavljene u novoj parametrizaciji) dobili bi identične rezultate.

Objektiv (65) je isto separabilan po kolonama kao i [34]. To se najlakše vidi kroz izvode. Slede prvo diferencijali (Δ_X matrica je matrica u kojoj su svi elementi nule izuzev odgovarajuce kolone j koja ima vrednosti j kolone matrice X, a vec je operator koji preslaže matricu u vektor, tj operator vektorizacije):

$$vec(\frac{\partial g(L,W)}{\partial L_{ij}})^T vec(\Delta_L) = tr\left(L{\Delta_L}^T S^{yy}\right) + tr(W{\Delta_L}^T S^{yx}) \tag{66}$$

Za elemente na dijagonali važi isti izraz, samo je dodata logaaritmovana determinanta (i naravno $i = j$). Diferencijali po promenljivama W:

$$vec(\frac{\partial g(L,W)}{\partial W_{lj}})^T vec(\Delta_W) = -tr(\Delta_W L^T S^{yx}) + tr(W{\Delta_W}^T S^{xx}) \tag{67}$$

Iz njih se lako dobijaju jednačine za izvode (a samim tim i uslovi optimalnosti):

$$\frac{\partial g(L,W)}{\partial L_{ij}} = \sum_{k=1}^{p} {S^{yy}}_{ik} L_{kj} - \sum_{k=1}^{m} {S^{xy}}_{ik} W_{kj} \tag{68}$$

$$\frac{\partial g(L,W)}{\partial W_{lj}} = \sum_{k=1}^{m} {S^{xx}}_{lk} W_{kj} - \sum_{k=1}^{p} {S^{xy}}_{lk} L_{kj} \tag{69}$$

Kao što se vidi, separabilan je po kolonama (u izrazu za gradijent za svaku kolonu, figuriraju isključivo samo elementi te kolone). Za dalji progres postoje dva pravca:

1. Da se penalizuje Σ i Θ matrica.

2. Da se penalizuju novo-uvedeni parametri L i W. U tom slučaju, objektiv ostaje i dalje separabilan po kolonama i dozvoljava jednostavnu paralelizaciju.

U nastavku se fokusiramo na ideju 2. Dodajemo dva penala na glavni objektiv (65) : $\lambda_L |L|_1 + \lambda_W |W|_1$

Sada su jednačine za izvode:

$$\frac{\partial g(L,W)}{\partial L_{ij}} = \sum_{k=1}^{p} {S^{yy}}_{ik} L_{kj} - \sum_{k=1}^{m} {S^{yx}}_{ik} W_{kj} + \lambda_L sign(L_{ij}) \tag{70}$$

$$\frac{\partial g(L,W)}{\partial W_{lj}} = \sum_{k=1}^{m} {S^{xx}}_{lk} W_{kj} - \sum_{k=1}^{p} {S^{xy}}_{lk} L_{kj} + \lambda_W sign(W_{lj}) \tag{71}$$

Izjednačavanjem ovih izraza sa nulom dobijamo nove jednacine prvog reda za zatvoreno rešenje gradijenta po pojedinačnim koordinatama. Na osnovu uslova optimalnosti sledi:

$$L_{ij} = -\frac{\sum_{k \neq i}^{p} {S^{yy}}_{ik} L_{kj} - \sum_{k=1}^{m} {S^{yx}}_{ik} W_{kj} + \lambda_L sign(L_{ij})}{{S_{ii}}^{yy}} \tag{72}$$

$$W_{lj} = -\frac{\sum_{k \neq l}^{m} {S^{xx}}_{lk} W_{kj} - \sum_{k=1}^{p} {S^{xy}}_{lk} L_{kj} + \lambda_W sign(W_{lj})}{S_{ll}^{xx}} \tag{73}$$

Jednačina (72) je slična kao i kod prethodno analiziranih SFCHL [71] i SNETCH [34] metoda, samo se konstantni član razlikuje. Tako da poseduje pogodne karakteristike kao što su brzina izracunavanja, jednostavno odredjivanje aktivnog skupa, i visok stepen paralelizacije.

4.2.2 Empirijska evaluacija

U cilju ispitivanja performansi predloženog algoritma na problemima učenja iz kondicionalno zavisnih podataka, generisali smo sintetičke primere različitih dimenzija. Prostor ulaznih varijabli smo izabrali da bude dva u svim primerima. Broj izlaznih varijabli (koje uslovno zavise od ulaznih) smo varirali od 10 do 500.

Graf povezanosti izlaznih varijabli je generisan tako što su neki elementi van dijagonale inverzne kovarijansne matrice slučajno popunjeni, dok su ostali ostali na nultim vrednostima. Obezeđeno je da inverzna kovarijansna matrica bude simetrična i pozitivno definitna. Linearno preslikavanje iz ulaznih u izlazne varijable je takođe proređeno i slučajno izabrano uniformno iz zatvorenog intervala.

Odbirci podataka su izvučeni iz normalne multivarijatne distribucije sa prosečnim vrednostima dobijenim iz linearnog preslikavanja, i kovarijansnom matricom dobijenom na prethodno opisan način. Po hiljadu ulazno-izlaznih parova je sačuvano za procese treniranja i testiranja.

U eksperimentima su poređena dva algoritma za učenje, naš predloženi SFGCRF, i prethodno publikovani model koji ćemo nazvati ALT (kao alternativni pristup). U Tabelama 12 i 13 je prikazano nekoliko indikatora performansi algoritama. RMSE i R2 su mere generalizacionih sposobnosti algoritama, tj koliko dobro je model naučio funkcionalno mapiranje iz ulaznih u izlazne varijable. RMSE je mera greške i što je manji broj to je algoritam bolji, dok je R2 više mera korelacije predviđenog i originalnog signala i što je bliže jedinici to je algoritam bolji. Preciznost i Odziv su mere koliko je dobro naučena struktura proređenosti inverzne kovarijansne matrice. Preciznost govori o tome koliki je procenat predviđenih nenultih elemenata

Tabela 12: Problemi dimenzija od 10 do 100 čvorova

Size	100	100	50	50	10	10
Model	ALT	SFGCRF	ALT	SFGCRF	ALT	SFGCRF
time	0.10415	0.74683	0.061973	0.22656	0.06722	0.037549
precision	0.44246	0.25659	0.49114	0.43534	0.45455	0.40909
recall	0.44246	0.16641	0.49871	0.25964	0.47619	0.42857
accuracy	0.8566	0.8308	0.8416	0.8324	0.77	0.75
lambda	0.07	0.21	0.09	0.24	0.24	0.32
nnz true	2672	2672	828	828	52	52
nnz est	2672	1768	840	514	54	54
rmse	0.70377	0.66491	0.79124	0.76242	1.7571	1.0033
r2	0.77017	0.79485	0.86427	0.87398	0.76177	0.92234

Tabela 13: Problemi dimenzija od 200 do 1000 čvorova

Size	1000	1000	500	500	200	200
Model	ALT	SFGCRF	ALT	SFGCRF	ALT	SFGCRF
Time	372.46	26.105	22.16	8.0577	0.65836	1.7452
recall	0.16262	0.085677	0.33782	0.10283	0.35501	0.14547
accuracy	0.97305	0.97523	0.95985	0.95247	0.93302	0.92868
lambda	0.08	0.12	0.06	0.14	0.08	0.17
nnz true	35128	35128	16994	16994	4352	4352
nnz est	31864	22268	15224	11164	4354	2962
rmse	0.69331	0.61155	0.6601	0.62178	0.69498	0.66072
r2	0.42322	0.55124	0.60437	0.64897	0.67863	0.70954

u matrici zapravo pogođen, dok Odziv govori o tome koliki je procenat od tačnih elemenata vraćen algoritmom. Obe mere su što veće to bolje po algoritam. I na kraju vreme izvršavanja, naravno što je manje to bolje.

4.2.3 Diskusija

Iz rezultata prikazanih u Tabelama 12 i 13 se može zaključiti da predlozeni SFGCRF algoritam postiže bolje prediktivne performanse jer je na svim primerima postizao manju grešku predikcije RMSE. Međutim izgleda da je alternativni algoritam bolji što se tiče otkrivanja strukture matrica. Tako da je bitno i koji je krajnji cilj upotrebe algoritma, da li samo kao prediktivni model ili za otkrivanje strukture. Konačno, na manjim primerima izgleda da je SFGCRF modelu potrebno više vremena (tabela 12). Međutim, izgleda da sa porastom dimenzionalnosti problema alternativni algoritam

radi sve sporije i sporije, što se može videti iz Tabele 13. To ukazuje da je predloženi algoritam skalabilniji i može se koristiti na velikim problemima. Ostaje za buduće istraživanje da se utvrdi na koliko velikim problemima je algoritam primenjiv, kao i da se istraže drugi slučajevi strukture matrica osim slučajne inverzne kovarijansne matrice.

5 Uslovna slučajna polja sa teškim repovima

U ovom poglavlju daćemo još jedan predlog za dalji nastavak istraživanja. Poznato je da se Gausovski modeli ne ponašaju adekvatno kada su podaci nastali iz distribucije sa teškim repovima. Takođe, Gausovski modeli predstavljeni do sada su simetrični, i ne modeluju asimetričnost distribucije. U nastavku, predložićemo jedan model koji uvažava pomenute pretpostavke.

5.1 NIGCRF

Proširili smo GCRF model sa dve nove osobine. Prvo, omogućili smo modelovanje teških i polu-teških repova distribucije uvođenjem skrivene varijable. Ova nova varijabla deluje kao faktor skaliranja kovarijanse, što omogućava korišćenje različite kovarijanse za različite uzorke.

Drugo, omogućili smo asimetričnost interakcionog potencijala uvođenjem novog člana u interakciju. Nakrivljenost (eng. skewness) novoformirane distribucije je direktna posledica novodefinisanog potencijala interakcije kao i dodatne skrivene slučajne varijable.

Predloženi GCRF model je predstavljen sledećim izrazom:

$$l(y; R, G, \beta, z) = \sum_{i}^{N} \sum_{j}^{P} \beta_j (y_{ij} - R_{ij}) + \frac{1}{z_i} \sum_{j}^{P} G_{jj} (y_{ij} - R_{ij})^2 + \frac{1}{z_i} \sum_{(j,k)} G_{jk} (y_j - y_k)^2 \tag{74}$$

Kada skrivena varijabla z ima Inverznu Gausovu distribuciju, lako se može pokazati da cela distribucija odgovara takozvanoj distribuciji Multivarijatne Normalne Inverzne Gausovske distribucije (eng. Multivariate Normal Inverse Gaussian distribution, skraćeno MNIG). Ova raspodela pripada familiji generalizovanih hiperboličnih distribucija, koje su definisane kao Normal-Variance mixture distribucije. Ova familija poseduje nekoliko atraktivnih osobina, kao što je beskonačna deljivost (eng. infinite divisibility) i zatvorena je nad skupom afinih transformacija. Ono što je obećavajuće u ovakvom pristupu, je to što su to upravo svojstva koja čine Gausovska slučajna polja atraktivnim, što ukazuje na potencijal ovakvog modela.

Generativni proces modela može se posmatrati kao uzorak iz sledeće mixture distribucije:

$$Y = \mu + Z\Sigma\beta + \sqrt{Z}\Sigma^{\frac{1}{2}}U \tag{75}$$

Ovde je promenljiva $U \sim N(0,1)$ normalno raspodeljena promenljiva, Σ je kovarijansna matrica, β je novouvedeni parametar koji određuje nakrivljenost distribucije, dok je Z skrivena promenljiva uzorkovana iz Inverzne Gausovske distribucije $IG(\delta^2, \alpha^2 - \beta^T\Sigma\beta)$.

Učenje parametara ovako postavljenog modela je nešto složenije od učenja Gausovskih slučajnih polja. Postoji nekoliko pristupa estimacije MNIG distribucije. Jedan od perspektivnijih za naš problem skalabilne estimacije retke MNIG, tj. NIGCRF-a, je predstavljen u [50] i zasnovan je na algoritmu maksimizacije očekivanja (eng. Expectation Maximization, skraćeno EM). Razlika između našeg modela i modela predstavljenog u [50] je u skalabilnosti i retkosti inverzne kovarijansne matrice. U problemima koji su od interesa za ovo istraživanje neophodno je naučiti retku matricu (graf interakcija), dok je model u [50] zasnovan na učenju inverzne kovarijansne matrice punog ranga. Takođe, model predstavljen u [50] je pogodan isključivo za malodimenzionalne probleme, usled velike računske složenosti algoritma. Istraživanje predstavljeno u ovoj tezi je fokusirano na visokodimenzionalne

probleme.

5.1.1 Metod

U nastavku ćemo opisati algoritam za učenje parametara predloženog modela. Usled postojanja skrivene promenljive, za učenje je pogodan EM algoritam. U ovom algoritmu, iterativno (do konvergencije) se primenjuju dva različita koraka: Estimacioni korak, u kojem se vrši estimacija skrivene promenljive na osnovu svih ostalih, i Maksimizacioni korak, u kojem se vrši maksimizacija funkcije verodostojnosti koristeći estimirane vrednosti skrivenih promenljivih.

Estimacioni korak skrivene promenljive (E korak u EM algoritmu) je zajednički i za originalni, i za naš izvedeni model, stoga su jednačine ažuriranja iste (preuzete iz [50]):

$$\varsigma_i = E\{Z_i|Y_i = y_i\} = \frac{q(y_i)}{\alpha}\frac{K_{(d-1)/2}[\alpha_q(y_i)]}{K_{(d+1)/2}[\alpha_q(y_i)]} \tag{76}$$

$$\varsigma_i = E\{Z_i|Y_i = y_i\} = \frac{q(y_i)}{\alpha}\frac{K_{(d-1)/2}[\alpha_q(y_i)]}{K_{(d+1)/2}[\alpha_q(y_i)]} \tag{77}$$

$$\phi_i = E\{Z_i^{-1}|Y_i = y_i\} = \frac{\alpha}{q(y_i)}\frac{K_{(d+3)/2}[\alpha_q(y_i)]}{K_{(d+1)/2}[\alpha_q(y_i)]} \tag{78}$$

za $i = 1, ..., N$. Zatim, moguće je izračunati potrebne momente skrivene promenljive $\xi = \sum_{i=1}^{N} \varsigma_i/N$ i $\upsilon = N(\sum_{i=1}^{N}(\phi_i - \xi^{-1}))^{-1}$.

M korak maksimizuje ukupnu verodostojnost tako što ažurira parametre koristeći očekivane vrednosti dovoljnih statistika pronađenih u E-koraku. Kao što smo napomenuli ranije, uslovna distribucija $Y_i|Z_i$ je Gausovski raspodeljena. U nastavku ćemo izvesti izraz za ovu uslovnu distribuciju i ujedno izvesti izraz za M korak algoritma za učenje.

Kanonski oblik GCRF modela (u kojem je G težinski graf veza između promenljivih)

je:

$$l(y; R, G) = \sum_{i}^{N} \sum_{j}^{P} G_{jj}(y_{ij} - R_{ij})^2 + \sum_{(j,k)} G_{jk}(y_{ij} - y_{ik})^2 \tag{79}$$

Predloženi model uvodi dodatni član koji predstavlja nakrivljenost distribucije:

$$\sum_{j}^{P} \beta_j^T (y_{ij} - R_{ij}) \tag{80}$$

Takodje ubacujemo mixture efekat pomoću slučajne promenljive z_i, koja ima različitu vrednost za različite odbirke (sempluje se iz Inverse Gaussian distribucije). Ovakav efekat služi da modeluje teške repove distribucije.

$$l(y; R, G, \beta, z) = \sum_{i}^{N} \sum_{j}^{P} \beta_j (y_{ij} - R_{ij}) + \frac{1}{z_i} \sum_{j}^{P} G_{jj}(y_{ij} - R_{ij})^2 + \frac{1}{z_i} \sum_{(j,k)} G_{jk}(y_j - y_k)^2 \tag{81}$$

Sada ako pogledamo kako izgleda mixture component distribucija za Multivariate normal inverse Gaussian distribuciju:

$$p(Y|\mu, \beta, z, \Sigma^{-1}) \sim \mathcal{N}(\mu + z\Sigma^{-1}\beta, z\Sigma^{-1}) \tag{82}$$

$$p(Y|\mu, \beta, z, \Sigma^{-1}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(y_i - \mu - z_i\Sigma\beta)^T \Sigma^{-1} (y_i - \mu - z_i\Sigma\beta)} \tag{83}$$

Sada posmatramo negative log-likelihood:

$$l(y; \mu, \beta, z, \Sigma^{-1}) = \frac{N}{2} \log |\Sigma^{-1}| + \frac{1}{2} \sum_{i}^{N} \frac{1}{z_i} (y_i - \mu - z_i\Sigma\beta)^T \Sigma^{-1} (y_i - \mu - z_i\Sigma\beta) \tag{84}$$

$$l(y;\mu,\beta,z,\Sigma^{-1}) = \frac{N}{2}\log|\Sigma^{-1}| + \frac{1}{2}\sum_{i}^{N}\frac{1}{z_i}y_i^T\Sigma^{-1}y_i^T + \frac{1}{z_i}\mu^T\Sigma^{-1}\mu + z_i\beta^T\Sigma\beta \\ -\frac{2}{z_i}y_i^T\Sigma^{-1}\mu - 2y_i^T\beta - 2\mu_i^T\beta \tag{85}$$

Kada izjednačimo jednacine (81) i (85) dobijamo sledeće:

Izjednačimo članove prvog reda (y_i):

$$\sum_{j}^{P}\sum_{k}^{P} -y_j\mu_k\Sigma_{jk}^{-1} - \sum_{j}^{P} y_j\beta = \sum_{j}^{P} -2G_{jj}y_jR - y_j\beta \tag{86}$$

$$\sum_{k}^{P}\mu_k\Sigma_{jk}^{-1} = 2G_{jj}R \tag{87}$$

U matričnom obliku:

$$\Sigma^{-1}\mu = 2G_dR \tag{88}$$

G_d je dijagonalna matrica koja sadrži glavnu dijagonalu G matrice.

Zatim izjednačavamo članove drugog reda (y_i^2):

$$\sum_{j}^{P}\sum_{k}^{P} y_jy_k\Sigma_{jk}^{-1} = \sum_{j}^{P} G_{jj}y_j^2 + \frac{1}{2}\sum_{j}^{P}\sum_{k\neq j}^{P} G_{jk}y_j^2 - 2G_{jk}y_jy_k + G_{jk}y_k^2 \tag{89}$$

Sada dobijamo inverznu kovarijansnu matricu izraženu preko parametara grafa:

$$\Sigma_{ij}^{-1} = \begin{cases} G_{ii} + \sum_{k\neq i} G_{ik}, & \text{if } i = j \\ -G_{ij}, & \text{if } i \neq j \end{cases} \tag{90}$$

Sada je lako uvideti da inverzna kovarijansna matrica ima isti obrazac retkosti kao i G matrica.

Matrica G je Laplacian grafa, i pretpostavka je da je retka matrica. Uvodimo još jednu dodatnu pretpostavku da je Cholesky faktorizacija grafa (a samim tim

i precision matrice) takodje retka, pretpostavka o kojoj je bilo reči u prethodnim poglavljima:

$$\Sigma_{ij}^{-1} = LL^T \tag{91}$$

gde je L retka donje trougaona matrica čiji obrazac popunjenosti se može izračunati na osnovu matrice susedstva (inicijalne vrednosti matrice G).

Sada je log-likelihood:

$$\begin{aligned} l(y;\mu,\beta,z,L) = \frac{N}{2}\log(\prod_{j=1}^{P} L_{jj}^2) + \frac{1}{2}\sum_{i}^{N}\frac{1}{z_i}y_i^T(LL^T)^{-1}y_i^T + \frac{1}{z_i}\mu^T LL^T\mu \\ +z_i\beta^T(LL^T)^{-1}\beta - \frac{2}{z_i}y_i^T(LL^T)^{-1}\mu - 2y_i^T\beta - 2\mu_i^T\beta \end{aligned} \tag{92}$$

Da bi uspešno modelirali repove mixture distribucije neophodno je da ukupna disperzija distribucije zavisi samo od mixture težinskih koeficijenata z. Stoga, uslov da bi ukupna distribucija bila Normal Inverse Gaussian je:

$$\det\Sigma = 1 \tag{93}$$

Ovo možemo postići tako što ćemo uvrstiti dodatni constraint u vidu Lagrange multiplier-a:

$$\lambda\log\prod_{j=1}^{P}(L_{jj}^2) = 2\lambda\sum_{j=1}^{P}\log(L_{jj}) \tag{94}$$

Kako je λ slobodna optimizaciona promenljiva, mozemo uprostiti izraz tako što ćemo optimizovati 2 λ (čime se efektivno gubi činilac 2 iz gornjeg izraza).

$$l(y;\mu,\beta,z,L) = \frac{1}{2}\sum_{i}^{N}(\frac{1}{z_i}y_i^T(LL^T)^{-1}y_i^T + \frac{1}{z_i}\mu^T LL^T\mu + z_i\beta^T(LL^T)^{-1}\beta \\ - \frac{2}{z_i}y_i^T(LL^T)^{-1}\mu - 2y_i^T\beta - 2\mu_i^T\beta) + (N-\lambda)\sum_{j=1}^{P}\log(L_{jj}) \tag{95}$$

Uvodimo smenu μ parametra:

$$l(y;R,\beta,z,L) = \frac{1}{2}\sum_{i}^{N}(\frac{1}{z_i}y_i^T(LL^T)^{-1}y_i^T + \frac{1}{z_i}R^T G_D LL^T G_D R + z_i\beta^T(LL^T)^{-1}\beta \\ - \frac{2}{z_i}y_i^T(LL^T)^{-1}G_D R - 2y_i^T\beta - 2R^T G_D\beta) + (N-\lambda)\sum_{j=1}^{P}\log(L_{jj}) \tag{96}$$

gde je G_D matrica izražena u L parametrima (na osnovu jednačina (90) i (91)):

$$G_{Dij} = \begin{cases} \sum_{j\neq i}\sum_k L_{ik}L_{jk}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{97}$$

Sada možemo definisati parcijalne izvode negativne log-likelihood funkcije po parametrima:

$$\frac{\partial l}{\partial \lambda} = \sum_i \log L_{ii} \tag{98}$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{2}\sum_i 2z_i Tr(E_j^T X_\beta) - 2y_i^T E_j + 2X_{RG}^T E_j \\ = E_j^T(z_i X_\beta - y_i - X_{RG}) \tag{99}$$

U gornjem izrazu, E_j predstavlja vektor čiji je element j jednak 1, dok svi ostali elementi imaju vrednost nula. Vektor X_β je jednak: $X_\beta = (LL^T)^{-1}\beta$. Ovaj vektor može efikasno da se izracuna (u $O(M)$ operacija, gde je M broj ne-nula u L matrici) pomocu Gausove eliminacije (backward-forward algoritam):

$$LL^T X_\beta = \beta \tag{100}$$

$$LX_t = \beta \tag{101}$$

$$L^T X_\beta = X_t \tag{102}$$

Obe jednačine (101) i (102) rešive su u $O(M)$ vremenu. Iste garancije na vreme izvršavanja važe i za vektor X_{RG}:

$$X_{RG} = (LL^T)^{-1} RG_D \tag{103}$$

M korak je konveksan, i stoga se može optimizovati metodom koordinatnog spusta.

6 Zaključak

U ovoj disertaciji, predmet istraživanja su Verovatnosni grafovski modeli, i to Markovljeva slučajna polja (Markov Random Field) i Uslovna slučajna polja (Conditional Random Field).

Cilj istraživanja je pronalaženje i analiza novih modela za struktuirano učenje i algoritama za njihovo obučavanje.

U sklopu doprinosa ove teze, najpre je osmišljena i realizovana generalizacija jedne postojeće GCRF formulacije. Zatim je izvršena eksperimentalna analiza predloženog modela na nekoliko sintetičkih i nekoliko stvarnih skupova podataka iz aktuelnih domena klimatologije i zdravstva, kao i teorijska analiza složenosti.

Zatim je uočen tip problema koji je veoma zastupljen a nije dovoljno obrađen u literaturi, i za koji je moguće uvesti dodatne pretpostavke za koje je zatim pokazano da mogu znatno olakšati problem učenja strukture iz mnogodimenzionalnih podataka. Predložena su i realizovana dva nova GMRF modela zasnovana na uvedenim pretpostavkama i L1 regularizacione norme, i razvijeni su brzi i skalabilni algoritmi za njihovo učenje.

Pokazana je prednost predloženih modela nad postojećim resenjima u vidu brzine i skalabilnosti. Predloženi algoritmi su eksperimentalno potvrđeni kroz poređenje sa najbrzim postojećim resenjima na nekoliko mnogodimenzionalnih sintetičkih problema kao i na pravim podacima iz oblasti ekspresije gena, DNK metilacije, i EEG signala. Zatim je odrađena teorijska analiza složenosti algoritama, i zatim eksperimentalno potvrđena sposobnost paralelizacije.

Za kraj su predložene su još dve dodatne ekstenzije GCRF modela i postavljena je teorijska osnova za njihovo učenje.

Literatura

[1] Gökhan Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J Smola, Ben Taskar, and S.V.N. Vishwanathan. *Predicting Structured Data*. MIT Press, 2007.

[2] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[3] Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, and Georges Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd ICML*, pages 89–96. ACM, 2006.

[4] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

[5] Albert-László Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413, 2009.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[7] Thomas E Bartlett, Sofia C Olhede, and Alexey Zaikin. A dna methylation network interaction measure, and detection of network oncomarkers. *PloS one*, 9(1):e84573, 2014.

[8] Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[9] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] Xi Hang Cao, Ivan Stojkovic, and Zoran Obradovic. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC bioinformatics*, 17(359):1–10, 2016.

[12] Joachim Dahl, Lieven Vandenberghe, and Vwani Roychowdhury. Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software*, 23(4):501–520, 2008.

[13] Aaron Defazio and Tiberio S Caetano. A convex formulation for learning scale-free networks via submodular relaxation. In *Advances in Neural Information Processing Systems*, pages 1250–1258, 2012.

[14] Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

[15] Adrian Dobra, Chris Hans, Beatrix Jones, Joseph Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

[16] Tatjana Dokic, Payman Dehghanian, Po-Chen Chen, Mladen Kezunovic, Zenon Medina-Cetina, Jelena Stojanovic, and Zoran Obradovic. Risk assessment of a transmission line insulation breakdown due to lightning and severe weather. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 2488–2497. IEEE, 2016.

[17] Richard C Dubes and Anil K Jain. Random field models in image analysis. *Journal of applied statistics*, 16(2):131–164, 1989.

[18] John Duchi, Stephen Gould, Daphne Koller, et al. Projected subgradient methods for learning sparse gaussians. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

[19] Mary J Dunlop, Robert Sidney Cox, Joseph H Levine, Richard M Murray, and Michael B Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature genetics*, 40(12):1493–1498, 2008.

[20] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[22] A. George and W. H. Liu. The evolution of the minimum degree ordering algorithm. *SIAM Rev.*, 31(1):1–19, March 1989.

[23] Jesse Glass and Zoran Obradovic. Structured regression on multiscale networks. *IEEE Intelligent Systems*, 32(2):23–30, 2017.

[24] Djordje Gligorijevic, Jelena Stojanovic, and Zoran Obradovic. Improving confidence while predicting trends in temporal disease networks. In *SIAM SDM 4th Workshop on Data Mining for Medicine and Healthcare*, 2015.

[25] Djordje Gligorijevic, Jelena Stojanovic, and Zoran Obradovic. Uncertainty Propagation in Long-term Structured Regression on Evolving Networks. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[26] Martin Charles Golumbic. *Algorithmic Graph Theory and Perfect Graphs (Annals of Discrete Mathematics, Vol 57)*. North-Holland Publishing Co., Amsterdam, The Netherlands, 2004.

[27] Lei Han, Yu Zhang, and Tong Zhang. Fast component pursuit for large-scale inverse covariance estimation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1585–1594. ACM, 2016.

[28] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[29] Jean Honorio and Tommi S Jaakkola. Inverse covariance estimation for high-dimensional data in linear time and space: Spectral methods for riccati and sparse models. In *Proceedings of the 29th Conference on Uncertainty in AI*, 2013.

[30] Cho-Jui Hsieh, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2011.

[31] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems*, pages 3165–3173, 2013.

[32] Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[33] Gabriel Huerta, Bruno Sansó, and Jonathan R Stroud. A spatiotemporal model for mexico city ozone levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(2):231–248, 2004.

[34] Vladisav Jelisavcic, Ivan Stojkovic, Veljko Milutinovic, and Zoran Obradovic. Fast learning of scale-free networks based on cholesky factorization. *International Journal of Intelligent Systems*, 2018.

[35] Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen,

et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471, 1996.

[36] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, December 1998.

[37] Seyoung Kim and Eric P Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet*, 5(8):e1000587, 2009.

[38] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[39] Sanjiv Kumar and Martial Hebert. Discriminative Fields for Modeling Spatial Dependencies in Natural Images. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2004.

[40] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, pages 282–289, 2001.

[41] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[42] Steffen L Lauritzen. *Graphical models*. Clarendon Press, 1996.

[43] Stan Z Li. *Markov random field modeling in computer vision*. Springer Science & Business Media, 2012.

[44] Qiang Liu and Alexander T Ihler. Learning scale free networks by reweighted l1 regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 40–48, 2011.

[45] Tzu-Yu Liu, Thomas Burke, Lawrence P Park, Christopher W Woods, Aimee K Zaas, Geoffrey S Ginsburg, and Alfred O Hero. An individualized predictor of

health and disease using paired reference and target samples. *BMC bioinformatics*, 17(1):1, 2016.

[46] Calvin McCarter and Seyoung Kim. Large-scale optimization algorithms for sparse conditional gaussian graphical models. In *Artificial Intelligence and Statistics*, pages 528–537, 2016.

[47] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[48] Matthew J Menne, Claude N Williams Jr, and Russell S Vose. The US Historical Climatology Network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, 90(7):993, 2009.

[49] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in NIPS*, 14:841–848, 2002.

[50] Tor Arne Øigård, Alfred Hanssen, and Roy Edgar Hansen. The multivariate normal inverse gaussian distribution: Em-estimation and analysis of synthetic aperture sonar data. In *Signal Processing Conference, 2004 12th European*, pages 1433–1436. IEEE, 2004.

[51] Grant P Parnell, Benjamin M Tang, Marek Nalos, Nicola J Armstrong, Stephen J Huang, David R Booth, and Anthony S McLean. Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. *Shock*, 40(3):166–174, 2013.

[52] Martin Pavlovski, Fang Zhou, Ivan Stojkovic, Ljupco Kocarev, and Zoran Obradovic. Adaptive skip-train structured regression for temporal networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 305–321. Springer, 2017.

[53] Fuchun Peng and Andrew McCallum. Information extraction from research papers using conditional random fields. *Information processing & management*, 42(4):963–979, 2006.

[54] Vladan Radosavljevic, Kosta Ristovski, and Zoran Obradovic. Gaussian Conditional Random Fields for Modeling Patients' Response to Acute Inflammation Treatment. In *ICML 2013 workshop on Machine Learning for System Identification*, 2013.

[55] Vladan Radosavljevic, Slobodan Vucetic, and Zoran Obradovic. Continuous Conditional Random Fields for Regression in Remote Sensing. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010) Lisbon, Portugal*, 2010.

[56] Vladan Radosavljevic, Slobodan Vucetic, and Zoran Obradovic. Neural Gaussian Conditional Random Fields. In *Machine Learning and Knowledge Discovery in Databases*, pages 614–629. Springer, 2014.

[57] Benjamin Rolfs, Bala Rajaratnam, Dominique Guillot, Ian Wong, and Arian Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2012.

[58] Kengo Sato and Yasubumi Sakakibara. RNA secondary structural alignment with conditional random fields. *Bioinformatics*, 21(suppl 2):ii237–ii242, 2005.

[59] Katya Scheinberg and Irina Rish. Learning sparse gaussian markov networks using a greedy coordinate ascent approach. In *Machine Learning and Knowledge Discovery in Databases*, pages 196–212. Springer, 2010.

[60] Volker Schmid and Leonhard Held. Bayesian extrapolation of space–time trends in cancer registry data. *Biometrics*, 60(4):1034–1042, 2004.

[61] Paul Sheridan, Takeshi Kamimura, and Hidetoshi Shimodaira. A scale-free structure prior for graphical models with applications in functional genomics. *PLoS One*, 5(11):e13580, 2010.

[62] Paul Sheridan, Yuichi Yagahara, and Hidetoshi Shimodaira. A preferential attachment model with poisson growth for scale-free networks. *Annals of the institute of statistical mathematics*, 60(4):747–761, 2008.

[63] HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). `www.hcup-us.ahrq.gov/sidoverview.jsp`, 2003–2011. Agency for Healthcare Research and Quality, Rockville, MD.

[64] Michael Sioutis and Jean-Francois Condotta. *Tackling Large Qualitative Spatial Networks of Scale-Free-Like Structure*. Springer International Publishing, Cham, 2014.

[65] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.

[66] Kyung-Ah Sohn and Seyoung Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Artificial Intelligence and Statistics*, pages 1081–1089, 2012.

[67] Jelena Stojanovic, Djordje Gligorijevic, and Zoran Obradovic. Modeling customer engagement from partial observations. In *CIKM*, pages 1403–1412, 2016.

[68] Jelena Stojanovic, Milos Jovanovic, Djordje Gligorijevic, and Zoran Obradovic. Semi-supervised learning for structured regression on partially observed attributed graphs. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015) Vancouver, Canada*. SIAM, 2015.

[69] Ivan Stojkovic, Mohamed F. Ghalwash, Xi Hang Cao, and Zoran Obradovic. Effectiveness of Multiple Blood-Cleansing Interventions in Sepsis, Characterized in Rats. *Scientific Reports*, 6(24719):1–11, 2016.

[70] Ivan Stojkovic, Vladisav Jelisavcic, Veljko Milutinovic, and Zoran Obradovic. Distance based modeling of interactions in structured regression. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence IJCAI-16*, pages 2032–2038, 2016.

[71] Ivan Stojkovic, Vladisav Jelisavcic, Veljko Milutinovic, and Zoran Obradovic. Fast sparse gaussian markov random fields learning based on cholesky factorization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence IJCAI-17*, pages 2758–2764, 2017.

[72] Ivan Stojkovic and Zoran Obradovic. Predicting sepsis biomarker progression under therapy. In *Proceedings of the 30th IEEE International Symposium on Computer-Based Medical Systems IEEE CBMS-17*, pages 19–24, 2017.

[73] Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[74] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[75] Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, pages 93–128, 2006.

[76] Qingming Tang, Siqi Sun, and Jinbo Xu. Learning scale-free networks by dynamic node specific degree prior. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2247–2255, 2015.

[77] Marshall F Tappen, Ce Liu, Edward H Adelson, and William T Freeman. Learning gaussian conditional random fields for low-level vision. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR'07)*, pages 1–8. IEEE, 2007.

[78] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[79] Eran Treister and Javier Turek. A block-coordinate descent approach for large-scale sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pages 927–935, 2014.

[80] Eran Treister, Javier Turek, and Irad Yavneh. A multilevel framework for sparse optimization with application to inverse covariance estimation and logistic regression. In *SISC*, 2016.

[81] Tijana Vujicic, Jesse Glass, Fang Zhou, and Zoran Obradovic. Gaussian conditional random fields extended for directed graphs. *Machine Learning*, pages 1–18, 2017.

[82] Min Wang, Jingjun Yan, Xingxing He, Qiang Zhong, Chengye Zhan, and Shusheng Li. Candidate genes and pathogenesis investigation for sepsis-related acute respiratory distress syndrome based on gene expression profile. *Biological research*, 49(1):1–9, 2016.

[83] Fabrice Wendling, Karim Ansari-Asl, Fabrice Bartolomei, and Lotfi Senhadji. From eeg signals to brain connectivity: a model-based evaluation of interdependence measures. *Journal of neuroscience methods*, 183(1):9–18, 2009.

[84] Matt Wytock and J Zico Kolter. Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *ICML*, pages 1265–1273, 2013.

[85] Chenyan Xiong, Taifeng Wang, Wenkui Ding, Yidong Shen, and Tie-Yan Liu. Relational click prediction for sponsored search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 493–502. ACM, 2012.

[86] Jieping Ye and Jun Liu. Sparse methods for biomedical data. *ACM SIGKDD Explorations Newsletter*, 14(1):4–15, 2012.

[87] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Biografija – Vladisav Jelisavčić

Vladisav Jelisavčić je rođen 17.07.1987. godine u Obrenovcu. Osnovne studije je završio na Elektrotehničkom fakultetu, smer Računarska tehnika i informatika, sa prosečnom ocenom 8.70. Diplomirao 2010. godine, uspešno odbranivši diplomski rad na temu: "Vizualni simulator ekspertskih sistema " kod profesora Boška Nikolića. Master studije je završio na istom smeru 2011., odbranivši master rad: "Evaluacija algoritama za modelovanje znanja na osnovu objavljenih radova" kod profesora Veljka Milutinovića. Doktorske studije je upisao 2011/2012. godine na modulu Računarska tehnika i informatika i položio je sve ispite predviđene Nastavnim planom i programom modula sa prosečnom ocenom 9.80.

Od 2012. godine je zaposlen kao istraživač u Matematičkom institutu Srpske akademije nauka i umetnosti. Trenutno je angažovan na projektu Ministarstva prosvete, nauke i tehnološkog razvoja, III44006: Razvoj novih informaciono-komunikacionih tehnologija korišćenjem naprednih matematičkih metoda, sa primenama u medicini, energetici, e-Upravi, telekomunikacijama i zaštiti nacionalne baštine.

Vladisav Jelisavčić je autor jednog M21 rada i koautor tri rada objavljena na međunarodnim konferencijama. Učestvovao je u izradi tri tehnička rešenja, kategorije M82 i M83.

Godine 2014. odlazi na stručno usavršavanje na Temple University, Philadelphia, Pennsylvania, USA. Tokom boravka učestvovao je na projektu Prospective Analysis of Large and Complex Partially Observed Temporal Social Networks, Defense Advanced Projects Agency, DARPA-GRAPHS, AFOSR award number FA 9550-12- 1-0406, gde je stekao iskustvo iz oblasti struktuiranog učenja.

Naučno-istraživačko iskustvo je stekao učestvovanjem u projektima Ministarstva za nauku i tehnološki razvoj (MNTR) Republike Srbije. Projekti su realizovani na Matematičkom institutu Srpske akademije nauka i umetnosti, ili u Inovacionom centru Elektrotehničkog fakulteta. Projekti pripadaju oblasti mašinskog učenja i data mininga. Bio je angažovan na EU FP7 projektu BALCON: Boosting EU-Western

Balkan Countries research collaboration in the Monitoring and Control area, GA: 288076. Takođe bio je angažovan kao istraživač na inovacionom projektu MPNTR Primena metoda za pronalaženje znanja nad velikom količinom podataka.

Takođe, poseduje i iskustvo rada na nekoliko softverskih projekata otvorenog koda vezanih za big data. Trenutno se vodi kao jedan od aktivnih komitera na Apache Ignite projektu za distribuirane softverske keš sisteme.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

СТРУКТУРАНО УЧЕЊЕ НАД ВЕЛИКИМ ПОДАЦИМА ЗАСНОВАНО НА ВЕРОВАТНОСНИМ ГРАФОВСКИМ МОДЕЛИМА

која је моје ауторско дело.
Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.
Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство
2. Ауторство - некомерцијално
3. Ауторство – некомерцијално – без прераде
4. Ауторство – некомерцијално – делити под истим условима
5. Ауторство – без прераде
6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 01.05.2018.

Владо Делић

Изјава о ауторству

Потписани-а Владислав Јелисавчић
број уписа 5033/11

Изјављујем

да је докторска дисертација под насловом

Структурално учење над великим подацима
засновано на вероватносним графовским моделима

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 01.05.2018

Владислав Јелисавчић

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Владисав Јелисавчић
Број уписа 5039/11
Студијски програм Рачунарска техника и информатика
Наслов рада Структурално учење над великим подацима засновано на вероватносним графовским моделима
Ментор проф. др. Бељко Милутиновић

Потписани Владисав Јелисавчић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.
Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 01.05.2018.

Владисав Јелисавчић

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.