UNIVERZITET U BEOGRADU

MATEMATIČKI FAKULTET

MASTER RAD

Varijaciono zaključivanje u bajesovskoj statistici

Autor:
Marko RADOSAVLJEVIĆ

Mentor:
doc. dr Bojana MILOŠEVIĆ

Sadržaj

1 Uvod

Pri kreiranju brojnih statističkih modela, nailazi se na problem ocene nepoznate gustine. Ovaj problem je naročito prisutan u bajesovskoj statistici, gde se ocena parametara svodi na nalaženje aposteriorne raspodele. Ocena gustine se značajno usložnjava sa porastom dimenzije domena gustine, odnosno, u kontekstu bajesovske statistike, sa porastom broja parametara koje je neophodno oceniti. Kako su savremeni modeli mahom određeni velikim brojem parametara, neophodno je pronaći alternativan način za ocene nepoznatih gustina. U ovom radu će biti predstavljen jedan od pristupa rešavanju problema aproksimacije gustine, koji najčešću primenu nalazi u modelima bajesovske statistike - *varijacioni* pristup. Pojam varijacionog računa nije nov u matematičkoj analizi, već datira iz 17. veka. Pod varijacionim računom se najčešće podrazumeva teorija optimizacije realnovrednosnih funkcionala. Počiva na diferenciranju funkcionala, tj. ispitivanjem kako mala promena funkcije iz domena funkcionala utiče na promenu vrednosti funkcionala. Pokazaćemo kako se ovaj pristup može oslikati u bajesovskoj statistici, svođenjem problema aproksimacije aposteriorne raspodele na optimizacioni problem.

Naglasimo da varijaciono zaključivanje nije metod koji se koristi isključivo u modelima bajesovske statistike, već u bilo kojoj situaciji gde je potrebno oceniti nepoznatu gustinu. Ipak, zbog integralnog značaja problema aproksimacije gustine u bajesovskoj statistici, posebno je važno izučiti osobine varijacionog zaključivanja pri kreiranju bajesovskih modela. Ovo se naročito odnosi na kompleksnije modele, koji su određeni velikim skupom parametara, jer u takvim slučajevima varijaciono zaključivanje tipično ima bržu konvergenciju i barem jednak kvalitet aproksimacije gustine, u poređenju sa drugim dostupnim metodima.

U kontekstu bajesovske statistike, varijacioni metodi, koji će biti opisani nadalje, problem nalaženja nepoznate aposteriorne raspodele svode na problem minimizacije funkcionala Kulbak-Lajblerovog (u daljem tekstu, KL) razilaženja, $KL(q||p)$, gde je p tražena aposteriorna raspodela, a q tzv. *varijaciona raspodela* po kojoj se vrši minimizacija, pretragom zadate familije varijacionih raspodela. Da bi ova minimizacija bila moguća, najčešće se pribegava ograničavanju prostora mogućih varijacionih raspodela, tj. parametrizaciji varijacione raspodele. Tada se iz skupa mogućih parametara biraju optimalni *varijacioni parametri*, koji odgovaraju varijacionoj raspodeli koja je najbliža traženoj aposteriornoj raspodeli, u smislu KL razilaženja. Ključan problem varijacionog zaključivanja jeste odabir familije varijacionih raspodela - šira familija, u teoriji, smanjuje grešku aproksimacije, ali značajno usporava i otežava proces optimizacije. U ovom radu ponudićemo nekoliko strategija za rešavanje ovog problema. Takođe, demonstriraćemo i primenu varijacionog zaključivanja na nekoliko modela bajesovske statistike, a zatim i uporediti ovaj metod sa nekim drugim metodima slične namene, poput onih zasnovanih na uzorkovanju. Konačno, zaključićemo u kojim problemima i u kom obliku treba koristiti varijaciono zaključivanje, a kada ga je pametno izbeći.

Struktura rada je sledeća: u poglavlju 2 će biti obrađen metod varijacionog zaključivanja u osnovnoj, i najkorišćenijoj, formi. Poglavlje 3 sadrži naprednije tehnike koje se primenjuju pri varijacionom zaključivanju i koje dozvoljavaju korišćenje opštijih familija varijacionih raspodela. Poglavlje 4 je pregled nekoliko statističkih modela, poput linearne regresije, parametarskih i neparametarskih modela mešavina. U svrhu primene varijacionog zaključivanja, ovi modeli će biti prikazani iz bajesovske perspektive. U primeni nekih od obrađenih modela (poput linearne regresije) aproksimacija aposteriorne raspodele varijacionim zaključivanjem nije uobičajena, ali su izabrani zbog svoje popularnosti i jednostavnosti, u svrhu jasne ilustracije metoda.

Poglavlje 5 demonstrira primenu varijacionog zaključivanja na obrađene modele, sa detaljno objašnjenim i implementiranim algoritmima. Poglavlje 6 daje osvrt na druge metode koji se koriste pri obučavanju bajesovskih modela, poput algoritma očekivanje-maksimizacija i Monte Karlo algoritma zasnovanog na Markovljevim lancima, i njihovu primenu na neke od već spomenutih bajesovskih statističkih modela. Ovo poglavlje je obogaćeno i implementacijom opisanih metoda u programskom jeziku R, kako bi se čitaocu približile mogućnosti upotrebe ovog jezika za ocenjivanje parametara bajesovskih statističkih modela. Rezultati uporedne analize ovih metoda dovešće nas do poglavlja 7, u kom ćemo konstatovati kvalitete i mane pojedinih varijanti varijacionog zaključivanja, kao i situacije u kojima su pokazale bolje rezultate od algoritama koji su u masovnijoj upotrebi, kao što je Monte Karlo algoritam zasnovan na Markovljevim lancima.

2 Varijaciono zaključivanje i aproksimacija srednjeg polja

Pre definisanja metoda varijacionog zaključivanja, podsetimo se pojma i osobina KL razilaženja, zbog njegove ključne uloge u aproksimaciji aposteriorne raspodele varijacionim metodima.

Neka su P i Q apsolutno neprekidne raspodele sa odgovarajućim gustinama p i q, definisanim na istom nosaču D. Tada se KL razilaženje definiše kao:

$$KL(P||Q) = \int_D p(x) \ln \frac{p(x)}{q(x)} dx. \tag{1}$$

Radi obuhvatanja slučaja diskretnih raspodela P i Q, u upotrebi je i opštija definicija KL razilaženja raspodela P i Q, definisanih na istom nosaču D:

$$KL(P||Q) = E_{x\sim P} \ln \frac{P(x)}{Q(x)}, \tag{2}$$

gde oznaka $x \sim P$ naglašava da se očekivanje računa po raspodeli P.

Navedimo ključna svojstva KL razilaženja koja će nam biti neophodna u razumevanju varijacionih metoda:

1. KL razilaženje između dve raspodele je nenegativno, i jednako je nuli ako i samo ako su dve raspodele skoro svuda jednake:

   $$-KL(P||Q) = E_p \ln \frac{q(x)}{p(x)} \leq \ln E_p \frac{q(x)}{p(x)} = \ln 1 = 0, \tag{3}$$

   gde nejednakost sledi iz Jensenove nejednakosti. Jednakost važi u slučaju $p = q$ skoro svuda, čime je tvrđenje pokazano.

2. KL razilaženje nije simetrično:

   $$KL(P||Q) \neq KL(Q||P). \tag{4}$$

   Odsustvo simetričnosti se može pokazati brojnim kontraprimerima. Jednostavnosti radi, prikazaćemo diskretan kontraprimer. Neka je $\Omega = \{A, B\}, P(A) = \frac{1}{5}, P(B) = \frac{4}{5}, Q(A) = Q(B) = \frac{1}{2}$. Tada je, po definiciji, $KL(P||Q) = P(A) \ln \frac{P(A)}{Q(A)} + P(B) \ln \frac{P(B)}{Q(B)} \approx 0.19$, a $KL(Q||P) = Q(A) \ln \frac{Q(A)}{P(A)} + Q(B) \ln \frac{Q(B)}{P(B)} \approx 0.22$.

Odavde možemo zaključiti da KL razilaženje nije metrika. Stoga se, iako se na njega može naići u literaturi, izbegava termin KL rastojanje.

2.1 Aproksimacija aposteriorne raspodele varijacionim metodima

Najpre, definišimo problem koji želimo da rešimo. Neka je X dati uzorak, a Z skup svih skrivenih (neopaženih) promenljivih, koje želimo da ocenimo. Skrivenim promenljivama se mogu smatrati svi neopaženi faktori koji na neki način utiču na raspodelu uzorka X. Neretko se izjednačavaju sa parametrima raspodele uzorka X, ali u pojedinim primerima ćemo demonstrirati razliku između ova dva pojma. Parametri raspodele uzorka jesu skrivene promenljive, ali obratno nije uvek slučaj. Za bajesovsku ocenu promenljivih Z, neophodno je odrediti aposteriornu raspodelu $p(Z|X)$, za koju, iz Bajesove teoreme, važi:

$$p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)}. \tag{5}$$

Jednom kada je poznata aposteriorna raspodela, za ocenu skrivenih promenljivih Z se uzima maksimum aposteriorne gustine - tzv. *MAP* ocena (eng. *Maximum a posteriori*). $p(X|Z)$ ćemo, u skladu sa standardnom terminologijom, često posmatrati kao funkciju skrivenih promenljivih, i zvati *verodostojnost*, dok ćemo raspodelu $p(Z)$ zvati *apriornom raspodelom*. Imenilac, $p(X)$, se obično naziva *verovatnoćom podataka* ili *marginalnom verodostojnošću* (eng. *evidence*). Bajesovski pristup, za razliku od frekventističkog, počiva na inicijalnim pretpostavkama o nekoj pojavi (sadržanim u apriornoj raspodeli), korigovanim opažanjem date pojave (sadržanom u verodostojnosti). Verodostojnost kazuje koliko opažena pojava odgovara našim inicijalnim pretpostavkama. Verovatnoća podataka sadrži informaciju o tome koliko je opažena pojava zaista verovatna - ako nije, možda ne treba da se oslanjamo na nju pri korekciji naših apriornih uverenja.

Upravo verovatnoća podataka zadaje najviše poteškoća pri ocenjivanju aposteriorne raspodele, naročito kod kompleksnijih modela, sa više skrivenih promenljivih, i u prisustvu masovnijih podataka. Nalaženje $p(X)$ obično zahteva izračunavanje kompleksnog integrala, koji često nema analitičko rešenje, dok su numeričke metode integracije u ovakvim slučajevima vremenski izuzetno zahtevne. Zato se pribegava aproksimaciji aposteriorne raspodele, pomoću varijacionog zaključivanja.

Ideja na kojoj se zasnivaju svi varijacioni metodi za ocenu gustine je odabir familije Q, koju nazivamo i *familijom varijacionih raspodela* ili *varijacionom familijom* (a elemente ove familije - *varijacionim raspodelama*), a onda i nalaženje one varijacione raspodele koja ima najmanje KL razilaženje od tražene aposteriorne raspodele p:

$$q^* = \arg\min_{q \in Q} KL(q||p). \tag{6}$$

Optimalna varijaciona raspodela q^* se zatim uzima za ocenu aposteriorne raspodele. Napomenimo da su varijacione raspodele definisane nad prostorom skrivenih promenljivih Z, kako bi ovakvo razmatranje KL razilaženja imalo smisla.

Prvi problem koji uočavamo je pitanje odabira dovoljno opšte optimalne familije varijacionih raspodela, tako da omogućava nalaženje precizne aproksimacije aposteriorne raspodele, a dovoljno jednostavne da optimizacioni postupak bude moguć. Ovaj problem će biti tema nekih od narednih poglavlja. Drugi, esencijalni, problem je činjenica da izračunavanje KL razilaženja između varijacione raspodele i aposteriorne raspodele zahteva izračunavanje verovatnoće podataka, $p(X)$, koje je, kao

što je rečeno, poteškoća zbog koje pribegavamo varijacionom zaključivanju.

U svrhu prevazilaženja problema nalaženja verovatnoće podataka, primetimo da važi:

$$\begin{aligned}\ln p(X) &= \ln \int_Z p(X,Z)dZ \\ &= \ln \int_Z q(Z)\frac{p(X,Z)}{q(Z)}dZ \\ &= \ln E_q \frac{p(X,Z)}{q(Z)} \\ &\geq E_q \ln \frac{p(X,Z)}{q(Z)} \\ &= E_q \ln p(X,Z) - E_q \ln q(Z),\end{aligned} \tag{7}$$

gde nejednakost sledi iz Jensenove nejednakosti. Dakle, izrazom $E_q \ln p(X,Z) - E_q \ln q(Z)$ data je donja granica logaritma verovatnoće podataka (eng. *ELBO - evidence lower bound*). Razmotrimo sada KL razilaženje između varijacione i aposteriorne raspodele:

$$\begin{aligned}KL(q(Z)||p(Z|X)) &= \int_Z q(Z) \ln \frac{q(Z)}{p(Z|X)}dZ \\ &= \int_Z q(Z) \ln \frac{q(Z)p(X)}{p(X,Z)}dZ \\ &= \int_Z q(Z)(\ln q(Z) + \ln p(X) - \ln p(X,Z))dZ \\ &= E_q \ln q(Z) + E_q \ln p(X) - E_q \ln p(X,Z) \\ &= \ln p(X) - (E_q \ln p(X,Z) - E_q \ln q(Z)).\end{aligned} \tag{8}$$

Kako $\ln p(X)$ ne zavisi od varijacione raspodele q, važi $E_q \ln p(X) = \ln p(X)$. Iz istog razloga, minimizacijom funkcionala $KL(q(Z)||p(Z|X))$ po raspodeli q indirektno se maksimizuje funkcional $ELBO(q) = E_q \ln p(Z,X) - E_q \ln q(Z)$. Stoga se problem minimizacije KL razilaženja između varijacione i aposteriorne raspodele može adekvatno zameniti problemom maksimizacije donje granice verovatnoće podataka. Primetimo još da je zbir $KL(q||p)$ i $ELBO(q)$ jednak logaritmu marginalne verodostojnosti, tj. upravo funkciji koju $ELBO$ ograničava. Kako je, iz (3), KL razilaženje nenegativno, i $KL(q(Z)|p(Z|X)) = 0$ za $q(Z) = p(Z|X)$ skoro svuda, maksimalna vrednost $ELBO$-a je $\ln p(X)$,i to onda kada familija varijacionih raspodela sadrži traženu aposteriornu raspodelu.

Uočimo i da važi:

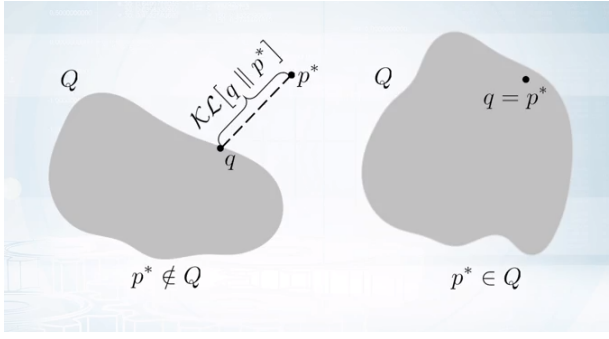
$$\begin{aligned}ELBO(q) &= E_q \ln p(X,Z) - E_q \ln q(Z) \\ &= E_q \ln p(X|Z) + E_q \ln p(Z) - E_q \ln q(Z) \\ &= E_q \ln p(X|Z) - E_q \ln \frac{q(Z)}{p(Z)} \\ &= E_q \ln p(X|Z) - KL(q(Z)||p(Z)).\end{aligned} \tag{9}$$

Sada je jasnija još jedna interpretacija donje granice verovatnoće podataka. Prvi sabirak predstavlja očekivanu verodostojnost podataka, i maksimizuje se maksimizacijom $ELBO$, dok drugi sabirak predstavlja negativno KL razilaženje varijacione

raspodele od apriorne raspodele, koji utiče na to da optimalna varijaciona raspodela, a time i ocena aposteriorne raspodele, ne bude daleko od apriorne raspodele u smislu KL razilaženja. Dakle, maksimizacija donje granice verovatnoće podataka predstavlja nalaženje balansa između verodostojnosti i apriorne raspodele, kao što je i uobičajeno u bajesovskom ocenjivanju.

2.2 Izbor familije varijacionih raspodela

Izbor familije varijacionih raspodela predstavlja najbitniji korak u procesu varijacionog zaključivanja. U idealnom slučaju, ukoliko familija varijacionih raspodela Q sadrži i aposteriornu raspodelu p, izraz $KL(q||p)$ minimum dostiže u nuli, te je rešenje maksimizacije donje granice verovatnoće podataka upravo tražena aposteriorna raspodela. Slika 1 (preuzeto iz [28]) ilustruje dva scenarija izbora varijacione familije.



Slika 1: KL razilaženje optimalne varijacione raspodele q i aposteriorne raspodele p^*, $KL(q||p^*)$ u slučaju kada aposteriorna raspodela ne pripada (levo) i pripada (desno) varijacionoj familiji

Opšti recept za optimalan izbor varijacione familije ne postoji. Generalno govoreći, izbor šire familije smanjuje mogućnost greške i povećava šansu pripadanja aposteriorne raspodele familiji, ali otežava izračunavanje donje granice verovatnoće podataka i dovodi do sporijih performansi. U narednim poglavljima, biće predstavljen jedan vid kompromisa između širine familije i kompleksnosti optimizacionog postupka.

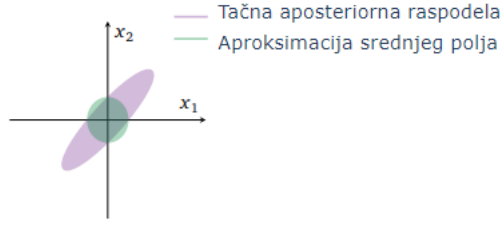
2.3 Aproksimacija srednjeg polja

Jedan od najkorišćenijih metoda za izbor familije varijacionih raspodela jeste aproksimacija srednjeg polja, koja se zasniva na pretpostavci da se varijaciona raspodela može faktorizovati, tj. pretpostavci o nezavisnosti skrivenih promenljivih. U najvećem broju slučajeva, ova pretpostavka nije zadovoljena za aposteriornu raspodelu, pa ovako izabrana familija Q ne sadrži aposteriornu raspodelu, te se aproksimacijom srednjeg polja teško dolazi do stvarne aposteriorne raspodele. Ipak, ovakva postavka omogućava maksimizaciju donje granice verovatnoće podataka po koordinatama i time znatno ubrzava optimizacioni postupak.

Dakle, pretpostavka je da za svaku raspodelu iz varijacione familije raspodela Q važi:

$$q(Z) = \prod_{i=1}^{d} q_i(Z_i), \tag{10}$$

gde je d broj skrivenih promenljivih koje se ocenjuju. Ove marginalne raspodele $q_i(Z_i)$ se još nazivaju i *varijacionim faktorima*. Svaki od varijacionih faktora se potom parametrizuje proizvoljno, u zavisnosti od svojstava skrivene promenljive kojoj odgovara. Promenljive po kojima se optimizacija vrši su upravo parametri raspodela q_i, koji se nazivaju i *varijacionim parametrima*. Faktorizacijom varijacione raspodele, metod srednjeg polja može proizvoljno dobro aproksimirati marginalne raspodele skrivenih promenljivih, ali gubi mogućnost prepoznavanja njihovih korelacija. Ovo se može videti na sledećem primeru:



Slika 2: Tačna aposteriorna raspodela i aproksimacija srednjeg polja

Na slici 2 (preuzeto iz [3]) su prikazani podaci generisani iz dvodimenzione normalne raspodele, sa visokom korelacijom između slučajnih veličina X_1 i X_2 i aproksimacija metodom srednjeg polja sa pretpostavkom normalnih marginalnih raspodela, dobijena maksimizacijom donje granice verovatnoće podataka. Možemo videti da je aproksimacija srednjih vrednosti precizna, disperzije aproksimativnih marginalnih raspodela niža od disperzija marginalnih raspodela aposteriorne (o čemu će biti reči nešto kasnije), i da ne postoji korelacija između marginalnih slučajnih veličina aproksimativne raspodele.

Iz prikazanog primera jasno je da metod srednjeg polja donosi određena ograničenja varijacionom zaključivanju. Sa druge strane, omogućava koordinatnu optimizaciju ELBO (koja će biti tema sledećeg poglavlja), te aproksimacija srednjeg polja omogućava lako nalaženje ocene aposteriorne raspodele, koja u određenim slučajevima može imati zadovoljavajuću preciznost.

2.4 Koordinatna optimizacija varijacionih parametara

Kao što znamo iz (7), važi:

$$ELBO(q) = E_q \ln p(Z, X) - E_q \ln q(Z).$$

Pođimo od pretpostavke aproksimacije srednjeg polja o nezavisnosti skrivenih promenljivih (10). Ako fiksiramo sve varijacione faktore osim jednog, q_j, možemo

predstaviti donju granicu verovatnoće podataka kao funkciju od q_j:

$$\begin{aligned} ELBO(q_j) &= E_q \ln P(Z,X) - E_q \ln q(Z) \\ &= E_j(E_{-j} \ln p(Z,X)) - E_j(E_{-j} \ln q(Z)) \\ &= E_j(E_{-j} \ln p(Z,X)) - E_j \ln q_j(Z_j) + const \\ &= E_j(\ln e^{E_{-j} \ln p(Z,X)} - \ln q_j(Z_j)) + const \\ &= -KL(q_j(Z_j), e^{E_{-j} \ln p(Z,X)}) + const, \end{aligned} \tag{11}$$

gde E_j označava očekivanje po j-tom varijacionom faktoru, a E_{-j} očekivanje po svim varijacionim faktorima sem j-tog. Maksimizacija ovog izraza ekvivalentna je minimizaciji KL razilaženja raspodela $q_j(Z_j)$ i $e^{E_{-j} \ln p(Z,X)}$. KL razilaženje dostiže minimum u nuli, pa se $ELBO$ maksimizuje za $q_j(Z_j) = e^{E_{-j} \ln p(Z,X)}$. Očito je da izračunavanje svakog varijacionog faktora zavisi od svih ostalih, te ovim nije dato eksplicitno rešenje optimizacionog problema, već se do rešenja dolazi iterativno, sa definisanim kriterijumom zaustavljanja.

Prema tome, koordinatna optimizacija varijacionih parametara kod metoda srednjeg polja se sprovodi kroz nekoliko koraka:

1. Inicijalizacija varijacionih faktora (tj. parametara varijacionih faktora).
2. Izračunavanje donje granice verovatnoće podataka $ELBO(q) = E_q \ln p(Z,X) - E_q \ln q(Z)$.
3. Ažuriranje varijacionih faktora, za svako $j \in \{1, 2, \ldots, d\}$, po pravilu $q_j(Z_j) = e^{E_{-j} \ln p(Z,X)}$.

Koraci 2 i 3 se ponavljaju naizmenično, do ispunjenja kriterijuma zaustavljanja. U praksi, optimizacija se obično zaustavlja postizanjem konvergencije donje granice verovatnoće podataka, ali su mogući i drugi kriterijumi (konvergencija varijacionih parametara ili zadat broj iteracija procesa).

Kako $p(Z_{-j}, X)$ ne zavisi od Z_j, zbog pretpostavke aproksimacije srednjeg polja, važi:

$$\begin{aligned} e^{E_{-j} \ln p(Z,X)} &= e^{E_{-j} \ln p(Z_j|Z_{-j},X)p(Z_{-j},X))} \\ &\propto e^{E_{-j} \ln p(Z_j|Z_{-j},X)}, \end{aligned} \tag{12}$$

što daje još jedan oblik pravila ažuriranja varijacionih parametara koordinatnom optimizacijom:

$$q_j(Z_j) \propto e^{E_{-j} \ln p(Z_j|Z_{-j},X)}. \tag{13}$$

Napomenimo još da se ovim metodom nalazi lokalni maksimum donje granice verovatnoće podataka. Ovakvih maksimuma može biti mnogo, te se preporučuje testiranje nekoliko različitih inicijalizacija varijacionih parametara (parametara raspodela $q_j(Z_j)$) jer konačna ocena može umnogome zavisiti od inicijalnih vrednosti. U tom slučaju, treba izabrati ocenu izvedenu iz najvećeg lokalnog maksimuma ELBO funkcije, koji odgovara minimalnom KL razilaženju između aproksimativne i stvarne aposteriorne raspodele.

Treba imati u vidu i da ažuriranje varijacionih parametara tako da važi (13) u pojedinim slučajevima zahteva netrivijalne kalkulacije, i da je ovaj korak najizazovnija prepreka u implementaciji aproksimacije srednjeg polja.

2.5 Varijaciono zaključivanje kod eksponencijalne familije raspodela

Razmotrimo sada aproksimaciju srednjeg polja i koordinatnu optimizaciju varijacionih parametara kod modela kod kojih su sve uslovne raspodele $p(Z_j|Z_{-j}, X)$ raspodele iz eksponencijalne familije (npr. normalna, eksponencijalna, gama, beta, Bernulijeva, Dirihleova raspodela, itd), tj. za svako Z_j važi:

$$p(Z_j|Z_{-j}X) = h(Z_j)e^{\eta_j(Z_{-j},X)^T Z_j - \alpha(\eta_j(Z_{-j},X)}. \tag{14}$$

Iz (13) sledi da se koordinatnom optimizacijom varijacioni parametri ažuriraju po pravilu $q_j(Z_j) = e^{E_{-j} \ln p(Z_j|Z_{-j},X)}$. Primenjeno na slučaj uslovnih raspodela iz eksponencijalne familije, ovo pravilo dobija oblik:

$$\begin{aligned} q_j(Z_j) &= e^{E_{-j} \ln p(Z_j|Z_{-j},X)} \\ &= e^{\ln h(Z_j) + E_{-j}(\eta_j(Z_{-j},X)^T)Z_j - E(\alpha(\eta_j(Z_{-j},X))} \\ &\propto h(Z_j)e^{E_{-j}(\eta_j(Z_{-j},X))^T Z_j}, \end{aligned} \tag{15}$$

te optimalna varijaciona raspodela takođe pripada eksponencijalnoj familiji.

Neka je varijaciona raspodela q_j parametrizovana varijacionim parametrima ν_j. Za novu vrednost varijacionih parametara se uzima očekivana vrednost varijacione raspodele dobijene iz (10), tj:

$$\nu_j = E_{-j}\eta_j(Z_{-j}, X). \tag{16}$$

Ovaj vid ažuriranja varijacionih parametara je vrlo primenljiv u praktično svim slučajevima gde uslovne raspodele $p(Z_j|Z_{-j}, X)$ pripadaju eksponencijalnoj familiji raspodela.

2.6 Kvaliteti i mane aproksimacije srednjeg polja

Prethodno je opisan metod varijacionog zaključivanja, i njegova najkorišćenija forma, aproksimacija srednjeg polja. Koje su prednosti, a koje mane aproksimacije srednjeg polja, i kada je treba koristiti?

Pretpostavka na kojoj se zasniva aproksimacija srednjeg polja jeste pretpostavka o nezavisnosti varijacionih parametara. Očigledno, ova pretpostavka je ograničavajuća, i u slučajevima kada je korelacija među skrivenim varijablama izražena, aproksimacija srednjeg polja nije pogodan izbor. Jedan ovakav primer je ilustrovan na slici 2, u poglavlju 2.3. Napomenimo da se pretpostavka o nezavisnosti može zameniti blažom pretpostavkom, gde se ne zahteva potpuna nezavisnost skrivenih promenljivih, već postojanje grupa, takvih da su sve skrivene promenljive unutar jedne grupe nezavisni od ostalih. Ovako ublažena pretpostavka računski ne otežava značajno proces ažuriranja varijacionih parametara, i oslanja se na varijantu koordinatne optimizacije (parametri iste grupe se ažuriraju zajedno), ali omogućava proširenje familije varijacionih raspodela, dozvoljavajući korelaciju unutar iste grupe. Ovaj metod je poznat i kao *strukturirana aproksimacija srednjeg polja*, i biće centralna tema poglavlja 3.1.

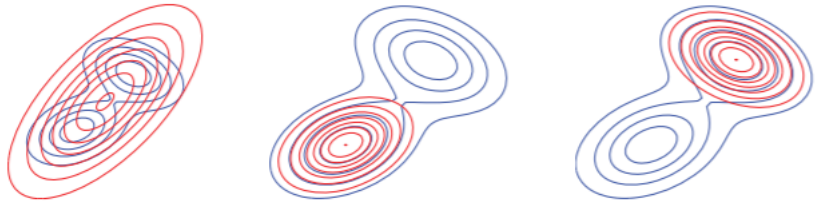
U spomenutom primeru (slika 2, poglavlje 2.3) je spomenuto da aproksimacija srednjeg polja teži potcenjivanju disperzije, tj. nije neuobičajeno da je disperzija optimalne varijacione raspodele niža od disperzije stvarne aposteriorne raspodele. Ovaj problem doseže do srži metoda varijacionog zaključivanja, tj. KL razilaženja. Kako je:

$$KL(q||p) = \int_D q(z) \ln \frac{q(z)}{p(z)} dZ, \tag{17}$$

minimizacija razilaženja $KL(q||p)$ pre svega teži da u regionima gde je raspodela p niska, postavi i raspodelu q nisko. Ovaj fenomen potiče od činjenice da je p u imeniocu, te bi u slučaju velike vrednosti raspodele q vrednosti raspodele p bliske nuli u nekom regionu, vrednost KL razilaženja drastično porasla. Zbog toga se procesom minimizacije KL razilaženja, raspodela q "prisiljava" da svu svoju težinu stavlja u okolinu moda raspodele p, a da je ne rasipa u regione gde raspodela p ima male vrednosti. Ovaj efekat za posledicu ima potcenjivanje stvarne disperzije aposteriorne raspodele, ali, sa druge strane, omogućava aproksimaciji srednjeg polja vrlo dobre performanse i dobru aproksimaciju barem jedne mode kod multimodalnih aposteriornih raspodela. Još jedna posledica minimizacije funkcionala $KL(q||p)$ je tzv. sužavanje domena aposteriorne raspodele, tj. u regionima gde je raspodela p jednaka nuli, i raspodela q mora biti jednaka nuli, kako bi KL razilaženje bilo konačno. Ako su ovo efekti upotrebe tzv. obrnutog razilaženja $KL(q||p)$, postavlja se pitanje šta ako se aproksimacija vrši minimizacijom razilaženja *unapred*, odnosno $KL(p||q)$? Na ovome se zasniva novi pristup aproksimaciji aposteriorne raspodele, poznat i kao *propagacija očekivanja* (v. [17]). Premda ovaj algoritam izlazi iz opsega ovog rada, nije zgoreg razmotriti manifestaciju pomenutih fenomena prilikom "obrtanja" razilaženja od interesa. Kako je:

$$KL(p||q) = \int_D p(Z) \ln \frac{p(Z)}{q(Z)} dZ, \tag{18}$$

minimizacija se ostvaruje kroz prisiljavanje varijacione raspodele da nema male vrednosti tamo gde tražena raspodela ima velike. Time se dobija proširenje domena aposteriorne raspodele, i nepoželjno raspoređivanje težine varijacione raspodele kod multimodalnih aposteriornih raspodela. Sledeća slika (preuzeto iz [18]) ilustruje razliku između ova dva pristupa:



Slika 3: Efekat upotrebe razilaženja unapred i obrnutog razilaženja u slučaju bimodalne raspodele. Plavom bojom su označene konture stvarne aposteriorne, a crvenom konture optimalne varijacione raspodele. a) Minimizacija $KL(p||q)$ dovodi do raspršenja težine između dveju moda. b,c) Minimizacija $KL(q||p)$ fokusira varijacionu raspodelu na jednu od moda

Konačno, konstatovan je efekat izbora inicijalnih vrednosti parametara varijacionih faktora na rešenje optimizacionog problema. Nažalost, ne postoje inteligentne strategije odabira inicijalnih vrednosti, već se obično biraju proizvoljne vrednosti, tako da parametri svakog varijacionog faktora odgovaraju apriornim uverenjima o skrivenoj promenljivi kojoj odgovara (v. [20, 27]), što će biti ilustrovano u nastavku rada, kroz primere.

3 Naprednije tehnike i primene varijacionog zaključivanja

Prethodno poglavlje opisuje varijaciono zaključivanje i njegovu najjednostavniju formu - metod aproksimacije srednjeg polja. Konstantovano je da se koordinatna optimizacija jednostavno implementira, i dovodi do lokalnog maksimuma ELBO-a, ali i da aproksimacija srednjeg polja ima svoje brojne nedostatke, naročito u smislu obuhvatnosti varijacione familije na kojoj počiva. Nadalje ćemo prikazati nekoliko naprednijih formi varijacionog zaključivanja, koje ne postavljaju tako striktne prepostavke (poput potpune nezavisnosti skrivenih promenljivih) kao aproksimacija srednjeg polja. Obradićemo i alternativni način optimizacije ciljne funkcije varijacionog zaključivanja - stohastičku optimizaciju. Mada će biti opisana za aproksimaciju srednjeg polja, stohastička optimizacija se može primeniti i na opštije konfiguracije varijacionog zaključivanja, unapređujući brzinu konvergencije metoda. Napomenimo da se na pojedinim mestima neće ulaziti u detalje nekih od metoda, zbog njihove kompleksnosti, već će se insistirati na pojašnjenju principa po kojima funkcionišu.

3.1 Strukturirano varijaciono zaključivanje

Ograničenja koja postavlja pretpostavka o potpunoj nezavisnosti skrivenih promenljivih, na kojoj se zasniva aproksimacija srednjeg polja, su do sada dovoljno apostrofirana. Metod koji se prirodno nastavlja na aproksimaciju srednjeg polja, relaksirajući pretpostavku o nezavisnosti, je *strukturirano varijaciono zaključivanje*.

Strukturirano varijaciono zaključivanje polazi od pretpostavke o postojanju strukture zavisnosti između skrivenih promenljivih, i mogućnosti particionisanja skupa skrivenih promenljivih na grupe koje su međusobno nezavisne. Neka je $Z_{1:n}$ skup skrivenih promenljivih i $\mathcal{P} = \{P_1, \ldots, P_c\}$ particionisanje ovog skupa, gde P_i označava skup indeksa skrivenih promenljivih koje pripadaju i-toj particiji. Uz pretpostavku o međusobnoj nezavisnosti particija, varijaciona raspodela se faktorizuje na sledeći način:

$$q(Z) = \prod_{i=1}^{c} q_i(Z_{P_i}), \tag{19}$$

gde je Z_{P_i} i-ta particija skupa skrivenih promenljivih.

Kao što je i spomenuto u poglavlju 2.6, strukturirano varijaciono zaključivanje se oslanja na postupak sličan koordinatnoj optimizaciji kod aproksimacije srednjeg polja, gde se istovremeno ažuriraju varijacioni parametri koji odgovaraju skrivenim promenljivama iste particije.

Fiksiranjem svih varijacionih faktora osim onih u particiji P_i, posmatrajmo $ELBO$ (7) kao funkciju varijacionih faktora q_{P_i}:

$$\begin{aligned} ELBO(q_{P_i}) &= E_{P_i}(E_{-P_i} \ln p(Z, X)) - E_{P_i}(E_{-P_i} \ln q(Z)) \\ &= E_{P_i}(E_{-P_i} \ln p(Z, X)) - E_{P_i} \ln q_{P_i}(Z_{P_i}) + const \\ &= E_{P_i}(\ln e^{E_{-P_i} \ln p(Z,X)} - \ln q_{P_i}(Z_{P_i})) + const \\ &= -KL(q_{P_i}(Z_{P_i}), e^{E_{-P_i} \ln p(Z,X)}) + const. \end{aligned} \tag{20}$$

Slično koordinatnoj optimizaciji aproksimacije srednjeg polja, $ELBO$ se maskimizuje za $q_{P_i} = e^{E_{-P_i} \ln p(Z,X)}$, te je pravilo ažuriranja varijacionih faktora

strukturiranog varijacionog zaključivanja identično pravilu koje prati koordinatna optimizacija aproksimacije srednjeg polja:

1. Definisanje particionisanja skupa skrivenih promenljivih $\mathcal{P} = \{P_1, \ldots, P_c\}$
2. Inicijalizacija varijacionih faktora $q_{1:c}$ (tj. parametara varijacionih faktora).
3. Izračunavanje donje granice verovatnoće podataka $ELBO(q) = E_q \ln p(Z, X) - E_q \ln q(Z)$.
4. Ažuriranje varijacionih faktora, za svako $i \in \{1, 2, \ldots, C\}$, po pravilu $q_{P_i}(Z_{P_i}) = e^{E_{-P_i} \ln p(Z,X)}$,

ponavljanjem koraka 3 i 4 do konvergencije $ELBO$.

Strukturirano varijaciono zaključivanje nudi unapređenje u odnosu na aproksimaciju srednjeg polja, zahvaljujući odustajanju od pretpostavke o potpunoj nezavisnosti skrivenih promenljivih, i mogućnosti aproksimacije korelacije među skrivenim promenljivama iste particije. Iako proces koordinatne optimizacije na prvi pogled deluje ekvivalentno kao u poglavlju 2.4, napomenimo da dimenzionalnost integrala koji se rešavaju u procesu ažuriranja varijacionih parametara raste sa većim particijama, što otežava ovaj, ionako netrivijalan, korak. Naravno, kao i u slučaju aproksimacije srednjeg polja, opasnost dostizanja lokalnog, a ne i globalnog, maksimuma postoji, te se savetuje optimizacija sa različitim inicijalnim vrednostima.

Uočavanje particija u skupu skrivenih promenljivih je, u većini slučajeva, trivijalno, i postoje modeli gde pretpostavka o međusobnoj nezavisnosti particija nije besmislena, te mogućnost pripadanja stvarne aposteriorne raspodele varijacionoj familiji strukturiranog varijacionog zaključivanja nije zanemarljiva. Jedan tipičan primer podele skrivenih promenljivih na particije je uočljiv u slučaju linearne regresije, gde se ocenjuju očekivanja i disperzije parametara koji učestvuju u linearnom modelu. Prirodno, parametri očekivanja se uzimaju za jednu particiju, a parametri disperzije za drugu. Ovakva, ali i drugačije podele, će biti prikazane u poglavljima 4 i 5, kroz primere.

3.2 Stohastičko varijaciono zaključivanje

Iako je, prethodno opisana, koordinatna optimizacija varijacionih parametara jednostavna za implementaciju, može biti jako spora, naročito u slučaju jako velikog obima uzorka X. Kako savremene primene najčešće zahtevaju obradu masovnih podataka, od koordinatne optimizacije se često odustaje, i pristupa se stohastičkoj optimizaciji ELBO-a. Stohastička optimizacija je poznat optimizacioni koncept, a čestu primenu nalazi u obučavanju kompleksnih modela mašinskog učenja koristeći veliku količinu podataka (poput neuronskih mreža), kod kojih je obračunavanje gradijenta vremenski izuzetno zahtevno. Pojmom stohastičke optimizacije obuhvaćene su brojne tehnike koje u optimizaciji generišu i koriste slučajne promenljive. Neki od primera upotrebe slučajnosti u optimizaciji su zamena gradijenta njegovom nepristrasnom ocenom (o kojoj će uskoro biti reči), optimizacija zasnovana na Monte Karlo simulacijama, i drugi. Ovakvi metodi obično svaki optimizacioni korak zamenjuju drugim, značajno jednostavnijim za izračunavanje, tako da se, uz povećanje iteracija postupka, konvergira ka istom rešenju.

U smislu varijacionog zaključivanja, od posebnog interesa je optimizacija stohastičkim gradijentom. Primenljiva je na svim optimizacionim problemima, gde se

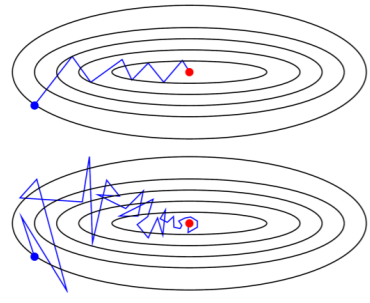
ciljna funkcija može predstaviti u obliku proseka jednostavnijih funkcija:

$$L(Z) = \frac{1}{N}\sum_{i=1}^{N} L_i(Z). \tag{21}$$

Iako se ova pretpostavka na prvi pogled može učiniti strogom, veliki broj problema moderne statistike se svodi na maksimizaciju ili minimizaciju ovakvih funkcija. Takva funkcija je i funkcija verodostojnosti, ili proizvoljna funkcija greške koja je definisana na pojedinačnim elementima uzorka (zaista, u većini primena, sumiranje iz (21) se vrši po elementima uzorka). Klasična gradijentna optimizacija u ovakvim primerima može biti vremenski skup proces, posebno za velike vrednosti N. Zato se, u stohastičkoj gradijentnoj optimizaciji, umesto gradijenta funkcije $L(Z)$, za ažuriranje parametara Z koristi njegova nepristrasna ocena, gradijent funkcije $L_i(Z)$, gde je i indeks slučajno izabranog elementa uzorka, tj. $P(i = k) = \frac{1}{N}$, za $k \in \{1, 2, \ldots, N\}$. Nepristrasnost se pokazuje jednostavno:

$$\begin{aligned} E\nabla L_i(Z) = \nabla E L_i(Z) &= \nabla \sum_{k=1}^{N} P(i=k) L_k(Z) \\ &= \nabla \sum_{k=1}^{N} \frac{1}{N} L_k(Z) = \nabla L(Z). \end{aligned} \tag{22}$$

Ova ocena je dobijena pomoću samo jedne komponente ciljne funkcije, što je čini daleko jednostavnijom za kalkulaciju. Tipično, stohastičku optimizaciju karakterišu haotično kretanje ka optimumu funkcije, i veći broj iteracija, u poređenju sa klasičnom gradijentnom optimizacijom, ali su vremenske uštede ostvarene upotrebom stohastičkog gradijenta impozantne. Sledećom slikom (preuzeto iz [19]) ilustrovane su tipične putanje gradijenta i stohastičkog gradijenta:



Slika 4: Tipično ponašanje klasične (gore) i stohastičke (dole) gradijentne optimizacije

Radi opisivanja stohastičke optimizacije ELBO, preformulisaćemo inicijalnu postavku, definicijom dve klase skrivenih promenljivih - lokalne i globalne. Najpre, za dati uzorak $X_{1:N}$, lokalnim skrivenim promenljivama $Z_{1:N}$ se smatraju one koje su vezane za pojedinačne tačke uzorka, dok se globalne skrivene promenljive β odnose na čitav uzorak. Preciznija diferencijacija između ovih klasa data je uslovnom nezavisnošću para X_i, Z_i od ostalih lokalnih skrivenih promenljivih i elemenata uzorka, kada su poznati globalne skrivene promenljive:

$$P(X_i, Z_i | X_{-i}, Z_{-i}, \beta) = P(X_i, Z_i | \beta). \tag{23}$$

Postojanje ovih tipova skrivenih promenljivih nije neuobičajeno u bajesovskih statističkim modelima, i biće očigledno u četvrtom poglavlju, u modelima Gausovih mešavina. Nad globalnim promenljivama se postavlja apriorna raspodela, i cilj je pronaći njihovu aposteriornu raspodelu. Tada se zajednička raspodela podataka i skrivenih promenljivih može razložiti na globalni činilac i proizvod lokalnih:

$$P(X, Z, \beta) = P(\beta) \prod_{i=1}^{N} P(X_i, Z_i|\beta). \tag{24}$$

Primenimo li ovako definisane skrivene promenljive u definiciju ciljne funkcije varijacionog zaključivanja, ELBO, dobijamo:

(25)
$$\begin{aligned} ELBO(q) &= E_q \ln p(X, Z, \beta) - E_q \ln q(Z, \beta) \\ &= E_q \ln p(X, Z|\beta) + E_q \ln p(\beta) - E_q \ln q(Z|\beta) - E_q \ln q(\beta) \\ &= E_q \ln p(\beta) - E_q \ln q(\beta) + \sum_{i=1}^{N} (E_q \ln p(X_i, Z_i|\beta) - E_q \ln q(Z_i)), \end{aligned}$$

gde je razvoj u sumu moguć zahvaljujući uslovnoj nezavisnosti (23) i pretpostavci aproksimacije srednjeg polja (10). Ovime je ciljna funkcija razložena na sabirak koji zavisi od globalnih promenljivih, i sumu koja zavisi od lokalnih promenljivih.

Optimizacija ovako definisane ciljne funkcije se vrši iterativno, naizmeničnim ažuriranjem varijacionih faktora koji odgovaraju globalnim, i varijacionih faktora koji odgovaraju lokalnim promenljivama. Duh stohastičke optimizacije se ogleda u ažuriranju globalnih promenljivih, kada se, za fiksirane lokalne varijacione faktore, desna suma zamenjuje jednim, slučajno izabranim, sabirkom. Egzaktna pravila ažuriranja lokalnih i globalnih varijacionih faktora prevazilaze opsege ovog rada, te se znatiželjni čitalac upućuje na [11].

3.3 Neparametarsko varijaciono zaključivanje

Neparametarsko varijaciono zaključivanje predstavlja strategiju odabira varijacione familije koja možda najbolje prikazuje bogatstvo teorije varijacionog zaključivanja. Ova strategija je motivisana vrlo popularnim metodom neparametarske statistike - ocenom gustine metodom jezgara (kernela), i sa lakoćom je inkorporira u postavku varijacionog zaključivanja. Suština je ocenjivanje aposteriorne raspodele metodom jezgara, tretirajući parametre jezgara kao varijacione parametre, koji se ocenjuju opisanim varijacionim metodima. Sa povoljnim odabirom oblika i dovoljno velikim brojem jezgara, ovako definisana familija može biti proizvoljno široka, te omogućava dobru aproksimaciju aposteriorne raspodele, koliko god kompleksan njen oblik bio. Ipak, optimizacioni proces nije jednostavan kao u slučaju aproksimacije srednjeg polja, što ćemo demonstrirati u nastavku.

Razmotrićemo slučaj neparametarskog varijacionog zaključivanja sa Gausovim jezgrima - varijacionu familiju definišemo kao familiju Gausovih mešavina:

$$q(Z) = \frac{1}{N} \sum_{n=1}^{N} g(Z; \mu_n, \sigma_n^2 I), \tag{26}$$

gde $g(Z; \mu_n, \sigma_n^2 I)$ označava gustinu normalne raspodele sa očekivanjem μ_n i kovarijacionom matricom $\sigma_n^2 I$.

Parametre μ_n i σ_n^2, očekivanje i disperziju n-te komponente mešavine, posmatramo kao varijacione parametre, i minimizacijom ELBO-a se vrši njihovo ocenjivanje. Naglasimo da ovaj slučaj odgovara neprekidno raspodeljenim skrivenim promenljivama Z. U slučaju diskretnih skrivenih promenljivih, nužno je izabrati drugi oblik jezgra. Prisetimo se da je, iz (7):

$$ELBO(q) = E_q \ln p(Z, X) - E_q \ln q(Z) = E_q \ln p(Z, X) + H(q), \tag{27}$$

gde je $H(q)$ entropija raspodele q.

Kada je q raspodela iz familije Gausovih mešavina, očekivanja $E_q \ln p(Z, X)$ i $E_q \ln q(Z)$ u opštem slučaju nemaju analitičko rešenje (v. [8]). Zato se, u ovom slučaju, pristupa novoj izmeni optimizacionog problema. Najpre se nalazi donja granica entropije $H(q)$:

$$\begin{aligned} H(q) &= -\int_D q(Z) \ln q(Z) dZ \\ &= -\int_D q(Z) \ln \frac{1}{N} \sum_{n=1}^{N} g(Z; \mu_n, \sigma_n^2 I) dZ \\ &\geq -\frac{1}{N} \sum_{n=1}^{N} \ln \int_D q(Z) g(Z; \mu_n, \sigma_n^2 I) dZ, \end{aligned} \tag{28}$$

gde je D domen skrivenih promenljivih Z.

Svaki integral u sumi sa desne strane nejednakosti je suma N konvolucija dve Gausove raspodele, koja je takođe Gausova raspodela, te je:

$$H(q) \geq -\frac{1}{N} \sum_{n=1}^{N} \ln q_n, \tag{29}$$

gde je $q_n = \frac{1}{N} \sum_{j=1}^{N} g(\mu_n; \mu_i, (\sigma_n^2 + \sigma_j^2) I)$.

Ostalo je još aproksimirati $E_q \ln p(Z, X)$. Označimo sa $f(Z) = p(Z, X)$. Važi:

$$E_q \ln f(Z) = \frac{1}{N} \sum_{n=1}^{N} \int_D g(Z; \mu_n, \sigma_n^2 I) f(Z) dZ. \tag{30}$$

Aproksimirajmo svaki sabirak sume pomoću razvoja Tejlorovog reda funkcije $f(Z)$ u okolini tačke μ_n:

$$f(Z) \approx \hat{f}_n(Z) = f(\mu_n) + \nabla f(\mu_n)(Z - \mu_n) + \frac{1}{2}(Z - \mu_n)^T \mathbf{H}_n (Z - \mu_n), \tag{31}$$

gde je $\mathbf{H}_n = \nabla_Z^2 f(Z)$ Hesijan, matrica drugih izvoda funkcije $f(Z)$. Tada je:

$$\begin{aligned} E_q(f(Z)) &\approx \frac{1}{N} \sum_{n=1}^{N} \int_D g(Z; \mu_n, \sigma_n^2 I) \hat{f}_n(Z) dZ \\ &= \frac{1}{N} \sum_{n=1}^{N} (f(\mu_n) + \frac{\sigma_n^2}{2} \operatorname{Tr}(\mathbf{H}_n)), \end{aligned} \tag{32}$$

gde je aproksimacija je dobijena delta metodom višeg reda za momente (v. [1]).

Sabiranjem ocena (29) i (32), dobijamo aproksimaciju ELBO-a:

$$ELBO(q) \approx L_2(q) = -\frac{1}{N} \sum_{n=1}^{N} (f(\mu_n) + \frac{\sigma_n^2}{2} \operatorname{Tr}(\mathbf{H}_n) - \ln q_n). \tag{33}$$

Primetimo da, sem pretpostavke da je logaritam zajedničke raspodele podataka i skrivenih promenljivih dva puta diferencijabilna funkcija, drugih pretpostavki nema. Takođe, valja napomenuti da nije potrebno računati celu matricu drugih izvoda, već samo njene dijagonalne elemente, kako u oceni ELBO-a figuriše samo trag ove matrice.

Postavlja se pitanje maksimizacije izraza (32) po parametrima μ_n i σ_n^2. Jedna opcija je, kao i obično, neka od gradijentnih metoda, međutim, ona bi zahtevala računanje gradijenta hesijana, tj. izračunavanje matrice trećih izvoda funkcije $f(Z)$. Sem što zahteva pooštravanje naših pretpostavki ($f(Z)$ tada mora biti tri puta diferencijabilna), ova opcija je i vremenski zahtevna.

Da bi se ovo izbeglo, upotrebićemo i aproksimaciju prvog reda ELBO-a. Do nje se dolazi na analogan način kao i do aproksimacije drugog reda, koristeći Tejlorov razvoj funkcije $f(Z)$ prvog reda:

$$L_1(q) = -\frac{1}{N}\sum_{n=1}^{N}(f(\mu_n) - \log q_n). \tag{34}$$

Sad proces optimizacije delimo u dve etape: maksimizacija aproksimacije ELBO-a prvog reda po parametrima μ_n, a zatim i maksimizacija aproksimacije drugog reda po parametrima σ_n^2. Ažuriranje varijacionih parametara se kroz ove dve etape vrši iterativno, do konvergencije aproksimacije drugog reda, $L_2(q)$.

Zbog prirode svoje definicije, ovaj metod je u stanju da, upotrebom dovoljnog broja komponenti (jezgara), dobro aproksimira i aposteriorne raspodele sa velikim brojem moda, koje često mogu predstavljati problem drugim metodima. Napomenimo da pretpostavka da je matrica kovarijacija oblika $\sigma_n^2 I$ može biti ograničavajuća, u smislu da otežava aproksimaciju asimetričnih aposteriornih raspodela sa teškim repovima. Ovaj problem se delimično prevazilazi povećanjem broja komponenti. Ipak, primetimo i da, povećanjem broj komponenti, broj parametara raste linearno, a time i vreme potrebno za njihovo ocenjivanje. Autori koji su predstavili ovaj metod (v. [8]) takođe procenjuju da KL razilaženje između stvarne aposteriorne i optimalne varijacione raspodele opada, u najboljem slučaju, logaritamski sa povećanjem broja komponenti, te je preporuka koristiti, ukoliko je to moguće, manji broj jezgara, kako bi se izbegli spomenuti kontraefekti.

4 Bajesovski statistički modeli

U naredna dva poglavlja demonstriraćemo primere upotrebe varijacionog zaključivanja na neke od popularnih statističkih modela - linearnu regresiju, kao i parametarske i neparametarske modele mešavina. Najpre ćemo, u četvrtom poglavlju, definisati spomenute modele iz bajesovske perspektive, a peto poglavlje sadrži implementirane metode varijacionog zaključivanja na ove modele, izvođenjem pravila ažuriranja skrivenih promenljivih ovih modela.

4.1 Linearna regresija

Linearna regresija je najkorišćeniji model u regresionim problemima, gde je cilj uspostaviti vezu između ciljne promenljive, Y, koja uzima vrednosti iz skupa R^d (nadalje ćemo razmatrati slučaj $d = 1$), i skupa prediktora $X \in R^{N\times n}$, gde je sa N označen obim uzorka, a sa n broj prediktora. Premda postoji nekoliko načina za uvođenje

linearne regresije, opredelićemo se za probabilistički pristup, jer najviše odgovara bajesovskoj perspektivi koju pokušavamo da istaknemo. Pretpostavka na kojoj počiva model linearne regresije je da je ciljna promenljiva normalno raspodeljena sa očekivanjem koje je linearno po parametrima modela:

$$Y|X \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n, \sigma^2), \tag{35}$$

gde n označava broj prediktora, a $\beta_0, \beta_1, \ldots, \beta_n$ parametre linearnog modela (skrivene promenljive) koje je cilj oceniti. Parametar σ^2 predstavlja disperziju modela, i pretpostavlja se jednakost disperzije u svakoj tački uzorka. Standardno odstupanje ciljne promenljive u praksi često zavisi od vrednosti prediktora, te ova pretpostavka predstavlja jedno od ograničenja modela linearne regresije. Ovaj parametar se može oceniti na više načina, a jedan od najčešćih se oslanja upravo na ocene parametara $\beta_0, \ldots, \beta_n$. Simplifikacije radi, nadalje ćemo smatrati ovu disperziju poznatom.

Naglasimo da sam naziv linearne regresije potiče od linearnosti po koeficijentima $\beta_0, \beta_1, \ldots, \beta_n$, dok linearnost po prediktorima nije nužna. Radi kompaktnijeg zapisa, skup prediktora se obično dopunjava prediktorom X_0, čije su sve vrednosti jednake 1, te se linearna kombinacija može predstaviti u obliku skalarnog proizvoda:

$$Y|X \sim \mathcal{N}(\beta \cdot X, \sigma^2). \tag{36}$$

Pre nego što se posvetimo bajesovskoj oceni parametara linearne regresije, osvrnućemo se na tradicionalniji, frekventistički pristup. Ovaj osvrt će nam ponuditi uvid u neke od poteškoća na koje se nailazi pri oceni parametara linearne regresije, a koje nećemo sresti pratimo li postupak bajesovskog ocenjivanja. Frekventističkim pristupom, kao što je i uobičajeno, parametri modela se ocenjuju metodom maksimalne verodostojnosti. Funkcija verodostojnosti jednaka je:

$$L(\beta) = P(Y_1, Y_2, \ldots, Y_N | X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(Y_i|X_i), \tag{37}$$

gde je N obim uzorka, a elementi uzorka su međusobno nezavisni. Da bismo pronašli ocenu maksimalne verodostojnosti, rešavamo ekvivalentan problem, minimizaciju izraza $-\log L(\beta)$:

$$\begin{aligned} -\log L(\beta) &= -\log \prod_{i=1}^{N} P(Y_i|X_i) = -\sum_{i=1}^{N} \log P(Y_i|X_i) = \\ &= -\sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta \cdot X_i)^2}{2\sigma^2}} = \\ &= -N \log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \beta \cdot X_i)^2. \end{aligned} \tag{38}$$

Kako N i σ^2 ne zavise od parametara, ocena maksimalne verodostojnosti se dobija minimizacijom izraza

$$\sum_{i=1}^{N} (Y_i - \beta \cdot X_i)^2, \tag{39}$$

ili matrično

$$||Y - X\beta||^2, \tag{40}$$

gde su Y i β kolona vektori dužina N i n, respektivno, a X matrica dimenzija $N \times n$. Dakle, ocena maksimalne verodostojnosti je jednaka oceni dobijenoj metodom najmanjih kvadrata. Diferenciranjem matričnog izraza dolazimo do ocene maksimalne verodostojnosti:

$$\begin{aligned} & X^T(Y - X\hat{\beta}) = 0 \\ & X^T X\hat{\beta} = X^T Y \\ & \hat{\beta} = (X^T X)^{-1} X^T Y. \end{aligned} \tag{41}$$

Ovime je data forma analitičkog rešenja koeficijenata linearne regresije. Međutim, u nekim slučajevima, inverz matrice $X^T X$ ne postoji, zbog linearne zavisnosti kolona matrice X. U praksi se ovaj slučaj retko sreće, ali se dešava da kolone matrice X imaju jaku linearnu korelisanost, te je matrica $X^T X$ loše uslovljena, i rešenje može biti jako nestabilno, odnosno male promene vrednosti matrice X dovode do značajnih promena vrednosti inverza $(X^T X)^{-1}$. Ovo je vrlo nepoželjno ponašanje, zbog toga što je, u praksi, X matrica podataka koji mogu biti, primera radi, rezultat merenja koja mogu i ne moraju biti potpuno precizna, ili pak mogu biti podložni ljudskoj grešci pri unosu podataka. U svrhu povećanja stabilnosti rešenja, obično se pristupa *regularizaciji*. Regularizacija podrazumeva promenu cilja optimizacionog problema, tj. umesto izraza $||Y - X\beta||^2$ se minimizuje izraz kom je dodat regularizacioni izraz, sa ciljem postizanja stabilnosti i robusnosti rešenja. Tipična regularizaciona ideja je upotreba neke od L-normi parametara β (najčešće $L2$ i $L1$), kojom se sprečavaju velike i nestabilne vrednosti parametara:

$$||Y - X\beta||^2 + \lambda||\beta||^2, . \tag{42}$$

Diferenciranjem izraza $(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$ po β dolazi se do stabilnijeg rešenja:

$$\begin{aligned} & -2X^T(Y - X\hat{\beta}) + 2\lambda\hat{\beta} = 0 \\ & X^T X\hat{\beta} + \lambda\hat{\beta} = X^T Y \\ & \hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y, \end{aligned} \tag{43}$$

gde je I jedinična matrica. Parametar λ se naziva i *regularizacionim parametrom*, i uobičajeno se nalazi pretraživanjem unapred zadatog skupa vrednosti. Regularizacija je opšti metod koji je poželjno (često i nužno) koristiti kod većine statističkih modela, a sem L-normi, u upotrebi su i drugi mehanizmi sa regularizacionom svrhom.

Poteškoće sa ovim rešenjem linearne regresije nastaju i u slučaju kada je matrica $X^T X$ visokodimenziona. Ovo se dešava kada X ima mnogo kolona, tj. kada je broj prediktora veliki. U praksi ovo nije redak slučaj, pa se zbog toga se do ocene parametara β često dolazi optimizacijom, primenom gradijentnih metoda na izraz (40), odnosno regularizovani izraz (42).

Primetimo da se pokazanim metodom mogu dobiti samo tačkaste ocene parametara β, ali se, u fazi predviđanja vrednosti ciljne promenljive za nove vrednosti prediktora, može doći do intervalne ocene za Y, na osnovu (36).

Vratimo se razmatranju linearne regresije iz bajesovske perspektive. Za razliku od frekventističkog pristupa, kojim se može doći samo do tačkastih ocena, bajesovski nudi mogućnost nalaženja intervalne ocene parametara β. Primenimo Bajesovu teoremu (1) na postavljeni problem:

$$P(\beta|Y, X) = \frac{P(Y|\beta, X)P(\beta)}{P(Y, X)}. \tag{44}$$

Iz (36) poznato je da $Y|\beta, X$ ima normalnu raspodelu sa poznatom disperzijom, i matematičkim očekivanjem $X\beta$, te se, radi postizanja konjugovanosti, za apriornu raspodelu parametara β uzima normalna raspodela $\mathcal{N}(m_\beta, \tau^{-1}I)$. Ukoliko je domensko znanje vezano za uticaj svakog od prediktora na ciljnu promenljivu poznato, može se integrisati kroz apriornu raspodelu parametra m_β. Ipak, kako to nije generalan slučaj, neka je $m_\beta = 0$. Primetimo da mogućnost inkluzije domenskog znanja nije predviđena frekventističkim pristupom (mada se može dodati optimizacionom problemu u vidu regularizacije). Takođe, apriornom raspodelom se kontroliše i problem koji je u frekventističkom pristupu bio inherentan - nestabilnost rešenja usled velikih vrednosti ocena parametara β. Stoga, apriorna raspodela vrši višestruku regularizacionu ulogu, te minimizacija norme vektora parametara nije potrebno.

Parametar τ se naziva i (hiper)parametrom preciznosti, i može se smatrati inverzom disperzije (što u jednodimenzionom slučaju i jeste). Premda je moguć izbor proizvoljne pozitivne raspodele, tipičan izbor apriorne raspodele za parametar preciznosti, koji omogućava konjugovanost, je gama raspodela (ekvivalent odabiru inverzne gama raspodele za parametar disperzije normalne raspodele [2]). Stoga, definišemo apriornu raspodelu parametra τ kao gama raspodelu, sa parametrima α_1 i α_2.

Dakle, potpun bajesovski model linearne regresije dat je pomoću:

$$\begin{aligned} Y &\sim \mathcal{N}(X\beta, \sigma^2) \\ \beta &\sim \mathcal{N}(0, \tau^{-1}) \\ \tau &\sim Gamma(\alpha_1, \alpha_2). \end{aligned} \tag{45}$$

Ovom definicijom bajesovskog modela linearne regresije ćemo se poslužiti u sledećem poglavlju, pri upotrebi varijacionog zaključivanja za ocenjivanje parametara β i τ.

4.2 Model Gausovih mešavina

Model Gausovih mešavina predstavlja jedan od najkorišćenijih metoda klasterovanja podataka. Pod klasterovanjem se podrazumeva dodeljivanje segmenta (ili klastera) svakom elementu uzorka, tako da slični elementi budu u istom segmentu. U parametarskoj postavci modela Gausovih mešavina, pretpostavlja se da svaki podatak pripada jednom od K klastera, koji su predstavljeni normalnom raspodelom sa očekivanjem μ_k i disperzijom σ_k^2. Cilj modela je da oceni parametre svakog klastera (tj. svake od K normalnih raspodela), kao i pripadnost svakog podatka klasteru: svakom X_i se pridružuje K-dimenzioni indikatorski vektor c_i, za koji je c_{ij} jednako 1, ako X_i pripada klasteru j, a jednako 0 u suprotnom. Dakle, skup skrivenih promenljivih je sačinjen od parametara μ_k, σ_k^2, $k \in \{1, 2, \ldots, K\}$ i promenljivih c_i, $i \in \{1, 2, \ldots, n\}$, gde n označava obim uzorka. U nastavku ćemo demonstrirati pojednostavljenu verziju ovog modela, za koju je $\sigma_k^2 = 1$, za svako k. Jednostavnosti radi, pretpostavićemo i da je uzorak X jednodimenzioni, tj. $X_i \in R, i \in \{1, 2, \ldots, n\}$. Prikazani metodi se bez poteškoća primenjuju i na višedimenzione skupove podataka.

Uočimo specifičnost skupa skrivenih promenljivih, koja je primer razlike između globalnih i lokalnih promenljivih, opisane u poglavlju 3.2. Naime, skrivena promenljiva c_i određuje samo i-tu tačku uzorka, dok parametri sredine i disperzije svake od normalnih raspodela utiču na raspodelu celog uzorka. Takođe, pripadnost tačke uzorka klasteru zavisi samo od njene vrednosti i globalnih parametara μ_k, σ_k^2, a ne i od ostalih tačaka uzorka, te uslovna nezavisnost definisana u poglavlju 3.2 važi.

Tipično, za apriornu raspodelu parametara μ_k se uzima normalna raspodela $\mathcal{N}(0, \sigma^2)$, gde je σ^2 unapred fiksiran hiperparametar, a za apriornu raspodelu vektora c_i uniformna diskretna raspodela na skupu $\{1, 2, \ldots, K\}$ sa zakonom verovatnoća $(\frac{1}{K}, \ldots, \frac{1}{K})$.

Razmotrimo sada združenu raspodelu skrivenih promenljivih i uzorka:

$$\begin{aligned} p(X, \mu, c) &= p(\mu)p(c)p(X|\mu, c) \\ &= \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(c_i)p(X_i|\mu, c_i). \end{aligned} \tag{46}$$

Kao što je navedeno, definišemo $p(\mu_k)$ kao gustinu normalne raspodele sa očekivanjem 0 i disperzijom σ^2, $p(c_i)$ kao uniformnu diskretnu raspodelu, a $p(X_i|\mu, c_i)$ kao normalnu raspodelu sa jediničnom disperzijom i očekivanjem μ_k, za ono k za koje je $c_{ik} = 1$, tj. očekivanjem $c_i^T \mu$.

Za egzaktno određivanje aposteriorne raspodele $p(\mu, c|X)$ neophodno je nalaženje verovatnoće podataka:

$$p(x) = \int_{R^K} p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i)p(X_i|\mu, c_i)d\mu. \tag{47}$$

U svakom činiocu proizvoda u podintegralnoj funkciji figurišu sve vrednosti $\mu_1, \ldots, \mu_K$, što znači da se ovaj integral ne može razložiti na proizvod jednostrukih integrala, te se može rešiti samo numerički, ali uz eksponencijalnu vremensku kompleksnost $O(K^n)$ (v. [3]). U slučaju velikog broja klastera, i, poslovično, velikog obima uzorka, ovo rešenje postaje neprimenljivo. U petom poglavlju biće demonstrirana primena varijacionog zaključivanja u aproksimaciji aposteriornih raspodela skrivenih promenljivih modela Gausovih mešavina, a neka od alternativnih rešenja biće spomenuta u šestom poglavlju.

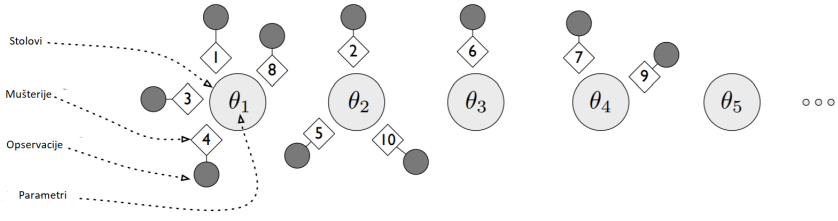
4.3 Neparametarski model mešavina

Pri kreiranju parametarskog modela mešavina, nameće se suštinski problem odabira parametra K, broja klastera. Ovaj problem se u većini slučajeva rešava heuristički, probanjem nekoliko vrednosti za parametar K, a zatim uporednom analizom dobijenih modela. Ipak, ovo rešenje nije zadovoljavajuće jer ne nudi garancije da je parametar dobijen ovim metodom zaista optimalan, ili je, pak, potrebno obučavanje velikog broja modela da bi se do garancija došlo. U nekim slučajevima, problem se prevazilazi ekspertskim znanjem i detaljnim poznavanjem podataka i raspodele iz koje podaci dolaze. Ipak, ovaj put je često težak, a nekada i nemoguć.

Čak i kada se ovaj problem prevaziđe, postavlja se pitanje klasifikacije novog podatka u jedan od klastera. U parametarskoj postavci se ne dozvoljava mogućnost nepripadanja nijednom klasteru, tj. kreiranju novog klastera, u zavisnosti od novopridošlih podataka. Ovo ponašanje nije poželjno u slučaju kada buduće podatke želimo da delimo u klastere, jer je moguće da se raspodela podataka, a time i optimalan broj klastera, vremenom menja, naročito kada je inicijalan skup klastera kreiran na osnovu malog uzorka. Takođe, ovakav model sprečava pojavu odudarajućih podataka (eng. *outliers*) u remećenju postojećih klastera, zbog teorijske mogućnosti da svaki odudarajući podatak dobije klaster.

Neparametarski oblik modela mešavina prevazilazi oba pomenuta problema. U

ovom slučaju, parametar K nije unapred zadat, već se procenjuje na osnovu uzorka, i nije ograničen, tj. dodavanjem novih podataka, broj klastera može da se povećava.
Evidentno je da, proširenjem modela mešavina na neparametarski oblik, više nije moguće koristiti diskretnu uniformnu raspodelu kao apriornu raspodelu za vektor pripadnosti, zbog neograničenog broja klastera. Stoga se za apriornu raspodelu vektora pripadnosti koristi tzv. *Dirihleov proces*, ili *proces kineskog restorana*. Ovaj naziv je motivisanim misaonim eksperimentom, koji pretpostavlja da postoji restoran sa neograničenim brojem stolova za kojima može sedeti neograničen broj mušterija, koji u restoran pristižu jedna za drugom. Prva mušterija ulazi i seda za jedan sto. Druga mušterija ulazi i bira za koji sto seda: pretpostavimo da seda za isti sto sa verovatnoćom $\frac{1}{1+\alpha}$, a za novi, prazan, sto, sa verovatnoćom $\frac{\alpha}{1+\alpha}$, gde je α proizvoljan pozitivan realan broj. Kada n-ta mušterija uđe u restoran, verovatnoća sa kojom seda za već popunjen sto je proporcionalna broju mušterija koje već sede za tim stolom, a verovatnoća sa kojom seda za prazan sto je proporcionalna parametru α. Intuitivno, sto sa više mušterija postaje "popularniji", i privlači više mušterija sebi, dok prazan sto ima konstantnu popularnost, određenu parametrom α.



Slika 5: Ilustracija procesa kineskog restorana (preuzeto iz [7])

U formalnijem tonu, neka je c_n vektor pripadnosti n-tog podatka (mušterije). Tada je:

$$P(c_n = k | c_1, \ldots, c_{n-1}) = \begin{cases} \frac{m_k}{n-1+\alpha}, & \text{za } k \leq K_{n-1} \\ \frac{\alpha}{n-1+\alpha}, & \text{za } k = K_{n-1}+1, \end{cases} \tag{48}$$

gde m_k označava broj podataka u k-tom klasteru (broj mušterija za k-tim stolom), a K_{n-1} broj klastera (stolova) u koje se grupišu prvih $n-1$ podataka, i $\alpha > 0$. Primetimo da k može uzimati vrednosti $1, 2, \ldots, K_{n-1}$ (neki od već formiranih klastera) i $K_{n-1}+1$ (novi klaster). Parametar α se naziva i parametrom koncentracije ili zgušnjavanja, i upravlja grupisanjem podataka u klastere. Veća vrednost parametra α povećava verovatnoću kreiranja novog klastera, i uzrokuje tendenciju podataka da se raspršе u klastere, a time i klastere sa malim brojem podataka. Male vrednosti α prisiljavaju podatke da se grupišu u što manji broj klastera.

Konstatujmo i da važi:

$$\begin{aligned}\sum_{k=1}^{K_{n-1}+1} P(c_n = k|c_1,\ldots,c_{n-1}) &= \sum_{k=1}^{K_{n-1}} \frac{m_k}{n-1+\alpha} + \frac{\alpha}{n-1+\alpha} \\ &= \frac{1}{n-1+\alpha}\sum_{k=1}^{K_{n-1}} m_k + \frac{\alpha}{n-1+\alpha} \\ &= \frac{n-1}{n-1+\alpha} + \frac{\alpha}{n-1+\alpha} = 1,\end{aligned} \tag{49}$$

te je, imajući u vidu i $\alpha > 0$, zaključujemo da je sa (26) opisana diskretna raspodela.

Iz same definicije Dirihleovog procesa nije očigledno kako raspored podataka utiče na formiranje klastera, tj. raspodelu $P(c_1, c_2, \ldots, c_n)$. Kako se, u opštem slučaju, raspored tačaka u uzorku može smatrati potpuno nasumičnim, njegov uticaj na konfiguraciju klastera bi bio izrazito nepoželjan. Pokažimo zato ključno svojstvo Dirihleovog procesa - invarijantnost na permutaciju uzorka. Važi:

$$P(c_1, c_2, \ldots, c_n) = P(c_1)P(c_2|c_1)\ldots P(c_n|c_1, c_2, \ldots c_{n-1}). \tag{50}$$

Označimo sa I_k skup indeksa svih podataka koji pripadaju klasteru k, sa $I_{k,1}$ indeks prvog podatka koji je pridružen klasteru k, sa $I_{k,2}$ indeks drugog podatka pridruženog klasteru k itd. Razmotrimo sada sve činioce proizvoda (55) koji odgovaraju elementima klastera k (tj. one c_i, za koje je $c_i = k$. Proizvod ovih činilaca jednak je:

$$\frac{\alpha \cdot 2 \cdot \ldots (m_k - 1)}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha)\ldots(I_{k,m_k} - 1 + \alpha)}, \tag{51}$$

gde, koristeći (53), činilac $\frac{\alpha}{I_{k,1}-1+\alpha}$ potiče od prvog podatka u klasteru k (onom koji je kreirao ovaj klaster), činilac $\frac{2}{I_{k,2}-1+\alpha}$ od drugog itd. Napišimo sada (55) u obliku proizvoda po svim klasterima k:

$$P(c_1, c_2, \ldots, c_n) = \prod_{k=1}^{K_n} \frac{\alpha(m_k - 1)!}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha)\ldots(I_{k,m_k} - 1 + \alpha)}. \tag{52}$$

Kako svaki podatak pripada tačno jednom klasteru, svaki indeks $i = 1, 2, \ldots, n$ se u ovom proizvodu pojavljuje tačno jednom, te imenilac možemo zapisati u obliku koji ne zavisi od indeksa podataka koji su pridruženi svakom od klastera:

$$P(c_1, c_2, \ldots, c_n) = \frac{\alpha^{K_n}\prod_{k=1}^{K_n}(m_k - 1)!}{\prod_{i=1}^{n}(i - 1 + \alpha)}. \tag{53}$$

Dakle, raspodela vektora pripadnosti zavisi od kardinalnosti svih klastera, i broja formiranih klastera, ali ne i od rasporeda podataka, te je Dirihleov proces invarijantan na permutaciju tačaka uzorka.

Parametar koncentracije direktno upravlja brojem formiranih klastera. Može se pokazati (v. [30]) da očekivana vrednost formiranih klastera raste logaritamski, $O(\alpha \log n)$, sa porastom obima uzorka n.

Upotrebom Dirihleovog procesa za apriornu raspodelu vektora pripadnosti, preostaje još da, slično parametarskom modelu mešavina, svakom klasteru dodelimo raspodelu podataka koji njemu pripadaju.

5 Primena varijacionog zaključivanja na bajesovske statističke modele

5.1 Linearna regresija

Dakle, cilj varijacionog zaključivanja jeste ocena aposteriorne raspodele skrivenih promenljivih. U slučaju linearne regresije, skrivene promenljive koje je potrebno oceniti su β i τ, prema (44), i važi:

$$\begin{aligned} P(\beta|\tau) &\sim \mathcal{N}(0, \tau^{-1}I) \\ P(\tau) &\sim Gamma(\alpha_1, \alpha_2). \end{aligned} \tag{54}$$

Radi primene strukturiranog varijacionog zaključivanja, pretpostavljamo nezavisnost parametara β od parametra τ, tj. faktorizaciju proizvoljne varijacione raspodele q:

$$q(\beta, \tau) = q(\beta)q(\tau). \tag{55}$$

Sada je cilj naizmenično ažurirati parametre β i τ, do konvergencije $ELBO$, postupkom opisanim u poglavlju 4.5. Prisetimo se da je pravilo ažuriranja:

$$q_{P_i}(Z_{P_i}) = e^{E_{-P_i} \ln p(Z,X)}, \tag{56}$$

tj. u slučaju koji se rešava:

$$\begin{aligned} \ln q_1(\beta) &= E_\tau \ln p(\beta, \tau, Y|X) + const \\ \ln q_2(\tau) &= E_\beta \ln p(\beta, \tau, Y|X + const. \end{aligned} \tag{57}$$

Pozabavimo se najpre raspodelom q_2:

$$\begin{aligned} \ln q_2(\tau) &= E_\beta \ln p(\beta, \tau, Y|X + const \\ &= E_\beta \ln p(Y|X, \beta)p(\beta|\tau)p(\tau) + const \\ &= E_\beta \ln p(Y|X, \beta) + E_\beta \ln p(\beta|\tau + E_\beta \ln p(\tau) + const \\ &= E_\beta \ln p(\beta|\tau) + \ln p(\tau) + const \\ &= \frac{N}{2} \ln \tau - \frac{\tau}{2} E_\beta \beta^T \beta + (\alpha_1 - 1) \ln \tau - \alpha_2 \tau \\ &= (\alpha_1 + \frac{N}{2} - 1) \ln \tau - (\alpha_2 + \frac{1}{2} E_\beta \beta^T \beta)\tau, \end{aligned} \tag{58}$$

što daje:

$$q_2(\tau) \sim Gamma(\alpha_1 + \frac{N}{2}, \alpha_2 + \frac{1}{2} E_\beta \beta^T \beta). \tag{59}$$

Odredimo sada i varijacioni faktor koji odgovara parametrima β:

$$
\begin{aligned}
\ln q_1(\beta) &= E_\tau \ln p(\beta, \tau, Y|X) + const \\
&= E_\tau \ln p(Y|X,\beta)p(\beta|\tau)p(\tau) + const \\
&= E_\tau \ln p(Y|X,\beta) + E_\tau \ln p(\beta|\tau) + E_\tau \ln p(\tau) + const \\
&= \ln p(Y|X,\beta) + E_\tau \ln p(\beta|\tau) + const \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 - \frac{1}{2}E_\tau \tau \beta^T\beta + const \\
&= -\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) - \frac{1}{2}\frac{\alpha_1}{\alpha_2}\beta^T\beta + const \\
&= -\frac{1}{2}\beta^T(\frac{\alpha_1}{\alpha_2}I + \frac{1}{\sigma^2}X^TX)\beta + \frac{1}{\sigma^2}\beta^T X^T Y + const.
\end{aligned} \tag{60}
$$

Kako je poslednji izraz kvadratna forma, raspodela $q_1(\beta)$ je normalna:

$$
\begin{aligned}
q_1(\beta) &\sim \mathcal{N}(m, S) \\
m &= \frac{1}{\sigma^2} S X^T Y \\
S &= (\frac{\alpha_1}{\alpha_2}I + \frac{1}{\sigma^2}X^TX)^{-1}.
\end{aligned} \tag{61}
$$

Konačno, neophodno je odrediti i pravilo za izračunavanja $ELBO$, radi provere konvergencije i kriterijuma zaustavljanja. Iz (7) je:

$$
\begin{aligned}
ELBO(q) &= E_q \ln p(X,Z) - E_q \ln q(Z) \\
&= \iint q(\beta,\tau)\ln p(Y,\beta,\tau|X)d\beta d\tau - \iint q(\beta,\tau)\ln q(\beta,\tau)d\beta d\tau \\
&= E_{\beta,\tau} lnp(Y,\beta,\tau|X) - E_{\beta,\tau}\ln q(\beta,\tau) \\
&= E_\beta \ln p(Y|X,\beta) + E_{\beta,\tau}\ln p(\beta|\tau) - E_\beta \ln q(\beta) - E_\tau \ln q(\tau).
\end{aligned} \tag{62}
$$

Svaki od sabiraka se sada može izračunati pojedinačno:

(63)

$$
\begin{aligned}
E_\beta \ln p(Y|X,\beta) &= \frac{1}{2}\ln\frac{1}{4\pi\sigma^2} - \frac{1}{2\sigma^2}Y^TY + \frac{1}{\sigma^2}m^TX^TY - \frac{1}{2\sigma^2}tr(X^TX(mm^T + S)) \\
E_{\beta,\tau}\ln p(\beta|\tau) &= -\frac{N}{2}\ln 2\pi + \frac{N}{2}(\psi(\alpha_\tau) - \ln\beta_\tau - \frac{\alpha_\tau}{2\beta_\tau}(m^Tm + tr(S)) \\
E_\tau \ln p(\tau) &= \alpha_1 \ln\alpha_2 + (\alpha_1 - 1)(\psi(\alpha_\tau) - \ln\beta_\tau) - \alpha_2\frac{\alpha_\tau}{\beta_\tau} - \ln\Gamma(\alpha_1) \\
E_\beta \ln q(\beta) &= -\frac{1}{2}\ln|S| - \frac{D}{2}(1 + \ln 2\pi) \\
E_\tau \ln q(\tau) &= -\ln\Gamma(\alpha_\tau) + (\alpha_\tau - 1)\psi(\alpha_\tau) + \ln\beta_\tau - \alpha_\tau.
\end{aligned}
$$

Sada je postupak optimizacije jasan. Inicijalizacijom vrednosti α_1, α_2, τ inicijalizuju se i varijacioni faktori, koji se iterativno ažuriraju pomoću (63) i (65), do konvergencije vrednosti $ELBO$, date u (66).

Očito je da ocena parametara linearne regresije upotrebom varijacionog zaključivanja nije jednostavan proces. Zato se, kao što je i pomenuto u uvodu, parametri linearne regresije tipično ocenjuju drugim metodima, ali je linearna regresija izabran kao prvi primer zbog svoje rasprostranjenosti i jednostavnosti.

5.2 Model Gausovih mešavina

Vratimo se ocenjivanju parametara modela mešavina. Govorićemo o modelu Gausovih mešavina sa jediničnom disperzijom, gde je pretpostavka da je združena raspodela podataka mešavina K normalnih raspodela. Potrebno je detektovati skrivenu strukturu podataka, tj. K skrivenih klastera, gde svakom klasteru odgovara normalna raspodela sa očekivanjem μ_k i disperzijom 1. Sem toga, svakoj tački iz uzorka X_i treba dodeliti K-dimenzioni vektor pripadnosti c_i koji ima sve nule sem na poziciji koja odgovara klasteru kom pripada X_i. Parametri μ_k i vektori pripadnosti c_i, $(k = 1, 2, \ldots, K,\ i = 1, 2, \ldots, n)$, predstavljaju skrivene promenljive modela koje ocenjujemo varijacionim zaključivanjem.

Prisetimo se da je ukupna raspodela podataka i skrivenih varijabli data pomoću:

$$p(\mu, c, X) = p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \mu). \tag{64}$$

Primenićemo aproksimaciju srednjeg polja, tj. pretpostavljamo nezavisnost varijacionih parametara. Aposteriornu raspodelu parametra μ aproksimiramo normalnom raspodelom $\mathcal{N}(m_k, s_k^2)$, a aposteriornu raspodelu vektora pripadnosti c_i aproksimiramo diskretnom (kategoričkom) raspodelom sa raspodelom verovatnoća $(\varphi_{i1}, \varphi_{i2}, \ldots, \varphi_{iK})$. Proces optimizacije ima za cilj nalaženje optimalnih varijacionih parametara m_k, s_k^2 i φ_i, $(k = 1, 2, \ldots, K,\ i = 1, 2, \ldots, n)$. Ažuriranje varijacionih parametara se vrši koordinatno, kao što je opisano u poglavlju 4.5. Ažuriranje se vrši do konvergencije ELBO, koja u ovom slučaju ima oblik:

$$\begin{aligned}
ELBO(m, s^2, \varphi) &= E(\ln p(m, s^2, \varphi, X)) - E(\ln q(m, s^2, \varphi)) \\
&= E(\ln p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \mu)) \\
&\quad - \sum_{i=1}^{n} E(\ln q(c_i; \varphi_i)) - \sum_{k=1}^{K} E(\ln q(\mu_k; m_k, s_k^2)) \\
&= \sum_{k=1}^{K} E(\ln p(\mu_k; m_k, s_k^2)) \\
&\quad + \sum_{i=1}^{n} (E(\ln p(c_i; \varphi_i) + E(\ln p(x_i | c_i, \mu; \varphi_i, k, s^2)) \\
&\quad - \sum_{i=1}^{n} E(\ln q(c_i; \varphi_i)) - \sum_{k=1}^{K} E(\ln q(\mu_k; m_k, s_k^2)).
\end{aligned} \tag{65}$$

U koordinatnoj optimizaciji, najpre ćemo ažurirati parametre φ_i, a zatim parametre m_k i s_k.

Ažuriranje parametara aposteriorne raspodele vektora pripadnosti. Prisetimo se da se, u koordinatnoj optimizaciji pri strukturiranom varijacionom zaključivanju, varijacioni parametri ažuriraju u skladu sa pravilom $q_j(Z_j) = e^{E_{-j} \ln p(Z, X)}$, do konvergencije vrednosti $ELBO$. U konkretnom slučaju ocene aposteriorne raspodele

vektora pripadnosti, važi:

$$
\begin{aligned}
q^*(c_i, \varphi_i) &\propto e^{E_{-c_i}(\ln p(\mu) \prod_{i=1}^n p(c_i) p(x_i|c_i,\mu))} \\
&\propto e^{E_{-c_i}(\ln p(\mu) p(c_i) p(x_i|c_i,\mu))} \\
&\propto e^{E_{\mu}(\ln p(\mu) p(c_i) p(x_i|c_i,\mu))} \\
&\propto e^{E_{\mu}(\ln p(c_i) p(x_i|c_i,\mu))} \\
&\propto e^{\ln p(c_i) + E_{\mu}(\ln(p(x_i|c_i,\mu)}.
\end{aligned} \tag{66}
$$

Rezultat je dobijen do na normalizaciju raspodele, eliminacijom svih faktora koji ne zavise od c_i i konstanti. Pod pretpostavkom uniformne apriorne raspodele za vektor pripadnosti, važi $\ln p(c_i) = -\ln K$, te je izraz $e^{\ln p(c_i)}$ konstantan. S obzirom na to da je c_i K-dimenzioni indikator, možemo zapisati:

$$
p(x_i|c_i, \mu) = \prod_{k=1}^{K} p(x_i|\mu_k)^{c_{ik}}.
$$

Odavde dalje sledi:

$$
\begin{aligned}
E_{\mu}(\ln p(x_i|c_i, \mu)) &= \sum_{k=1}^{K} c_{ik} E_{\mu}(\ln p(x_i|\mu_k)) \\
&= \sum_{k=1}^{K} c_{ik} E_{\mu}(-(x_i - \mu_k)^2/2) + const \\
&= \sum_{k=1}^{K} c_{ik} (E(\mu_k) x_i - E(\mu_k^2)/2) + const.
\end{aligned} \tag{67}
$$

μ_k ima normalnu raspodelu sa očekivanjem m_k i standardnim odstupanjem s_k, te je $E(\mu_k) = m_k$, a $E(\mu_k^2) = m_k^2 + s_k^2$. Odavde sledi da se parametri φ ažuriraju po pravilu:

$$
\varphi_{ik} \propto e^{m_k x_i - (m_k^2 + s_k^2)/2}. \tag{68}
$$

Ažuriranje parametara aposteriorne raspodele sredine komponenti mešavine Gausovih raspodela. Preostali delić slagalice je pravilo ažuriranja parametara aposteriorne raspodele sredine Gausovih raspodela mešavine, tj. varijacionih parametara m_k i s_k.

Slično kao u prethodnom odeljku, parametri se ažuriraju po pravilu definisanom u poglavlju 2.4:

$$
q(\mu_k) \propto e^{\ln p(\mu_k) + \sum_{i=1}^n E(\ln p(x_i|c_i,\mu)} \tag{69}
$$

Prisetimo se da je φ_{ik} verovatnoća da i-ti element uzorka pripada k-tom klasteru, te je $\varphi_{ik} = E(c_{ik})$. Stoga je:

$$
\begin{aligned}
\ln q(\mu_k) &= \ln p(\mu_k) + \sum_{i=1}^{n} E_{\varphi_i, m_{-k}, s^2_{-k}}(\ln p(x_i|c_i, \mu)) + const \\
&= \ln p(\mu_k) + \sum_{i=1}^{n} E_{\varphi_i}(c_{ik} \ln p(x_i|\mu_k)) + const \\
&= -\frac{mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} -\varphi_{ik} \frac{(x_i - \mu_k)^2}{2} + const \\
&= -\frac{mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} (\varphi_{ik} x_i \mu_k - \varphi_{ik} \frac{mu_k^2}{2}) + const \\
&= (\sum_{i=1}^{n} \varphi_{ik} x_i)\mu_k - (\frac{1}{2\sigma^2} + \sum_{i=1}^{n} \frac{\varphi_{ik}}{2})\mu_k^2 + const.
\end{aligned}
\tag{70}
$$

Zaključujemo da je $q(\mu_k)$ raspodela iz eksponencijalne familije, sa dovoljnim statistikama $\{\mu_k, \mu_k^2\}$ i parametrima $\{\sum_{i=1}^{n} \varphi_{ik} x_i, \frac{1}{2\sigma^2} + \sum_{i=1}^{n} \frac{\varphi_{ik}}{2}\}$, te su parametri Gausove raspodele $q(\mu_k)$:

$$
\begin{aligned}
m_k &= \frac{\sum_{i=1}^{n} \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^{n} \varphi_{ik}}, \\
s_k^2 &= \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^{n} \varphi_{ik}}.
\end{aligned}
\tag{71}
$$

Sada je preostalo još sprovesti koordinatnu optimizaciju, po postupku opisanom u 2.4 - najpre se inicijalizuju parametri m, s^2, φ, a potom se naizmenično ažuriraju po pravilima opisanim u (68) i (71), do konvergencije $ELBO$, čija se vrednost jednostavno računa iz (65).

Uočimo da ažuriranje parametara m i s^2 zavisi samo od parametara φ, i poznatih vrednosti σ^2 i X, te se postupak može sprovesti i u duhu strukturiranog varijacionog zaključivanja, grupnim ažuriranjem svih parametara φ_i, naizmenično sa ažuriranjem grupe parametara m_k, s_k^2.

Primetno je da je postupak varijacionog zaključivanja u slučaju modela Gausovih mešavina daleko jednostavniji za izvođenje, u poređenju sa do sada pokazanim, te nije čudno što se upravo ovaj model može pronaći u brojnim izvorima kao reprezentativni primer primene varijacionog zaključivanja.

5.3 Neparametarski model mešavina

Iako vrlo intuitivan, zahvaljujući reprezentaciji pomoću procesa kineskog restorana, neparametarski model mešavina zasnovan na Dirihleovom procesu nije najpogodniji za primenu aproksimacije srednjeg polja. Razlog za to je neadekvatnost pretpostavke o faktorizaciji varijacione raspodele po parametrima pripadnosti $q(c) = \prod_{i=1}^{n} q_i(c_i)$, s obzirom na činjenicu da različite vrednosti vektora pripadnosti mogu generisati isto klasterovanje podataka (npr. klasterovanja $c_1 = 1, c_2 = 1, c_3 = 2$ i $c_1 = 3, c_2 = 3, c_3 = 2$ predstavljaju isto particionisanje podataka), koja implicira izraženu strukturu zavisnosti između parametara pripadnosti. Zbog toga se pribegava konačnodimenzionim aproksimacijama Dirihleovog procesa, kojim se modifikuje apriorna raspodela

vektora pripadnosti (v. [4, 15]).

Jedan od pristupa aproksimaciji Dirihleovog procesa se zasniva na upotrebi simetrične Dirihleove raspodele. Vektor pripadnosti se aproksimira diskretnom raspodelom π, za koju definišemo apriornu Dirihleovu raspodelu $\mathcal{D}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$, za konačno K. Označimo sa Z_k parametre k-tog klastera (recimo očekivanje i disperzija normalne raspodele, u slučaju neparametarskog modela Gausovih mešavina). Tada je zajednička raspodela uzorka, pripadnosti, i parametara koji se ocenjuju:

$$P(X, c, Z, \pi) = \left[\prod_{i=1}^{n} p(X_i|Z_{c_i})p(c_i|\pi)\right]\left[\prod_{k=1}^{K} p(Z_k)\right]\mathcal{D}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}), \tag{72}$$

a faktorizacija varijacione raspodele:

$$q(Z, c, \pi) = \left[\prod_{i=1}^{n} q(c_n)\right]\left[\prod_{k=1}^{K} q(Z_k)\right]q(\pi). \tag{73}$$

Integralni korak koordinatne optimizacije varijacionih parametara je evaluacija $ELBO$ funkcionala, koji za neparametarski model mešavina ima oblik:

$$\begin{aligned} ELBO(q) &= E_q \ln p(X, Z) - E_q \ln q(Z) \\ &= \sum_{i=1}^{n}\sum_{c_i=1}^{K}\int_{Z_{c_i}} q(c_i)q(Z_{c_i}) \ln p(X_i|Z_{c_i})dZ_{c_i} \\ &+ \sum_{k=1}^{K}\int_{Z_k} q(Z_k) \ln \frac{p(Z_k)}{q(Z_k)}dZ_k - \sum_{i=1}^{n}\sum_{c_i=1}^{K} q(c_i) \ln q(c_i) \\ &+ \sum_{i=1}^{n}\sum_{c_i=1}^{K}\int q(c_i)q(\pi) \ln p(c_i|\pi)d\pi + \int q(\pi) \ln \frac{p(\pi)}{q(\pi)}d\pi. \end{aligned} \tag{74}$$

Konkretan oblik funkcionala $ELBO$ zavisi od pretpostavki modela (apriornih raspodela $p(\pi)$ i $p(Z_k)$, kao i raspodele mešavine $p(X_i|Z_{c_i})$). Potrebno je još odrediti generalan oblik pravila ažuriranja varijacionih parametara.

$$\begin{aligned} q(Z_k) &\propto p(Z_k)e^{\sum_{i=1}^{n} q(c_i=k) \ln p(X_i|Z_k)}, \\ q(c_i) &\propto e^{\sum_{c_{-i}} \prod_{j\neq i} q(c_j) \ln p(c_i|c_{-i})} e^{\int_{Z_{c_i}} q(Z_{c_i}) \ln p(X_i|Z_{c_i})dZ_{c_i}}. \end{aligned} \tag{75}$$

Detalji izvođenja pravila ažuriranja i evaluacije $ELBO$ prevazilaze opsege ovog rada, ali su detaljnije obrađeni u [4].

Upotrebom (74) i (75) i definisanjem pomenutih pretpostavki, strukturirano varijaciono zaključivanje sa koordinatnom optimizacijom varijacionih parametara se može primeniti na neparametarske modele mešavina.

6 Poređenje varijacionog zaključivanja sa alternativnim metodima

Radi potpunijeg razumevanja prednosti i mana varijacionog zaključivanja, predstavićemo i dva metoda koji su česte alternative pri rešavanju istog problema - algoritam očekivanje-maksimizacija i Monte Karlo metodi zasnovani na Markovljevim lancima,

a potom i uporediti Monte Karlo metod zasnovan na Markovljevim lancima sa varijacionim zaključivanjem na primeru linearne regresije.

Algoritam očekivanje-maksimizacija (eng. *Expectation-Maximization - EM* algoritam) je iterativan metod upotrebljiv za nalaženje (lokalne) ocene maksimalne verodostojnosti ili MAP ocene skrivenih promenljivih, onda kada to nije moguće tradicionalnim metodima. Vrlo je popularan za određene klase statističkih modela (skriveni Markovljevi modeli, modeli mešavina i drugi), zbog svoje jednostavnosti i mogućnosti da se sa relativno malo muke dođehttps://www.spe.org/events/en/2021/conference/21rptc/schedule-overview.html do lokalnog maksimuma tražene aposteriorne raspodele.

Monte Karlo metodi zasnovani na Markovljevim lancima su izuzetno popularan metod za aproksimaciju nepoznate gustine, koji kreira Markovljev lanac čija je stacionarna gustina raspodele upravo tražena. Ova klasa metoda je tema brojnih istraživanja od sredine dvadesetog veka, i naprednije metode ove klase se ubrajaju u najuspešnije metode za ocenu nepoznate gustine. Njegova mana je spora konvergencija, koja se naročito ispoljava u onim slučajevima kada nije esencijalno doći do globalnog maksimuma aposteriorne raspodele. Bazični metod iz ove klase, Metropolis-Hejstings algoritam, će biti opisan u poglavlju 6.2.

6.1 Algoritam očekivanje-maksimizacija

Koristeći se sličnom notacijom kao i do sada, neka je X dati uzorak, i Z skup skrivenih promenljivih. Dalje, neka je ω skup parametara pretpostavljene raspodele skrivenih promenljivih. Cilj algoritma očekivanje-maksimizacija je upravo ocena ovih parametara.

U najopštijem slučaju, za logaritam funkcije verodostojnosti parametara ω važi:

$$\log \mathcal{L}(\omega) = \sum_{i=1}^{N} \log p_\omega(x_i) = \sum_{i=1}^{N} \log \int p_\omega(x_i, Z) dZ. \tag{76}$$

Zbog činjenice da je Z skup skrivenih promenljivih, logaritam verodostojnosti se može smatrati slučajnom veličinom, te se može razmatrati i njegovo očekivanje. Algoritam očekivanje-maksimizacija problem nalaženja globalnog maksimuma logaritma funkcije verodostojnosti zamenjuje maksimizacijom njegovog očekivanja, i to cikličnim ponavljanjem dva koraka - koraka očekivanja ili E koraka, i koraka maksimizacije ili M koraka (otud i inspiracija za ime algoritma).

Polazeći od nasumično zadate vrednosti parametara ω, ω_0, smenjivanjem E i M koraka se vrednosti ažuriraju, kretanjem ka maksimumu očekivanja logaritma funkcije verodostojnosti. Za neko ω_{t-1}, cilj E koraka je nalaženje očekivanja:

$$E(\log \mathcal{L}(\omega)|\omega_{t-1}) = \int \log \mathcal{L}(\omega) p_{\omega_{t-1}}(z) dZ, \tag{77}$$

dok je cilj M koraka njegova maksimizacija po ω:

$$\omega_t = \arg\max_{\omega} E(\log \mathcal{L}(\omega)|\omega_{t-1}). \tag{78}$$

Iako maksimizacija očekivanja intuitivno odgovara maksimizaciji logaritma funkcije verodostojnosti, dokazaćemo korektnost algoritma očekivanje-maksimizacija. Neka je ω_{n-1} ocena parametara ω u $n-1$-toj iteraciji. Težimo da nova ocena ω_n bude

ona koja maksimizuje razliku $\Delta\mathcal{L}(\omega_n, \omega_{n-1})$:

(79)
$$\begin{aligned}\Delta\mathcal{L}(\omega_n, \omega_{n-1}) &= \log \mathcal{L}(\omega_n) - \log \mathcal{L}(\omega_{n-1}) \\ &= \log \int p(X, Z|\omega_n) dZ - \log \mathcal{L}(\omega_{n-1}) \\ &= \log \int p(X|Z, \omega_n) p(Z|\omega_n) dZ - \log \mathcal{L}(\omega_{n-1}) \\ &= \log \int p(X|Z, \omega_n) p(Z|\omega_n) \frac{p(Z|X, \omega_{n-1})}{p(Z|X, \omega_{n-1})} dZ - \log \mathcal{L}(\omega_{n-1}) \\ &= \log \int p(Z|X, \omega_{n-1}) \frac{p(X|Z, \omega_n) p(Z|\omega_n)}{p(Z|X, \omega_{n-1})} dZ - \log p(X|\omega_{n-1}) \\ &\geq \int p(Z|X, \omega_{n-1}) \log \frac{p(X|Z, \omega_n) p(Z|\omega_n)}{p(Z|X, \omega_{n-1}) p(X|\omega_{n-1})} dZ,\end{aligned}$$

odnosno:
(80)
$$\log \mathcal{L}(\omega_n) \geq \log \mathcal{L}(\omega_{n-1}) + \int p(Z|X, \omega_{n-1}) \log \frac{p(X|Z, \omega_n) p(Z|\omega_n)}{p(Z|X, \omega_{n-1}) p(X|\omega_{n-1})} dZ.$$

Definišimo:
(81)
$$l(\omega_n|\omega_{n-1}) = \log \mathcal{L}(\omega_{n-1}) + \int p(Z|X, \omega_{n-1}) \log \frac{p(X|Z, \omega_n) p(Z|\omega_n)}{p(Z|X, \omega_{n-1}) p(X|\omega_{n-1})} dZ,$$

radi kompaktnijeg zapisa:

(82)
$$\log \mathcal{L}(\omega_n) \geq l(\omega_n|\omega_{n-1}).$$

Dakle, $l(\omega_n|\omega_{n-1})$ je ograničena odozgo logaritmom funkcije verodostojnosti, i primetimo da važi:

(83)
$$\begin{aligned}l(\omega_{n-1}|\omega_{n-1}) &= \log \mathcal{L}(\omega_{n-1}) + \int p(Z|X, \omega_{n-1}) \log \frac{p(X, Z|\omega_{n-1})}{p(X, Z|\omega_{n-1})} dZ \\ &= \log \mathcal{L}(\omega_{n-1}).\end{aligned}$$

Prema tome, izborom sledeće vrednosti $\omega_n = \omega_{n-1}$ logaritam verodostojnosti se ne menja, a izbor ω_n koji povećava vrednost funkcije $l(\omega_n|\omega_{n-1})$ indirektno povećava i vrednost logaritma funkcije verodostojnosti.

Kako je:

(84)
$$\begin{aligned}\hat{w_n} &= \arg\max_{w_n} l(w_n|w_{n-1}) \\ &= \arg\max_{w_n} \int p(Z|X, \omega_{n-1}) \log \frac{p(X|Z, \omega_n) p(Z|\omega_n)}{p(Z|X, \omega_{n-1}) p(X|\omega_{n-1})} dZ \\ &= \arg\max_{w_n} \int p(Z|X, \omega_{n-1}) \log p(X, Z|w_n) dZ \\ &= \arg\max_{w_n} E(\log \mathcal{L}(\omega)|\omega_{t-1}),\end{aligned}$$

iterativnom primenom opisanih E i M koraka se povećava vrednost logaritma funkcije verodostojnosti.

Naravno, opisanim postupkom se dolazi do aproksimacije ocene maksimalne verodostojnosti parametara ω. Metod se može preformulisati tako da aproksimira i modu aposteriorne raspodele parametara (MAP ocenu) [10].

Opisana svojstva algoritma očekivanje-maksimizacija demonstriraju njegovu upotrebljivost, jednostavnu postavku i odsustvo kompleksnih kalkulacija, za mnoge probleme bajesovske statistike. Ipak, odlikuju ga spora konvergencija, koja se često završava u lokalnom optimumu, kao i nemogućnost merenja kvaliteta ocene. Ne pruža mogućnost predviđanja intervalnih ocena parametara, te ocene dobijene ovim metodom često treba uzimati sa rezervom, i njihov kvalitet proceniti drugim metodima. Ova karakteristika čini algoritam očekivanje-maksimizacija inferiornim u odnosu na varijaciono zaključivanje i Monte Karlo metode zasnovane na Markovljevim lancima, i upotrebljivim samo u slučajevima kada ocena raspodele nije od značaja.

6.2 Monte Karlo metodi zasnovani na Markovljevim lancima

Tokom rada, u više navrata je naglašena široka primena metoda za aproksimaciju aposteriorne raspodele putem uzorkovanja - naročito Monte Karlo metoda zasnovanih na Markovljevim lancima (eng. *Monte Carlo Markov Chain*, u daljem tekstu MCMC). Ova klasa metoda zamajac dobija još 90-ih godina 20. veka, od kada je predmet temeljnog izučavanja, kako u teorijskom, tako i u praktičnom smislu, te se može smatrati i najrazvijenijom klasom metoda za ocenu aposteriorne raspodele.

MCMC metodi su u stanju da generišu uzorak iz nepoznate raspodele, uz neophodnu mogućnost izračunavanja nepoznate gustine u proizvoljnoj tački. Pre formalnijeg definisanja varijanti MCMC metoda, ilustrujmo ideju uzorkovanja iz nepoznate raspodele preko primera (v. [24]).

Pretpostavimo da profesor želi da sazna raspodelu rezultata ispita u populaciji studenata. Neka profesor zna da rezultati imaju normalnu raspodelu sa disperzijom 15, ali mu je srednja vrednost nepoznata, i želi da je oceni putem MCMC algoritma. Pretpostavimo i da je profesor pregledao ispit jednog studenta, koji je imao 50 bodova na ispitu. Dakle, početna ocena srednje vrednosti je 50, tj. ocena raspodele je $\mathcal{N}(50, 15)$. Tada se algoritam sprovodi sledećim koracima:

1. Generisanje prve uzorkovane vrednosti - ova vrednost se generiše iz raspodele $\mathcal{N}(50, 15)$; pretpostavimo da je generisana vrednost 57;

2. Generisanje predloga nove vrednosti - ova vrednost se dobija dodavanjem slučajnog šuma na aktuelnu vrednost. Definišimo slučajan šum kao slučajnu promenljivu sa raspodelom $\mathcal{N}(0, 5)$. Predlog nove vrednosti se dobija sabiranjem uzorka šuma sa poslednjom generisanom vrednosti; pretpostavimo da je predlog nove vrednosti 54;

3. Generisanje verovatnoće prihvatanja - u ovom koraku se porede verodostojnosti aktuelne i predložene vrednosti. Izračunavanjem vrednosti tačke iz poznatog uzorka (poeni sa pregledanog ispita - 50) pod pretpostavkama raspodele $\mathcal{N}(57, 15)$ i $\mathcal{N}(54, 15)$. Novi predlog se usvaja ili odbacuje, u skladu sa verovatnoćom prihvatanja predloga, koja je proporcionalna količniku verodostojnosti raspodele sa novim predlogom srednje vrednosti, i verodostojnosti raspodele sa poslednjom generisanom srednjom vrednošću. Količnik se tipično poredi sa slučajno generisanom vrednošću, iz intervala $(0, 1)$. U ovom slučaju, količnik

je približno jednak $\frac{0.39}{0.32} \approx 1.22$, te bi novi predlog bio usvojen. Cilj ovog koraka je omogućavanje novim elementima koji daju veće vrednosti verodostojnosti inkluziju u lanac, i obrnuto;

4. Ako je generisani predlog prihvaćen, uzima se za poslednji element lanca, i verodostojnost novog predloga sa poredi sa njegovom. U suprotnom, prethodno generisana vrednost ostaje aktuelna.

Ovim koracima je opisana jedna iteracija Metropolis-Hejstings algoritma, najpoznatijeg primerka porodice MCMC algoritama. Iteracije se nastavljaju od drugog koraka, dok se ne dosegne zadovoljavajuća veličina uzorka.

Primer ilustruje i poreklo imena ove klase metoda, odnosno komponente koje ga sačinjavaju. Monte Karlo komponenta se ogleda u upotrebi slučajno generisanog uzorka za nalaženje nepoznate srednje vrednosti, u skladu sa drugim Monte Karlo metodima. Sa druge strane, primetimo da nova vrednost uzorka zavisi isključivo od prethodne - jednom kada se usvoji nova vrednost, stare vrednosti ne učestvuju u verovatnoći prihvatanja, te ne utiču na nove vrednosti uzorka. Dakle, opisanim postupkom se kreira lanac koji ispunjava Markovljevo svojstvo.

Ilustrujmo opisani metod i na formalniji način. Za ocenu aposteriorne raspodele $p(Z|X)$, gde je Z skup skrivenih promenljivih, a X dati uzorak, generiše se Markovljev lanac, iste dimenzionalnosti kao i skup skrivenih promenljivih, a vrednosti lanca se tretiraju kao uzorak iz tražene raspodele. Neka je q unapred zadata raspodela, tzv. *raspodela predloga*, kojom se postojeći elementi lanca modifikuju u nove, i neka je Z_0 skup inicijalnih vrednosti lanca. Tipično, zahteva se da je raspodela predloga simetrična oko nule, mada postoje varijante Metropolis-Hejstings algoritma koje koriste asimetričnu raspodelu predloga, i modifikovani korak prihvatanja. Radi definisanja algoritma u svojoj osnovnoj formi (tzv. Metropolis-Hejstings algoritam sa slučajnim lutanjem), pretpostavićemo da je raspodela predloga normalna sa očekivanjem $0, \mathcal{N}(0, \sigma^2)$. Tada se Metropolis-Hejstings sprovodi kroz sledeće korake, za iteraciju t, i zadatu vrednost lanca u prethodnoj iteraciji, Z_{t-1}:

1. $Z' \sim \mathcal{N}(0, \sigma^2)$
2. $\alpha_t = \frac{p(X|Z')}{p(X|Z_{t-1})}$
3. $U_t \sim \mathcal{U}[0, 1]$
4. $Z_t = \begin{cases} Z', & \alpha_t \geq U_t \\ Z_{t-1}, & \text{inače.} \end{cases}$

Zahtevani broj elemenata u lancu je hiperparametar koji se uobičajeno određuje pre generisanja lanca, i direktno utiče na performanse algoritma, u smislu broja iteracija i brzine.

Već se iz uvodnog primera mogu uočiti neka od ograničenja MCMC metoda. Najpre, očigledna je zavisnost lanca od disperzije raspodele šuma koji se dodaje aktuelnim vrednostima. Neadekvatan odabir ovog parametra lako dovodi do vrednosti lanca koje su besmislene (recimo, negativnih poena na ispitu studenata iz primera), u slučaju kada je uzeta vrednost prevelika, ili pak, premalog rasejanja vrednosti generisanog lanca, i spore konvergencije ka traženoj raspodeli. Takođe, MCMC metodi neretko pate od sporog zagrevanja (eng. *burn-in*), koji se ogleda u nereprezentativnosti prvih nekoliko vrednosti lanca, te se one tipično odbacuju. Broj početnih elemenata lanca je takođe

parametar koji valja pažljivo podesiti. Ovaj problem se može prevazići i opreznijim izborom prve vrednosti lanca. Teoretski, vrednosti bliže modi aposteriorne raspodele su pogodnije inicijalne vrednosti, te se, u praksi, inicijalna vrednost može dobiti i primitivnijom ocenom mode aposteriorne raspodele, ili se koristi više lanaca sa različitim inicijalnim vrednostima, a onda se bira onaj najstabilniji, čime se utvrđuje pogodnija inicijalna vrednost. Navedeni su neki od faktora koji MCMC metode čine zahtevnim za dijagnostiku, zbog kompleksnog procesa podešavanja parametara, koji često zahteva iskustvo u radu sa ovom klasom metoda, ali i vremenske kompleksnosti, zbog potrebe za generisanjem većeg broja lanaca sa velikim brojem iteracija, radi ublažavanja spomenutih problema.

6.3 Poređenje metoda za ocenu aposteriorne raspodele na primeru linearne regresije

Konačno, uporedićemo metod varijacionog zaključivanja sa opisanim alternativama, na primeru linearne regresije. Metodi će biti implementirani u programskom jeziku R, te ćemo, pored tačnosti aproksimacije, kao kriterijum poređenja posmatrati i brzinu konvergencije.

Pravila ažuriranja varijacionih parametara i kalkulacija $ELBO$ funkcionala u slučaju linearne regresije kod strukturiranog varijacionog zaključivanja su već opisani, u poglavlju 5.1.

Krenimo od jednostavnog sintetičkog primera. Neka je X dati uzorak, sačinjen od pet prediktora, sa 1000 elemenata, a Y ciljna promenljiva data jednostavnim linearnim modelom $Y = -0.5 + 2X_1 - 3X_2 - 1.5X_3 + X_4 - 5.5X_5$, sa parametrom preciznosti $\tau = 0.5$:

```
set.seed(42)
n     <- 1000          # Broj elemenata uzorka
d     <- 5            # Broj prediktora
# Koeficijenti linearne regresije
coefs <- c(-0.5, 2, -3, -1.5, 1, -5.5)
tau  <- 0.5      # Parametar inverza disperzije

# Generisanje uzorka
x <- cbind(1, replicate(d, rnorm(n)))
# Generisanje ciljne promenljive
y <- x %*% coefs  + rnorm(n, sd = sqrt(1/tau))
```

Dalje, oslanjajući se na prikazane rezultate (64), (66) i (68), definišemo funkciju koja iterativno ažurira varijacione parametre i kalkuliše vrednost $ELBO$. Funkcija, kao ulazne parametre, prima uzorak, ciljnu promenljive, parametar preciznosti, inicijalne vrednosti varijacionih faktora, maksimalan broj iteracija optimizacije, parametar *eps* kojim se definiše kriterijum zaustavljanja, odnosno maksimalna razlika u uzastopnim vrednostima $ELBO$ kojom se optimizacioni proces prekida, i *verbose* parametar, koji omogućava monitoring optimizacionog procesa i štampanje vrednosti $ELBO$ u svakoj iteraciji. Izlaz funkcije je struktura koja sadrži sve neophodne rezultate. Promenljivama izlazne strukture m i S su dati vektor srednjih vrednosti i kovarijaciona matrica normalno raspodeljenih parametara linearnog modela β (66).

```
linreg_vi_fit <- function(X, Y, tau,
                      a_0, b_0,
                      max_iter = 500, eps = 1e-5,
```

```
                        verbose = TRUE)
{
  X = as.matrix(X)
  D = ncol(X)
  N = nrow(X)
  XX <- crossprod(X, X)      # Kalkulacija XtX
  XY <- crossprod(X, Y)      # Kalkulacija XtY
  YY <- c(crossprod(Y))      # Kalkulacija YtY

  # Vektor vrednosti ELBO kroz iteracije
  elbo_list <- rep(-Inf, max_iter)

  alpha <- a_0 + D / 2

  E_tau_0 = a_0 / b_0

  for (i in 2:max_iter)
  {

    S <- solve(diag(E_tau_0, D) + tau * XX)

    m <- tau * S %*% XY

    E_bb <- as.vector(crossprod(m) + matrix.trace(S))

    beta <- b_0 + E_bb / 2

    E_tau <- alpha / beta

    elbo_py <- 0.5 * N * log(tau / (2 * pi))
    - 0.5 * tau * YY +
      tau * crossprod(m, XY) - 0.5 * tau *
      matrix.trace(XX %*% (tcrossprod(m, m) + S))
    elbo_pb <- -0.5 * D * log(2*pi) + 0.5 * D *
    (digamma(alpha) - log(beta)) -
      0.5 * E_bb * E_tau
    elbo_pt <- a_0*log(b_0) + (a_0 - 1) *
    (digamma(alpha) - log(beta))
    - b_0*E_tau - log(gamma(a_0))
    elbo_qb <- -0.5 * log(det(S)) - 0.5 * D *
    (1 + log(2*pi))
    elbo_qt <- -lgamma(alpha) + (alpha - 1)*digamma(alpha)
    + log(beta) - alpha

    elbo_list[i] <- elbo_py + elbo_pb + elbo_pt
    + elbo_qb + elbo_qt

    if (verbose) { cat("Iteration:\t",i,"\tELBO:\t",
    elbo_list[i],"\tELBO_diffa:\t",
    elbo_list[i] - elbo_list[i - 1],"\n")}

    # Provera konvergencije
    if (elbo_list[i] - elbo_list[i - 1] < eps)
    { break }

    if (i == max_iter)
    {warning("Algoritam nije iskonvergirao.\n")}
  }

  obj <- structure(list(m = m, S = S, alpha = alpha,
  beta = beta, tau = tau, X = X, N = N, D = D,
  elbo_list = elbo_list[2:i]),
```

```
                    class = "vi_linreg")

  return(obj)

}
```

Kako je ranije obrazloženo, izbor inicijalnih vrednosti varijacionih parametara je od značajnog uticaja na konačnu ocenu, te se ova funkcija poziva više puta, sa različitim vrednostima parametara *a_0* i *b_0*, i bira se ona konfiguracija sa najvećom vrednosti $ELBO$.

```
max_elbo = -Inf
optimal_a_0 = 0
optimal_b_0 = 0
for (a_0 in c(1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 3, 5, 10))
{
  for (b_0 in c(1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 3, 5, 10))
  {
    vi_model <- linreg_vi_fit(X = x, Y = y, tau = tau,
                              a_0 = a_0, b_0 = b_0,
                              max_iter = 500, eps = 1e-7,
                              verbose = TRUE)
    elbo <- vi_model$elbo_list[length(vi_model$elbo_list)]
    if (elbo > max_elbo)
    {
      optimal_a_0 <- a_0
      optimal_b_0 <- b_0
      optimal_model <- vi_model
      max_elbo <- elbo
    }
  }
}
```

Rezultujuće ocene parametara m su:

	m
β_0	-0.467
β_1	1.979
β_2	-2.979
β_3	-1.435
β_4	0.972
β_5	-5.418

Parametrima m i S je data ocena raspodele parametara β, ali ćemo se, radi upotrebe jednostavne srednjekvadratne greške, osloniti na MAP ocenu parametara. Sledeća tabela sadrži konačne rezultate strukturiranog varijacionog zaključivanja, u smislu tačnosti aproksimacije i performansi:

Srednjekvadratna greška	1.174
Utrošeno vreme [s]	0.26

Primenimo i Metropolis-Hejstings algoritam na istim podacima. Kao što je opisano u prethodnom poglavlju, potrebno je definisati simetričnu raspodelu predloga nad parametrima koje valja oceniti (u ovom slučaju, to su parametri β), kao i nulte vrednosti lanca koji se generiše. Takođe, potrebno je definisati broj iteracija procesa, i broj početnih uzoraka koji se zanemaruju, radi prevazilaženja problema sporog

početka. U našem eksperimentu, definisana raspodela početka je normalna raspodela sa jediničnom disperzijom, broj iteracija je 10000, i zanemaruje se prvih 1000 elemenata lanca. Sledeća funkcija implementira Metropolis-Hejstings algoritam za definisanu aposteriornu i raspodelu predloga:

```
run_metropolis_MCMC <- function(startvalue, iterations){
  chain = array(dim = c(iterations+1,7))
  chain[1,] = startvalue

  for (i in 1:iterations){
    proposal = proposalfunction(chain[i,])

    probab = exp(posterior(proposal)
                 - posterior(chain[i,]))
    if (runif(1) < probab){
      chain[i+1,] = proposal
    }else{
      chain[i+1,] = chain[i,]
    }
  }
  return(chain)
}
```

Sledećom tabelom su dati rezultati Metropolis-Hejstings algoritma:

Srednjekvadratna greška	8.84
Utrošeno vreme [s]	0.63

Inferiornost Metropolis-Hejstings algoritma u odnosu na strukturirano varijaciono zaključivanje, i u smislu brzine, i u smislu tačnosti aproksimacije je evidentna. Zaključimo da je varijaciono zaključivanje jednostavniji algoritam za implementaciju, zbog manjeg broja hiperparametara koje treba podesiti, jednom kada su pravila za ažuriranja varijacionih parametara i kalkulaciju $ELBO$ funkcionala poznati. Ipak, zahtevnost ovog koraka je primetna u petom poglavlju, i brojni su modeli za koje ovaj korak i dalje nije prevaziđen, zbog potrebe za računanjem kompleksnih integrala.

Iako je u demonstriranom primeru varijaciono zaključivanje pokazalo superiornost, napomenimo da postoje brojne naprednije varijante metoda iz MCMC klase, koji nude bolje performanse od Metropolis-Hejstings algoritma, ali prevazilaze opseg ovog rada. Konačno, naglašavamo da jednostavan recept ne postoji, već se savetuje upotreba oba algoritma, barem u slučajevima gde računska kompleksnost varijacionog zaključivanja nije preterana.

7 Zaključak

U radu je predstavljena lepeza metoda varijacionog zaključivanja, od najjednostavnijeg - aproksimacije srednjeg polja, do kompleksnijih, poput strukturiranog i neparametarskog varijacionog zaključivanja. Svaki metod se može evaluirati kroz izbor varijacione familije, kojim se postiže balans između fleksibilnosti familije (a time i kvaliteta aproksimacije aposteriorne raspodele) i jednostavnosti optimizacionog postupka, odnosno brzine konvergencije. Ovi faktori, zajedno sa prirodom parametara čija se aposteriorna raspodela ocenjuju, su predviđeni da pomognu čitaocu u izboru metoda kojim će se poslužiti. Takođe, prikazan je metod stohastičkog varijacionog zaključivanja, kojim se, po potrebi, optimizacioni postupak može ubrzati. Pregledom nekoliko modela bajesovske statistike, i primenom varijacionog zaključivanja u oceni njihovih skrivenih promenljivih, čitaocu su ponuđeni primeri, koji približavaju slučajeve u kojima se opisani metodi mogu upotrebiti. Uz teorijsko izvođenje pravila za ažuriranje varijacionih parametara pojedinih modela, priložen je i programski kod u jeziku R, koji predstavlja dodatnu olakšicu u primeni ovih metoda u praksi. Konačno, razrađeni su i metodi koji predstavljaju alternativu varijacionom zaključivanju - algoritam očekivanje-maksimizacija i Monte Karlo metodi zasnovan na Markovljevim lancima. Uvidom u njihova svojstva, dobijena je široka ponuda metoda za ocenu aposteriorne raspodele, ali je i konstatovano da zlatni standard ne postoji, te da svaki od metoda ima spektar slučajeva ocenjivanja parametara bajesovskih modela u kom ga je poželjno primeniti. Zaključimo da se čitalac upućuje na literaturu i upotrebu varijacionog zaključivanja svaki put kada je potrebno oceniti parametre metoda koji su već obrađeni u kontekstu ovog metoda, zbog činjenice da postupak nalaženja pravila ažuriranja varijacionih parametara i kalkulaciju $ELBO$ funkcionala može biti jako kompleksan i vremenski zahtevan za pojedine modele. Stoga, skup bajesovskih modela kod kojih je varijaciono zaključivanje primenljivo u praksi je ograničen, ali dovoljno širok da ovaj metod zaslužuje svoje mesto u arsenalu bajesovskih statističara, zbog široke rasprostranjenosti modela u kojima je ovaj metod već pokazao zavidne rezultate.

Literatura

[1] Bickel, Peter J., and Kjell A. Doksum. Mathematical statistics: basic ideas and selected topics, volumes I-II package. CRC Press, 2015.

[2] Bishop, Christopher M. "Pattern recognition." Machine learning 128.9 (2006).

[3] Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

[4] Blei, David M., and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." Bayesian analysis 1.1 (2006): 121-143.

[5] Borman, Sean. "The expectation maximization algorithm-a short tutorial." Submitted for publication 41 (2004).

[6] Drugowitsch, Jan. "Variational Bayesian inference for linear and logistic regression." arXiv preprint arXiv:1310.5438 (2013).

[7] Gershman, Samuel J., and David M. Blei. "A tutorial on Bayesian nonparametric models." Journal of Mathematical Psychology 56.1 (2012): 1-12.

[8] Gershman, Samuel, Matt Hoffman, and David Blei. "Nonparametric variational inference." arXiv preprint arXiv:1206.4665 (2012).

[9] Grimmer, Justin. "An introduction to Bayesian inference via variational approximations." Political Analysis 19.1 (2011): 32-47.

[10] Gupta, Maya R., and Yihua Chen. Theory and use of the EM algorithm. Now Publishers Inc, 2011.

[11] Hoffman, Matthew D., et al. "Stochastic variational inference." Journal of Machine Learning Research 14.5 (2013).

[12] Hoffman, Matthew D., and David M. Blei. "Structured stochastic variational inference." Artificial Intelligence and Statistics. 2015.

[13] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[14] Knoblauch, Jeremias, Jack Jewson, and Theodoros Damoulas. "Generalized variational inference: Three arguments for deriving new posteriors." arXiv preprint arXiv:1904.02063 (2019).

[15] Kurihara, Kenichi, Max Welling, and Yee Whye Teh. "Collapsed Variational Dirichlet Process Mixture Models." IJCAI. Vol. 7. 2007.

[16] Lu, Chenguang. "Understanding and Accelerating EM Algorithm's Convergence by Fair Competition Principle and Rate-Verisimilitude Function." arXiv preprint arXiv:2104.12592 (2021).

[17] Minka, Thomas P. "Expectation propagation for approximate Bayesian inference." arXiv preprint arXiv:1301.2294 (2013).

[18] Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.

[19] Nikolić Mladen, Zečević Anđelka. "Mašinsko učenje". 2019.

[20] Ostwald, Dirk. "Variational Inference". (2019).

[21] Rasmussen, Carl Edward. "The infinite Gaussian mixture model." NIPS. Vol. 12. 1999.

[22] Roberts, Stephen J., et al. "Bayesian approaches to Gaussian mixture modeling." IEEE Transactions on Pattern Analysis and Machine Intelligence 20.11 (1998): 1133-1142.

[23] Salimans, Tim, Diederik Kingma, and Max Welling. "Markov chain monte carlo and variational inference: Bridging the gap." International Conference on Machine Learning. PMLR, 2015.

[24] Van Ravenzwaaij, Don, Pete Cassey, and Scott D. Brown. "A simple introduction to Markov Chain Monte–Carlo sampling." Psychonomic bulletin & review 25.1 (2018): 143-154.

[25] Zhang, Cheng, et al. "Advances in variational inference." IEEE transactions on pattern analysis and machine intelligence 41.8 (2018): 2008-2026.

[26] Zhao, Hui. "Variational Bayesian Learning and its Applications." (2014).

[27] https://rpubs.com/cakapourani/variational-bayes-lr

[28] https://www.coursera.org/learn/bayesian-methods-in-machine-learning/home/week/3

[29] https://khayatrayen.github.io/MCMC.html

[30] https://www.cs.princeton.edu/courses/archive/fall07/cos597C/scribe/20070921.pdf

Biografija

Marko Radosavljević je rođen u Valjevu 3. oktobra 1994. godine, gde završava Osnovnu školu "Andra Savčić", sa Vukovom diplomom. Završio je Valjevsku gimnaziju, specijalizovano matematički smer, sa Vukovom diplomom. 2013. godine upisuje Matematički fakultet u Beogradu, na modulu Statistika, aktuarska i finansijska matematika. Osnovne akademske studije završava u julu 2017. godine, sa prosečnom ocenom 9.64. 2017. godine upisuje master studije, na istom modulu. Položio je sve ispite na master studijama, sa prosečnom ocenom 9.33. Od 2018. godine, svoju profesionalnu karijeru gradi u oblastima nauke o podacima, mašinskog učenja i digitalne obrade signala.