SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

Filip Bilandžija

Utjecaj izbora razreda u histogramu na signifikantnost signala

Zagreb, 2022.

Ovaj je rad izrađen u Laboratoriju za fiziku elementarnih čestica na Institutu Ruđer Bošković u Zagrebu pod komentorstvom izv. prof. dr. sc. Vuke Brigljevića te je pod mentorstvom prof. dr. sc. Mirka Planinića predan na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2021./2022.

Sadržaj

1 Uvod

Motivacija za ovo istraživanje bio je studentski projekt na temu "Zajednička produkcija Higgsovog i teškog Z bozona". U tom su se projektu istraživali stvaranje Higgsovog bozona koji se raspada u 2 b kvarka zajedno sa Z bozonom koji se raspada u 2 leptona.

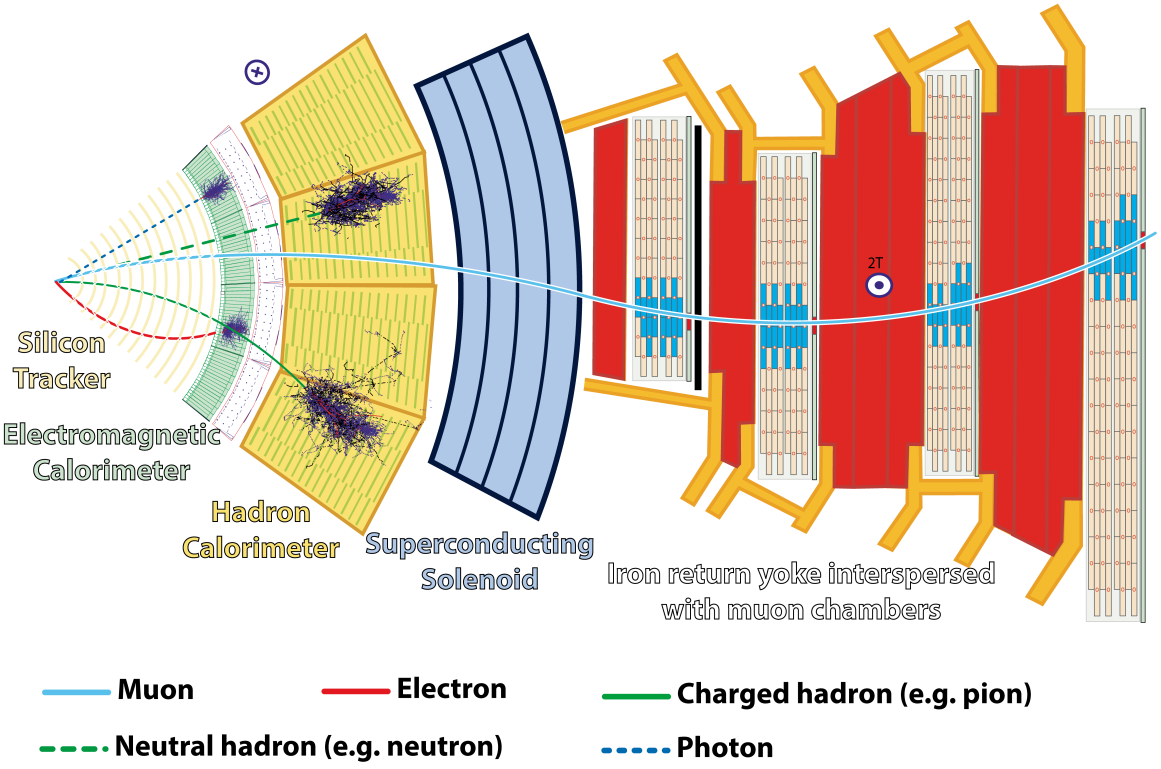
1.1 Proton-proton sudari

Na velikom hadronskom sudarivaču (engl. Large Hadron Collider, LHC), kao što mu i ime govori, sudaraju se hadroni, specifično protoni, ali i neki teški ioni poput iona olova i zlata. Pritom se snopovi protona ubrzavaju na brzinu blizu brzine svjetlosti, te se ukrštavaju na nekoliko mjesta duž sudarivača gdje se nalaze detektori. Jedan od tih detektora je i kompaktni mionski solenoid (engl. Compact Muon Solenoid, CMS) s kojega smo i proučavali podatke. Energija centra mase proton-proton sudara iznosila je $\sqrt{s} = 13$ TeV u razdoblju između 2016. i 2018. godine (od 2022. ona iznosi 13.6 TeV), što nam daje vrlo veliku energiju za kreiranje novih čestica na raspolaganju.

Proton je kompozitna čestica koja se sastoji od partona - kvarkova i gluona. Osim tri valentna kvarka (uud), ima i kvarkove mora (parovi kvarkova u, d, s, c, i u manjoj mjeri b) te gluone koje vežu te kvarkove. Pri sudaru protona, ne sudaraju se protoni kao cjeline, nego samo jedan od njegovih dijelova - partona, dok ostali bivaju raspršeni. Interakcija partona, osnovna reakcija, naziva se tvrdi proces (engl. hard process), te iz njega izlaze elementarne čestice.

Ako su te izlazne čestice kvarkovi ili gluoni, ne možemo ih detektirati same zbog svojstva bojnog zatočenja, nego ih detektiramo u vezanim stanjima - bezbojnim hadronima. Nakon tzv. tvrdog procesa u kojem kvarkovi nastaju, oni se vrlo brzo vežu u bezbojne hadrone, a ti se hadroni, jer su velikom većinom nestabilni, raspadaju u više drugih hadrona, pa naposljetku detektiramo veliki broj hadrona koji je nastao iz početnog kvarka. Tu nakupinu hadrona zovemo mlaz hadrona (engl. jet), a proces njegovog nastajanja hadronizacija. Točan proces hadronizacije još uvijek ne razumijemo u potpunosti[1], no postoje fenomenološki modeli inspirirani kvantnom kromodinamikom koji mogu objasniti eksperimentalne podatke [1].

[1] To je posljedica temeljnog problema kvantne kromodinamike: konstanta vezanja α_s je relativno velika na niskim energijama, pa račun nije moguće provesti perturbativno.



Slika 1: Transverzalni presjek CMS detektora. Preuzeto iz [2].

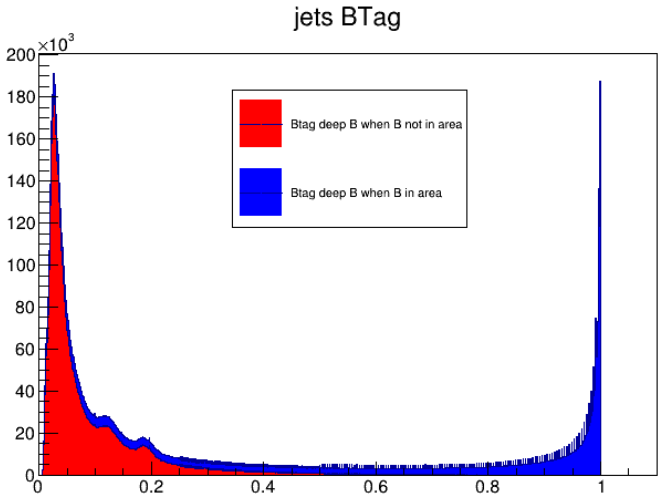
1.2 Detekcija čestica na CMS detektoru

CMS detektor jedan je od dva velika detektora opće namjene na LHC-u te je cilindričnog oblika čime zatvara skoro cijeli prostorni kut, što je iznimno važno za detekciju čestica. Njegov je presjek prikazan na slici 1. Sami je detektor izgrađen od više slojeva poddetektora.

U sloju najbližem točki sudara nalazi se silicijski detektor tragova koji mjeri putanju nabijenih čestica. U kombinaciji s magnetskim poljem od 3.8 T koje proizvodi supravodljiva zavojnica (solenoid), čime se nabijene čestice zakreću, detektor tragova može precizno odrediti impuls nabijenih čestica. U drugom se sloju nalaze kalorimetri - elektromagnetski i hadronski, detektori koji mogu precizno odrediti energiju čestica. U elektromagnetskom, unutarnjem, mjere se energije elektrona i fotona, dok se u hadronskom, vanjskom, kalorimetru mjere energije svih hadrona. Čestice se određuju prema mjestu detekcije i obliku signala koji ostavljaju u detektoru.

1.3 Etiketiranje b mlazova

Mlazove hadrona dobivamo tako da grupiramo čestice koje su međusobno "blizu". Kada definiramo što nam "blizu" znači, zapravo definiramo algoritam koji grupira čestice. Identificiranje kvarka ili gluona iz kojeg mlaz potječe posve je netrivijalan problem. Nas je zanimala samo identifikacija



Slika 2: Rezultat algoritma za etiketiranje b mlazova koji algoritam daje za mlazove koji u simulacijama imaju b kvark u blizini (plava) i za one koji nemaju (crvena).

mlazova poteklih iz b kvarkova - b mlazova, koja ima neke svoje posebnosti. Nakon tvrdog procesa iz kojeg izlazi b kvark, on brzo formira B hadron. Taj B hadron, iako nestabilan, putuje dovoljno dugo da možemo izmjeriti udaljenost njegovog mjesta raspada (engl. secondary vertex) od centra sudara koja obično iznosi oko nekoliko milimetara [3]. Uzimajući u obzir taj pomak u mjestu raspada, dobivamo već dobru separaciju mlazova poteklih iz b kvarka od ostalih. Najbolja se separacija b mlazova od ostalih mlazova postiže koristeći algoritam baziran na strojnom učenju, koji uzima, uz pomak mjesta raspada, još parametara u obzir poput broja hadrona u mlazu, udaljenost druge točke raspada i sl. te nam daje informaciju koliko je siguran da mlaz potječe iz b kvarka. Taj algoritam (engl. b-tagging), dakle, daje broj koliko je on siguran da li su ti mlazovi potekli iz b kvarka ili nisu, a mi pomoću tog rezultata stavljamo etiketu na mlazove izabirući određeni broj kao diskriminantu (u našem slučaju bio je 0.5).

Kao što vidimo sa slike 2, takav algoritam nam daje diskriminantu između b mlazova i ostalih mlazova, gdje su prikazane vjerojatnosti da nam mlaz potječe iz b kvarka od kojih su plavom prikazani mlazovi koji stvarno potječu iz b kvarka, a crvenom mlazovi koji potječu iz ostalih kvarkova ili gluona. Možemo primijetiti da je iznimno koristan alat u identifikaciji mlazova poteklih iz b kvarka.

1.4 Analiza podataka u fizici elementarnih čestica

U fizici elementarnih čestica, obično nema jednostavnog izraza pomoću kojeg možemo teoriju usporediti s eksperimentom. Zbog toga, eksperimentalne podatke uspoređujemo sa simulacijama koje

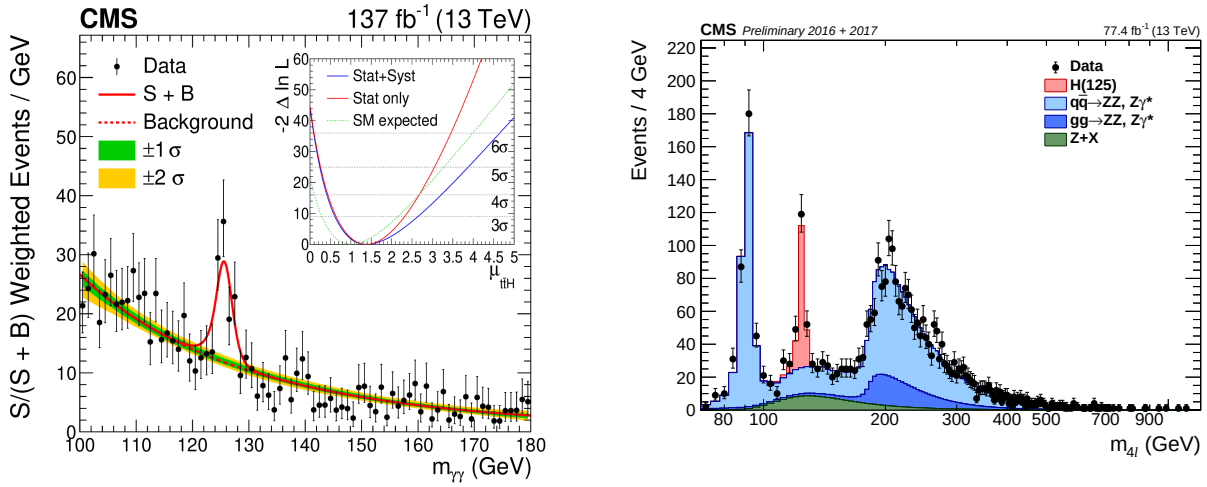
uzimaju u obzir teorijski dobivene veličine. Te simulacije ne generiraju samo događaje (npr. čestice nastale u sudarima). Čak štoviše, one simuliraju i kakav će signal detektor imati nakon prolaska generirane čestice kroz njega. Daljnja analiza rekonstruira događaje, pritom uzimajući u obzir i odgovor detektora u eksperimentu i simulirani odgovor detektora u simulacijama. Nakon rekonstrukcije događaja, slijedi statistička analiza u kojima se mjere učestalosti određenih procesa, testiraju hipoteze, daju procjene na granice teorija i, općenito, dolazi do novih spoznaja.

Simulacije se općenito odvijaju na 3 nivoa. Na prvom, najnižem, nivou gledamo partone - kvarkove i gluone - koji izlaze iz tvrdog procesa ili se raspršuju iz protona. Taj nivo simulacija odvija se pomoću algoritama stvaranja događaja (engl. event generators)[2]. Na drugom nivou gledamo fotone, leptone i generirane hadrone, nastale iz partona, koji će živjeti dovoljno dugo da ostave trag u detektoru te njih grupiramo u generirane hadronske mlazove. Na posljednjem nivou gledamo rekonstruirane objekte, gdje je simuliran odgovor detektora te je provedena računalna rekonstrukcija događaja. Dakle, na tom, posljednjem, nivou simuliran je odgovor detektora i to je jedini koji možemo izravno usporediti s eksperimentom. Prilikom same analize podataka, imamo pristup svakom nivou simulacija pojedinačno.

[2]Pod događajem se podrazumijeva prolazak jednog protonskog snopa kroz drugi pri kojem se dogodi bar 1 tvrdi proces.

Tablica 1: Tablica raspada Higgsovog bozona za masu $m_H = 125.6$ GeV. Preuzeto iz [4].

Kanal raspada	Udio ukupnih raspada [%]
$H \longrightarrow b\bar{b}$	57.1 ± 1.9
$H \longrightarrow WW^*$	22.0 ± 0.9
$H \longrightarrow gg$	8.53 ± 0.85
$H \longrightarrow \tau\tau$	6.26 ± 0.35
$H \longrightarrow c\bar{c}$	2.88 ± 0.35
$H \longrightarrow ZZ^*$	2.73 ± 0.11
$H \longrightarrow \gamma\gamma$	0.228 ± 0.011
$H \longrightarrow Z\gamma$	0.157 ± 0.014



Slika 3: Invarijantna masa sustava (lijevo) dva fotona i (desno) 4 leptona u analizama za $H \to \gamma\gamma$ i $H \to ZZ^* \to 4l$ [10, 11].

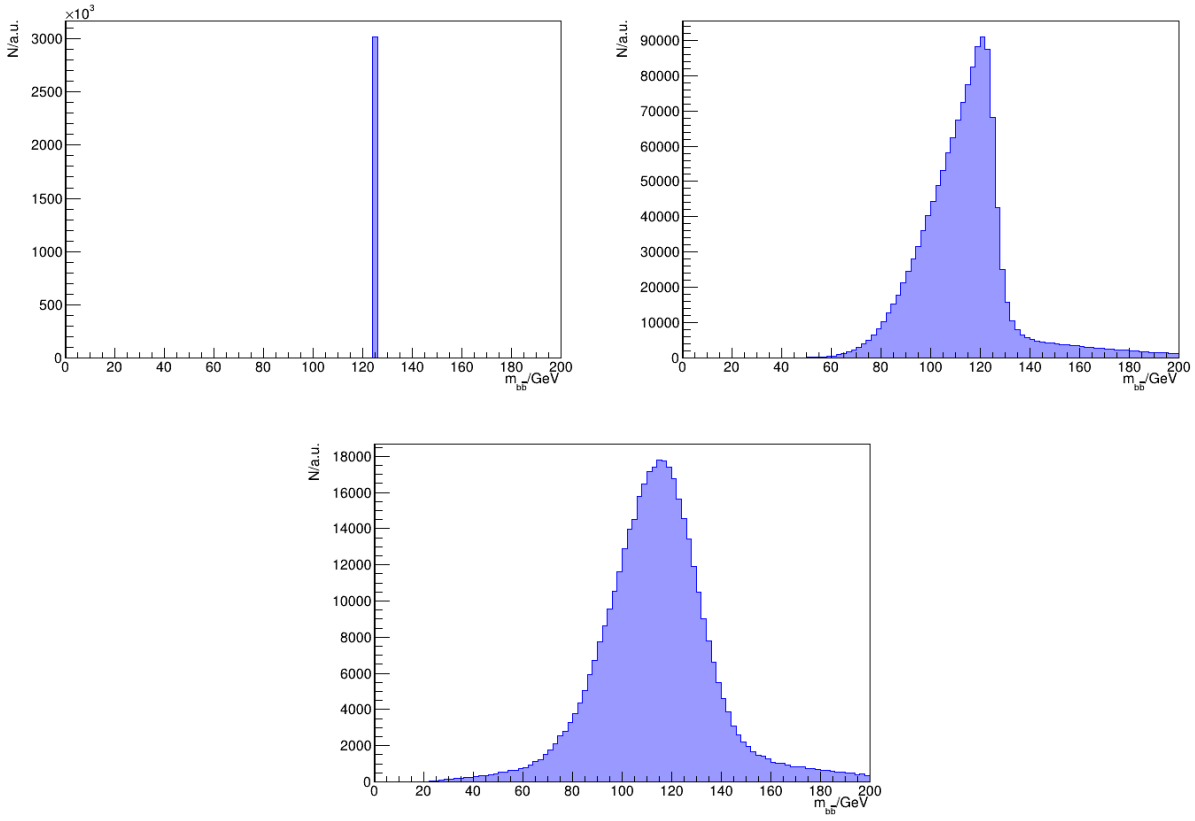
1.5 Raspadi Higgsovog bozona

Higgsov je bozon, kao i velika većina čestica, nestabilan te ima vrijeme poluživota reda veličine 10^{-22} s i masu od oko 125 GeV [5, 6]. Najvjerojatniji raspadi Higgsovog bozona prikazani su u tablici 1. Kao što vidimo daleko najvjerojatniji raspad je raspad u b kvark i antikvark. No, 2012. godine, kada je Higgsov bozon prvi puta otkriven, nije opažen u tom kanalu raspada, već u kanalima $H \longrightarrow \gamma\gamma$ i $H \longrightarrow ZZ^* \longrightarrow 4l$ prikazanima na slici 3. Kao što se i vidi iz tablice 1, ti su raspadi znatno manje vjerojatni od raspada u b kvark i antikvark.

Za masu Higgsovog bozona $m_H = 125$ GeV, očekuje se da će udio raspada u b kvark i antikvark biti oko 58% [7]. Precizno mjerenje udjela tog raspada omogućuje testiranje hipoteze da je interakcija nabijenih fermiona standardnog modela s Higgsovim poljem izvor generacije mase tih fermiona [8, 9].

Raspade općenito tražimo tako da iz čestica u konačnom stanju rekonstruiramo događaj i slijed raspada te time zaključimo koje su čestice bile u početnom stanju. To možemo postići traženjem vrhova u spektru invarijantne mase. Ako dvije čestice dolaze iz raspada početne čestice, invarijantna masa sustava tih dvaju čestica bit će jednaka masi početne čestice te će se u spektru vidjeti kao vrh.

Dva su razloga zašto je Higgsov bozon opažen u ta dva kanala raspada, a ne u glavnom, najvjerojatnijem, kanalu raspada: pozadina konačnog stanja i rezolucija detektora. Naime, pozadina je za 2 b kvarka daleko veća od odgovarajućih pozadina za 2 fotona ili 4 leptona, dok je rezolucija za mjerenje energija fotona i leptona oko 1 %, dok je za kvarkove, odnosno hadrone, čak oko 10 %. Stoga vrh koji dolazi od iste čestice će biti širi, pa ga je zbog toga i teže uočiti. Pritom rezoluciju u slučaju kvarkova

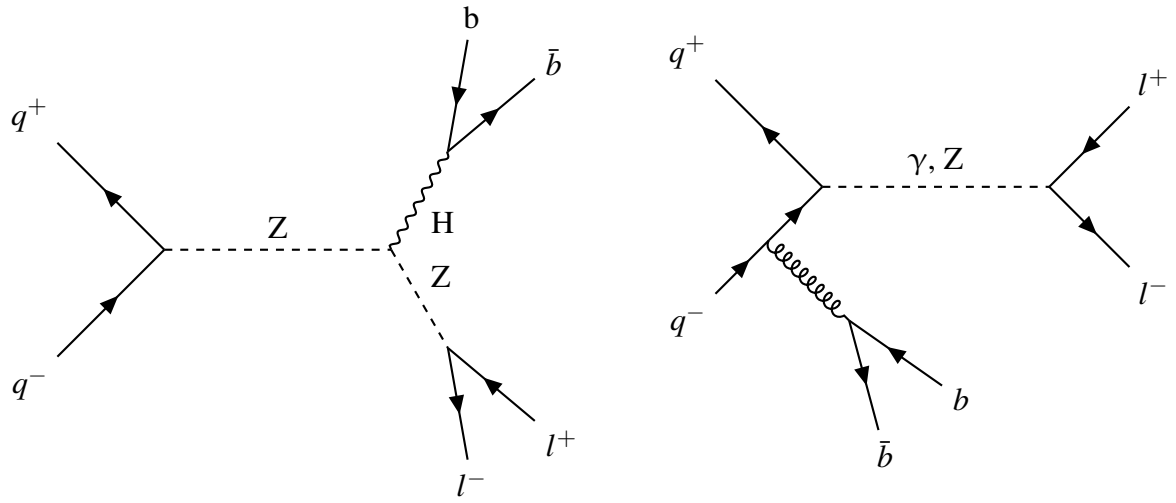


Slika 4: Utjecaj hadronizacije i detektora na rezoluciju. Histogrami nisu normalizirani na luminozitet. (Gore lijevo) Invarijantna masa sustava 2 b kvarka u najnižem nivou simulatora. (Gore desno) Invarijanta masa sustava 2 generirana hadronska mlaza s najvećim b-tagom. (Dolje) Invarijantna masa sustava 2 rekonstruirana hadronska mlaza s najvećim b-tagom.

ne definira samo eksperimentalni postav, nego i fizika sama.

Utjecaj hadronizacije na rezoluciju, kao i utjecaj detektora, prikazani su na slici 4. Na gornjem lijevom histogramu, prikazana je invarijanta masa sustava 2 b kvarka te vidimo da svi događaji imaju invarijantu masu od oko 125 GeV, što je i masa Higgsovog bozona. Gornji desni graf prikazuje invarijantnu masu sustava 2 generirana hadronska mlaza iz 2 b kvarka. Na tom histogramu možemo vidjeti utjecaj fizike, tj. hadronizacije na rezoluciju. Naime, tada se raspodjela invarijantnih masa širi. Utjecaj detektora možemo vidjeti na donjem histogramu, na kojem je prikazana invarijantna masa sustava 2 rekonstruirana hadronska mlaza. Raspodjela se dalje širi te poprima oblik gausijana.

Utjecaj rezolucije ne možemo značajnije popraviti, no ono na što možemo utjecati je odnos pozadine i signala, tj. odabir samog procesa koji promatramo. Kako bismo smanjili pozadinu s obzirom na snagu signala, umjesto promatranja konačnog stanja 2 b kvarka, možemo promatrati slučaj u kojemu je uz Higgsov bozon produciran i Z bozon pa konačno stanje koje tražimo, uz 2



Slika 5: Feynmanovi dijagrami s konačnim stanjem 2 b kvarka i 2 leptona za (lijevo) zajedničku tvorbu Higgsovog i Z bozona i (desno) Drell-Yan proces pri kojem se zrači gluon koji se raspada u 2 b kvarka.

b kvarka, ima i 2 leptona s invarijantnom masom teškog Z bozona. Time efektivno promatramo zajedničku tvorbu Higgsovog bozona i teškog Z bozona prikazanu Feynmanovim dijagramom na slici 5 lijevo. Iako su i CMS i ATLAS kolaboracije već mjerile raspad Higgsovog bozona u b kvarkove, ta mjerenja nisu imala dovoljno malu statističku nesigurnost da budu proglašena otkrićem [12, 13].

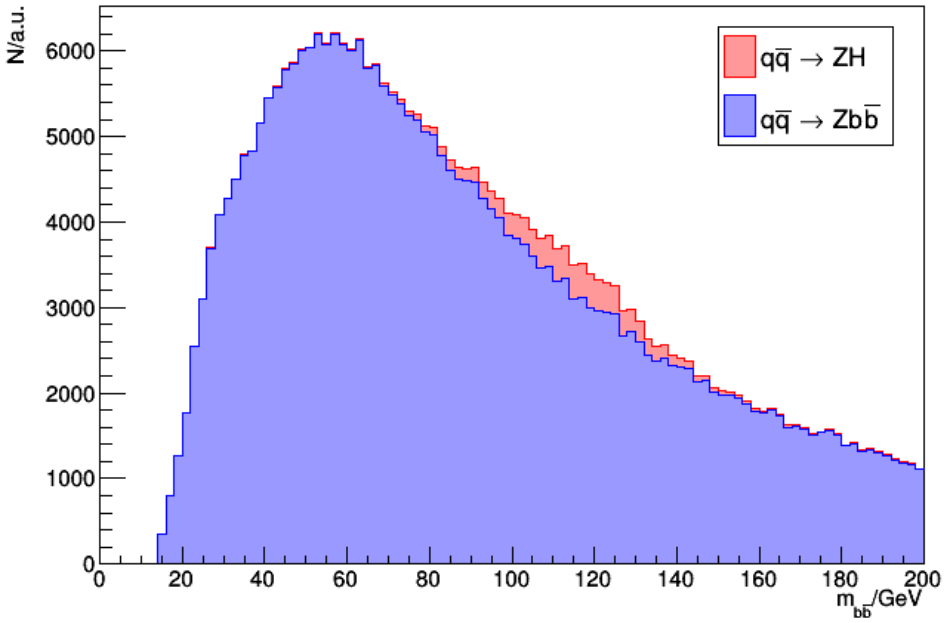
1.6 Pozadina

Glavna pozadina za proces u kojem tražimo u konačnom stanju 2 b kvarka (odnosno mlaza) i 2 leptona je tzv. Drell-Yan proces, prikazan na slici 5 desno. U Drell-Yan procesu iz sudarenog kvarka i antikvarka preko bozonskog se propagatora stvaraju 2 leptona. Pritom jedan od kvarkova može, prije samog tvrdog procesa, zračiti gluon koji se raspada u par b kvarkova. Iz tih kvarkova nastaju hadronski mlazovi koji će biti obilježeni kao nastali iz b-kvarka. Dakako, ima i drugih doprinosa procesu, ali ovaj koji smo upravo opisali je najdominantniji.

1.7 Ukupni histogram signala i pozadine

Nakon što smo proveli isti algoritam selekcije događaja i za signal i za pozadinu, s istim kinematičkim i b-tag kriterijima, te sve doprinose skalirali na luminozitet[3], dobili smo ukupni histogram

[3]Luminozitet je definiran kao faktor proporcionalnosti između broja događaja u vremenu i udarnog presjeka, koji predstavlja vjerojatnost da se određeni proces dogodi. Luminozitet je svojstvo sudarivača i ovisi o karakteristikama samog snopa koji se sudara.



Slika 6: Ukupni graf pozadine i signala u slučaju kada promatramo zajedničku tvorbu Higgsovog i Z bozona.

prikazan na slici 6. Kao što vidimo, čak i uz odabir zajedničke tvorbe Higgsovog i teškog Z bozona, signal nije ni izbliza jasan kao što je on bio u slučaju za raspade u 2 fotona ili 4 leptona. Iz slike je jasno da količina signala naspram pozadine nije velika te da će prilagodba podataka imati veliku nesigurnost, pa se može postaviti pitanje: možemo li drugačijim izborom razreda poboljšati prilagodbu i smanjiti statističku nesigurnost?

2 Statistika

2.1 Matematički formalizam

2.1.1 Vjerojatnost

Postoje više definicija vjerojatnosti, ali sve zadovoljavaju Kolmogorove aksiome [14]. Mi ćemo implicitno koristiti frekventističku. Neka je (Ω, $F \subseteq 2^{\Omega}$, P) prostor mjere, gdje je P funkcija koja preslikava elemente iz F, podskupa skupa 2^{Ω}, u realne brojeve. P je mjera vjerojatnosti ako su ispunjeni sljedeći uvjeti:

1. $P(E) \geq 0, \forall E \in F$
2. $P(\Omega) = 1$
3. $\forall (E_1, \ldots, E_n) \in F^n : E_i \cup E_j = 0, P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

Ova definicija vrijedi i u diskretnom i u kontinuiranom slučaju [15]. Varijabla koja poprima specifičnu vrijednost za svaki element skupa Ω zovemo slučajna varijabla. Pritom pojedini elementi mogu biti karakterizirani s više veličina, pa je u tom slučaju slučajna varijabla višedimenzionalni vektor [16].

2.1.2 Funkcija gustoće vjerojatnosti

Promotrimo neko mjerenje čiji je ishod određen kontinuiranom varijablom x koja može poprimiti vrijednosti u nekom intervalu. Vjerojatnost da je vrijednost koju poprimi unutar infinitezimalnog intervala $[x, x + \mathrm{d}x]$, dana je s funkcijom gustoće vjerojatnosti (engl. probability density function, p.d.f.) $f(x)$:

$$p(x \in [x, x + \mathrm{d}x]) = f(x)\mathrm{d}x. \tag{1}$$

Funkcija gustoće vjerojatnosti je normalizirana tako da je ukupna vjerojatnost svih ishoda jednaka jedan:

$$\int_S f(x)\mathrm{d}x = 1, \tag{2}$$

gdje je S interval svih mogućih ishoda za x [16].

Iako smo definirali funkciju gustoće vjerojatnosti (ili kraće, gustoću vjerojatnosti) za kontinuiranu varijablu, definicija za diskretnu varijablu nije bitno drugačija, ali pritom se ne koristi gustoća vjerojatnosti nego vjerojatnost. Umjesto da se mora promatrati neki interval kada se gleda vjerojatnost, za diskretnu varijablu se može promatrati i točno određena vrijednost. Također, kod normalizacije umjesto integrala pojavljuje se suma koja ide preko svih mogućih ishoda.

Nama će u ovome radu od posebne važnosti biti dvije gustoće vjerojatnosti: Poissonova raspodjela i Gaussova (normalna) raspodjela. Poissonova raspodjela je raspodjela diskretne varijable definirana po formuli:

$$P(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, \tag{3}$$

gdje je λ očekivana vrijednost, koja je ujedno i varijanca.

Poissonova je raspodjela iznimno bitna u fizici jer ona zapravo opisuje svaki eksperiment u kojem se mjeri broj događaja, primjerice broj raspada nestabilnog elementa u jedinici vremena. Očekivana vrijednost u eksperimentu može se odrediti mjereći te ju usporediti s teorijskom vrijednošću. Također, posjeduje i jako bitno svojstvo da je zbroj Poissonovih distribucija i sama Poissonova distribucija, gdje je pritom očekivana vrijednost zbroja zapravo zbroj očekivanih vrijednosti.

Gaussova raspodjela, za razliku od Poissonove, je raspodjela kontinuirane varijable te je dana formulom:

$$p(x;\mu,\sigma) = \frac{1}{2\pi\sigma^2}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \tag{4}$$

gdje je μ srednja vrijednost, dok je σ varijanca.

2.1.3 Procjenitelji

Za razliku od svakodnevnog života gdje može označavati aproksimaciju, procjena je u statistici tehnički pojam koji je precizno definiran. Ona označava preciznu i točnu proceduru koja vodi do rezultata i određivanja nesigurnosti tog rezultata [17].

Procjenitelj je procedura koja, kada se primjeni na podatke uzorka, daje numeričku vrijednost za svojstvo populacije ili, ako je prikladno, svojstvo ili parametar gustoće vjerojatnosti populacije. Ovo je posve općenita definicija, te se koristi kada želimo procjenu za populacije iz koje smo uzeli uzorak [17]. Mi ćemo se u ovome radu fokusirati na jedan takav procjenitelj: metodu najveće izglednosti.

2.1.4 Funkcija izglednosti

Metoda najveće izglednosti bazira se konstrukciji združene distribucije vjerojatnosti svih mjerenja u nekom skupu podataka.

Promotrimo slučajnu varijablu x raspodijeljenu po nekoj gustoći vjerojatnosti $f(x;\theta)$, gdje po pretpostavci znamo oblik gustoće vjerojatnosti, ali vrijednost barem jednog parametra θ (ili više njih $\vec{\theta} = (\theta_1,\ldots,\theta_m)$) je nepoznata. Metoda najveće izglednosti nam procjenjuje vrijednost parametara u slučaju konačnog uzorka podataka. Pretpostavimo da smo mjerili vrijednost slučajne varijable x n

puta, te da su nam rezultati tog mjerenja vrijednosti $x_1,\ldots,x_n$. Ovdje x nije nužno skalar, nego može biti i višedimenzionalni vektor gdje svaka komponenta vektora predstavlja drugačiju mjerenu veličinu [16].

Vjerojatnost i-tog mjerenja da bude u intervalu $[x_i, x_i + \mathrm{d}x_i]$ je $f(x_i;\theta)\mathrm{d}x_i$. Pod pretpostavkom da su mjerenja nezavisna, vjerojatnost prvog mjerenja da bude u intervalu $[x_1, x_1 + \mathrm{d}x_1]$, drugog da bude u intervalu $[x_2, x_2 + \mathrm{d}x_2]$, i tako do n-tog dana je s izrazom:

$$p(x) = \prod_{i=1}^{n} p(x_i \in [x_i, x_i + \mathrm{d}x_i]) = \prod_{i=1}^{n} f(x_i;\theta)\mathrm{d}x_i. \tag{5}$$

Ako su oblik pretpostavljene gustoće vjerojatnosti i vrijednosti parametra točne, može se očekivati da će vjerojatnost za promatrane podatke biti znatno viša nego ako je vrijednost parametra daleko od stvarne. Kako $\mathrm{d}x_i$ ne ovise o parametrima, možemo definirati tzv. funkciju izglednosti (engl. likelihood function, likelihood) koja ne ovisi o $\mathrm{d}x_i$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{6}$$

Možemo primijetiti da je to samo združena funkcija gustoće vjerojatnosti slučajnih varijabli $x_1,\ldots,x_n$, iako se tretira kao funkcija parametra θ. Pritom se vrijednosti x_i tretiraju kao fiksirane jer je eksperiment završio [16].

Kako bismo pronašli set parametara koji daje maksimalnu vrijednost funkcije izglednosti, dovoljno je derivirati funkciju po svim parametrima:

$$\frac{\partial L}{\partial \theta_i} = 0, i = 1,\ldots,m. \tag{7}$$

U praksi, često baratamo s vrlo niskim vrijednostima funkcije izglednosti ($< 10^{-100}$), pa nam račun postaje numerički nestabilan. Kako bismo doskočili tom problemu, u računima se koristi logaritam funkcije izglednosti $\ln L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$.

Kada procijenimo željeni parametar θ metodom najveće izglednosti, trebamo odrediti i interval pouzdanosti koji odgovara pokrivenošću od 68.27% ("1σ") oko procijenjene vrijednosti. Pritom koristit ćemo aproksimaciju velike statistike u kojoj je nesigurnost izglednosti dana kao razliku vrijednosti za koju logaritam izglednosti za jednu polovinu od logaritma najveće izglednosti i vrijednosti za koju imamo najveću izglednost:

$$lnL_{max} - lnL_{\sigma} = \frac{1}{2}. \tag{8}$$

Koristimo li neki od numeričkih alata (npr. MINUIT), on će, tražeći maksimalnu izglednost, odmah odrediti i grešku, pa ćemo te rezultate usporediti s našom aproksimacijom.

2.2 Histogram

Ukoliko imamo konačan set podataka, možemo ga grafički prikazati na histogramu. Na histogramu, x-os je podijeljena u m podintervala ili razreda (engl. bins) širine Δx_i, $i = 1,\ldots,m$, gdje su Δx_i obično, ali ne nužno, isti za svaki razred. Pritom je broj događaja n_i u razredu i prikazan na y-osi, a površina ispod histograma A jednaka je $A = \sum_{i=1}^{m} n_i \cdot \Delta x_i$ [16].

Podatke raspodjeljujemo po razredima jer ih je tako jednostavnije gledati, ali pritom gubimo neku informaciju. Kako je teorija koja stoji iza fizike elementarnih čestica u srži kvantna, ona nam ne daje potpuno determinirane opservable, nego vjerojatnost njihovog nalaženja. Zbog toga ne bi imalo smisla promatrati događaj po događaj, nego trebamo gledati njihovu učestalost ili raspodjelu mjerenih vrijednosti. Upravo tako i uspoređujemo teoriju s eksperimentom: gledajući broj eksperimentalno izmjerenih podataka po razredu i njihove očekivane vrijednosti dobivene iz teorije.

2.3 Generiranje nasumičnih brojeva i Monte Carlo simulacija

Iako je stvarna nasumičnost nerješiv problem u klasičnoj računalnoj znanosti[4], postoje načini da se generiraju nasumični brojevi. Postoje algoritmi koji mogu generirati redove brojeva koji imaju svojstva bliska pravim nasumičnim varijablama. Pritom ti algoritmi, koji generiraju pseudonasumične brojeve, moraju zadovoljavati nekoliko uvjeta.

Dobri generatori pseudonasumičnih brojeva moraju, u limesu velikih brojeva, imati željene statističke karakteristike stvarnih nasumičnih varijabli. Također, moraju moći generirati sekvence brojeva koje su statistički neovisne o prijašnjim pozivima, iako će naravno svaki broj biti matematički determiniran, kroz algoritam generatora, iz prijašnjih brojeva [15].

Tehnika koja koristi sekvence nasumičnih brojeva kako bi simulirala mjerenja zove se Monte Carlo metoda. Monte Carlo metoda može se primijeniti kada god je problem parametriziran s nekom vjerojatnosnom distribucijom. Uzmimo za primjer eksperiment u kojem se elektron raspršuje na nekoj fiksnoj meti i zatim se detektira. Pritom nam teorija daje neka predviđanja o vjerojatnosti da se događaj zbije na određenom kutu raspršenja (drugim riječima, o diferencijalnom udarnom presjeku). Iz te distribucije, možemo primijeniti Monte Carlo metodu kako bismo generirali kuteve raspršenja i pripadajuće impulse. Takvi se programi zovu generatori događaja (engl. event generators), a u fizici elementarnih čestica koriste se kako bi opisali široki raspon čestičnih interakcija [16].

[4] Korištenjem kvantnih sustava, moguće je dobiti stvarnu nasumičnost.

2.4 Predlošci histograma

U eksperimentu, cilj nam je proučiti određene fizikalne procese i karakterizirati ih: izmjeriti njihov udarni presjek (učestalost tog procesa) ili postaviti granice na njega, izmjeriti raspodjele različitih opservabli (diferencijalni udarni presjek), kao što mogu biti kutne raspodjele ili raspodjele impulsa čestica u konačnom stanju, itd. Proces se najčešće proučava izborom jedne ili više opservabli koje nam dopuštaju izdvojiti proučavani proces od drugih. Mjerene vrijednosti izabrane opservable, npr. invarijantne mase, pohranjene su u histogramima koji se koriste za proučavanje raspodjele te opservable. Histogram stvarno mjerenih vrijednosti d_i, gdje i označava i-ti razred histograma, uspoređuje se s histogramom očekivanih vrijednosti iz Monte-Carlo simulacije MC_i, koja dobiva doprinos iz simulacije proučenog procesa u određenom mjerenju, signala s_i, i iz simulacija drugih procesa, pozadine b_i, koji doprinose mjerenju:

$$MC_i = b_i + \mu \cdot s_i. \tag{9}$$

Ovdje smo uveli normalizacijski faktor μ za signal, koji je jednak 1 ako je teorija na kojoj se simulacija temelji točna.

Simulirane raspodjele signala i pozadine, s_i i b_i, u kojima je pohranjen očekivani oblik histograma, zovemo predlošci (engl. template histograms, templates), te imamo jedan posebno za signal, a jedan posebno za pozadinu. U usporedbi histograma podataka, d_i, sa simulacijama MC_i variramo te predloške unutar njihovih nesigurnosti kako bi došli do što boljeg slaganja između podataka i simulacije. Spomenute varijacije mogu uključivati i varijacije normalizacije i varijacije oblika predložaka. U ovom radu ćemo pretpostaviti da je predložak za pozadinu savršeno poznat i razmatrati jedino nesigurnost na normalizaciju za signal, sadržana u varijabli μ, koju ćemo zvati snagu signala. Kao što smo i vidjeli u uvodu i prethodnim potpoglavljima, u centru naše pažnje bit će histogrami te ćemo promatrati postoji li optimalan izbor raspodjele razreda - njihovog broja, širina te granica - koji maksimizira snagu signala.

2.5 Otkriće nove fizike

Otkriće novih fenomena u fizici glavni je cilj eksperimenata. Ako eksperiment producira uvjerljivo odstupanje od dosadašnjih saznanja ili očekivanja na temelju poznatih zakona, može se tvrditi da je otkriveno nešto novo. No čak i u slučaju nedostatka značajnog odstupanja, ono može dati granice na moguću novu fiziku [15]. Dobar primjer postavljanja granica na moguću novu fiziku je mjerenje raspada protona. Naime, standardni model čestica predviđa da je proton stabilna čestica, dok razne

Tablica 2: Signifikantnosti izražene kao $Z\sigma$ i pripadne vjerojatnosti (p-vrijednosti). Preuzeto iz [15].

Z [σ]	p [%]
1.00	15.9
1.28	10.0
1.64	5.00
2.00	2.28
2.32	1.00
3.00	0.14
3.09	0.10
3.71	0.01
4.00	$3.17 \cdot 10^{-3}$
5.00	$2.87 \cdot 10^{-5}$
6.00	$9.87 \cdot 10^{-8}$

teorije fizike izvan standardnog modela (engl. beyond standard model, BSM) predviđaju da on to nije. Mjerenjem raspada protona postavlja se granica na teorije koje predviđaju njegov raspad. Zasad, nije pronađen ni jedan dokaz da je proton nestabilna čestica pa imamo samo donju granicu na njegovo vrijeme poluživota koja iznosi 3.6×10^{29} godina.

Kako pridjev "uvjerljivo" nije egzaktan, trebali bismo ga kvantificirati i reći što nam "uvjerljivo" znači u kontekstu fizike elementarnih čestica. Općenito, gledajući raspodjelu npr. invarijantnih masa, imat ćemo pozadinu u obliku eksponencijalno padajuće krivulje te signal u obliku Gaussove funkcije. Ako imamo neki set podataka, kako bismo mogli tvrditi da smo otkrili signal trebamo pokazati da su nam podatci, odnosno mjerenja, dovoljno nekonzistentni s hipotezom da imamo samo pozadinu, koju po pretpostavci znamo gotovo savršeno. Naravno, uvijek postoje procesi za koje imamo nesigurnosti, kao i nepouzdanosti pri generiranju same pozadine, te to trebamo uključiti u analizu. Kvantitativna mjera nepoklapanja podataka s hipotezom samo pozadine dana je sa signifikantnošću, definiranoj kao vjerojatnosti određene slučajne varijable da imaju vrijednost ekstremniju od one dane podatcima pod hipotezom da je samo pozadina prisutna.[5]

Signifikantnost, iako definirana kao vjerojatnost (p-vrijednost), najčešće se ne navodi kao broj nego kao broj standardnih devijacija koji odgovaraju p-vrijednosti na desnom repu normalne dis-

[5]Obično signal i pozadina daju očekivanu vrijednost od same pozadine, no moguće je i da signal smanjuje očekivanu vrijednost pozadine.

tribucije. Dakle, signifikantnost je moguće iskazati kao "$Z\sigma$", koju dobivamo transformacijom iz vjerojatnosti:

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x = 1 - \Phi(Z) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{Z}{\sqrt{2}}\right)\right] \tag{10}$$

U fizici elementarnih čestica, prag za otkriće novog fenomena je signifikantnost od 5σ što odgovara vjerojatnosti od $p = 2.87 \cdot 10^{-7}$ [15].

3 Rezultati i diskusija

U daljnjoj ćemo diskusiji pretpostaviti, motivirani histogramom na slici 6, da nam je pozadina u obliku padajuće eksponencijale, a signal u obliku gausijana. Pritom ćemo, donekle proizvoljno, ali opet s motivacijom iz histograma na slici 6, odrediti da nam se srednja vrijednost gausijana (ili masa) nalazi na 120 GeV, a da je standardna devijacija 20 GeV koju ćemo zvati širina signala i označavati sa σ_s. Te parametre nećemo mijenjati u daljnjoj diskusiji. Ono što ćemo mijenjati su luminozitet te omjer signala i pozadine. Dakle, skalirat ćemo i pozadinu i signal te njihov omjer. Kasnije ćemo koristiti i očekivanu pozadinu koja ima velike statističke fluktuacije, a naposljetku ćemo i primijeniti naučeno na histogram na slici 6.

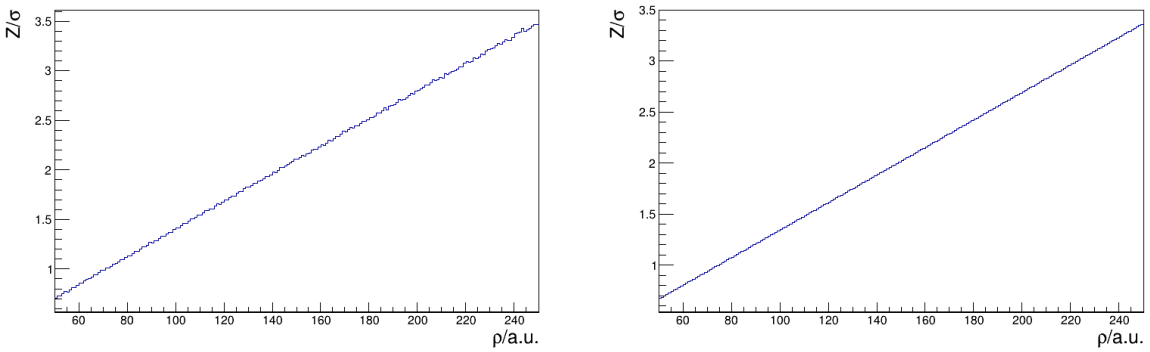
3.1 Statističke fluktuacije pozadine

Prvo što smo promatrali su same statističke fluktuacije pozadine tako što smo generirali pseudonasumične uzorke s očekivanim vrijednostima pozadine i brojali koliko puta nam je integral ispod krivulje veći od integrala ispod krivulje za očekivane vrijednosti pozadine i signala. Pritom smo promatrali ovisnost signifikantnosti o očekivanoj vrijednosti signala, intervalu integracije i luminozitetu. Za svaku vrijednost generirano je 10^7 pseudonasumičnih uzoraka, a signifikantnost je izražena u obliku $Z\sigma$ kako je i navedeno u prethodnom poglavlju.

Rezultate ćemo usporediti s aproksimacijom da je signifikantnost zapravo omjer količine signala i korijena količine pozadine u jedinicama $Z\sigma$. U pozadini te aproksimacije stoji činjenica da je standardna devijacija Poissonove raspodjele zapravo korijen iz očekivane vrijednosti pozadine. Kako Poissonovu raspodjelu možemo aproksimirati Gaussovom raspodjelom za visoke očekivane vrijednosti, to odstupanje možemo izravno povezati sa standardnom devijacijom σ. Dakle, ono što mjerimo u ovoj aproksimaciji je koliko nam signal odstupa od očekivane vrijednosti u jedinicama $Z\sigma$ s vrlo jednostavnom formulom $Z[\sigma] = \frac{S}{\sqrt{B}}$, koja će biti bolja što je očekivana vrijednost pozadine veća.

3.1.1 Ovisnost signifikantnosti o očekivanoj vrijednosti signala

Ovisnost signifikantnosti o očekivanoj vrijednosti signala prikazana je na slici 7. Na lijevoj slici prikazana je signifikantnost dobivena Monte Carlo metodom, dok je na desnoj slici prikazana signifikantnost dobivena aproksimacijom $\frac{S}{\sqrt{B}}$. Očekivana vrijednost signala dana je u proizvoljnim jedinicama (engl. arbitrary units, a.u.), a signifikantnost u jedinicama $Z\sigma$. Možemo primijetiti da se te dvije slike jako dobro poklapaju (do na statističke fluktuacije), te time potvrđujemo da je i



Slika 7: Ovisnost signifikantnosti o očekivanoj vrijednosti signala. (Lijevo) Signifikantnost dobivena Monte Carlo simulacijama za pojedinu snagu signala. (Desno) Signifikantnost dobivena aproksimacijom za pojedinu snagu signala.

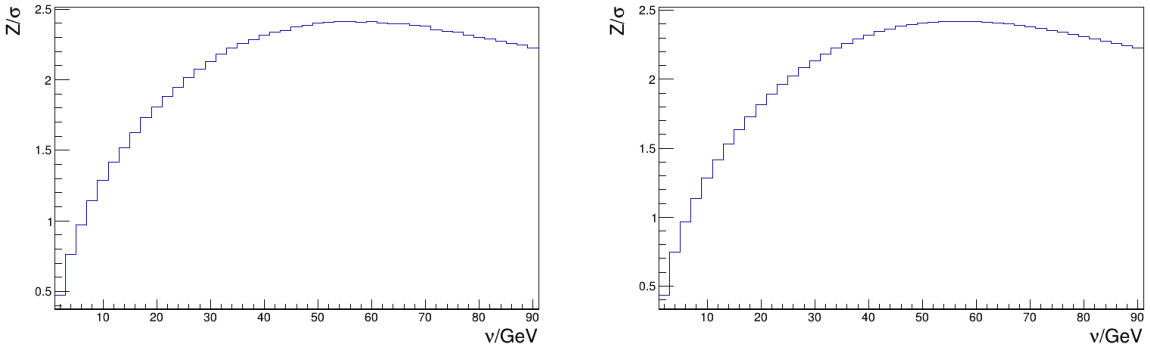
aproksimacija dobra. Također, možemo primijetiti i da nam signifikantnost raste približno linearno s očekivanom vrijednosti signala u jedinicama $Z\sigma$. Dakle, što je signal jači, to ćemo biti sigurniji da je taj signal zaista prisutan, a ne samo statistička fluktuacija pozadine, što je i očekivano.

3.1.2 Ovisnost signifikantnosti o intervalu integracije

Za razliku od ovisnosti signifikantnosti o očekivanoj vrijednosti signala, za ovisnost o intervalu integracije nemamo monotono ponašanje kao što se može vidjeti sa slike 8. Kao i prethodnom slučaju, rezultati aproksimacija se jako dobro slažu s rezultatima dobivenim Monte Carlo metodom. U ovome slučaju imamo jednu točku za koju imamo maksimum signifikantnosti, a to je ako je širina intervala 54 GeV. Dakle, optimalni interval je malo širi od standardne devijacije našeg signala. To je i očekivano, tom širinom intervala obuhvaćamo veliku većinu doprinosa signala. Kada su vrijednosti niže od toga, ne obuhvaćamo cijeli signal, a kada su vrijednosti više, signal više ne doprinosi, a fluktuacije pozadine rastu.

3.1.3 Ovisnost signifikantnosti o luminozitetu

Slično kao i u slučaju ovisnosti o očekivanoj vrijednosti signala, i za ovisnost o luminozitetu imamo monotono ponašanje što se vidi na slici 9. Kao i u prethodnim slučajevima, rezultati aproksimacija se slažu s rezultatima dobivenim Monte Carlo metodom. Ovisnost signifikantnosti raste s luminozitetom i to s korijenom luminoziteta, što je i očekivano iz aproksimacije $\frac{S}{\sqrt{B}}$. Naime, kako luminozitet skalira i pozadinu i signal, on ne mijenja njihov omjer. Kasnije ćemo i eksplicitno provjeriti potenciju ovisnosti o luminozitetu za histogram na slici 6. Pritom statističke fluktuacije pozadine, za

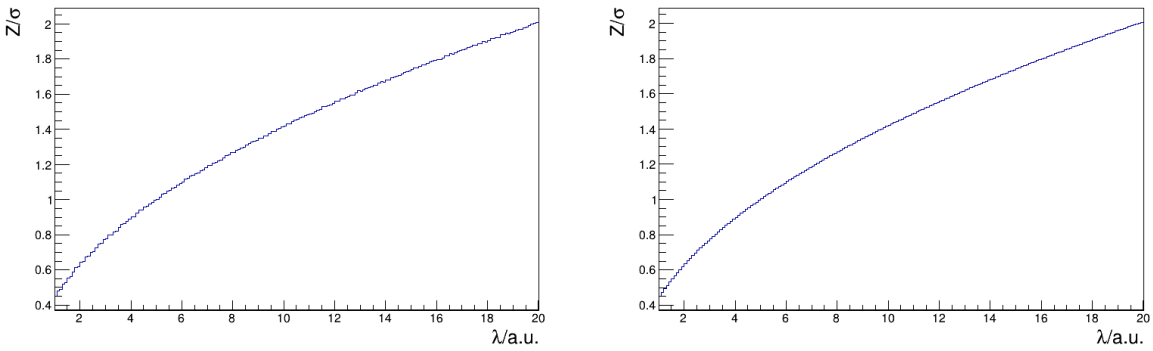


Slika 8: Ovisnost signifikantnosti o intervalu integracije. (Lijevo) Signifikantnost dobivena Monte Carlo simulacijama za pojedini interval integracije. (Desno) Signifikantnost dobivena aproksimacijom za pojedini interval integracije.

veće očekivane vrijednosti, postaju apsolutno veće, ali relativno manje (u odnosu na iznos očekivane vrijednosti) te zbog toga raste signifikantnost.

3.2 Utjecaj širine razreda

Nakon proučavanja pozadinski fluktuacija možemo si postaviti pitanje koje je i tema ovog rada: postoji li optimalna širina razreda za koju imamo najmanju nesigurnost snage signala? Kako bismo proučili kako spajanje razreda utječe na signifikantnost signala, prvo generiramo pseudonasumični uzorak (ili, skraćeno, pseudouzorak) u obliku koji smo naveli početkom poglavlja, te zatim radimo prilagodbu tim podatcima. Prilikom prilagodbe te krivulje na podatke, koristimo funkciju izglednosti te tražimo njen maksimum i interval nepouzdanosti. Dakle, radimo metodu najveće izglednosti. Za-



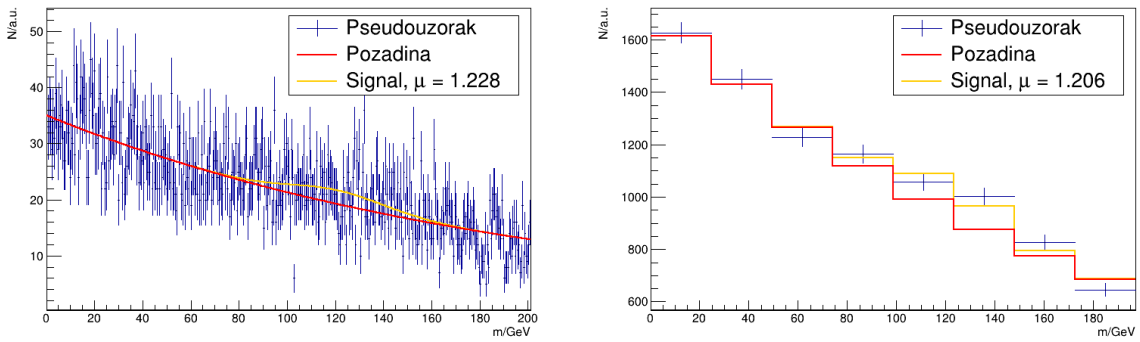
Slika 9: Ovisnost signifikantnosti o luminozitetu. (Lijevo) Signifikantnost dobivena Monte Carlo simulacijama za pojedini iznos luminoziteta. (Desno) Signifikantnost dobivena aproksimacijom za pojedini iznos luminoziteta.

tim, tom istom uzorku spajamo određeni broj razreda u jedan, te ponavljamo taj proces dok širina spojenih razreda ne bude znatno veća od širine signala. Primjer pseudouzorka originalne raspodjele i pseudouzorka nastalog spajanjem susjednih razreda originalne raspodjele, zajedno s pripadnim prilagođenim krivuljama, prikazan je na slici 10.

Za histograme u kojima proučavamo ovisnost o širini razreda, rezultate prikazujemo tako da na horizontalnoj osi imamo vrijednost omjera širine razreda i širine signala, a na vertikalnoj osi vrijednost na kojoj imamo najveće izglednost i pripadnu nepouzdanost. Pritom možemo promatrati kako nam rezultati ovise o početnom luminozitetu i početnoj očekivanoj vrijednosti signala.

3.2.1 Variranje luminoziteta

Prvo što ćemo gledati je utjecaj širine razreda kada je fiksirana početna očekivana vrijednost signala, a luminozitet između pseudouzoraka mijenjamo. Na slici 11 vidimo rezultate takvog postupka. Između pojedinih luminoziteta postoji razlika u izmjerenim snagama signala, no ona je posljedica statističkih fluktuacija prilikom generiranja pseudouzorka, dok su sami rezultati unutar greške konzistentne s jedinicom, pa na to ne trebamo previše obraćati pažnju. Naime, u području gdje očekujemo signal, Monte Carlo metoda u jednom pseudouzorku može generirati brojeve koji su u prosjeku niži ili viši od ukupnih očekivanih vrijednosti. Stvarna snaga signala μ koju ovdje očekujemo, za sve grafove, iznosi 1, ali, kao što smo rekli, izmjerena može biti i viša i niža. Važnije je promotriti kako se izmjerena snaga signala mijenja prilikom spajanja određenog broja razreda. S lijeve strane slike 11, možemo vidjeti da nam srednja vrijednost u svim slučajevima ostaje otprilike konstantna bar do oko 1.3 σ_s, s tim da u slučaju većeg luminoziteta ostaje otprilike konstanta i do 1.7 σ_s. Snaga signala pritom manje oscilira za veće luminozitete, što smo mogli i očekivati jer spajanje razreda na manjim



Slika 10: Primjeri pseudouzorka i prilagodbe signala na podatke za: (Lijevo) početnu raspodjelu gdje je širina razreda 0.5 GeV i (Desno) raspodjelu nastalu spajanjem 40 susjednih razreda.

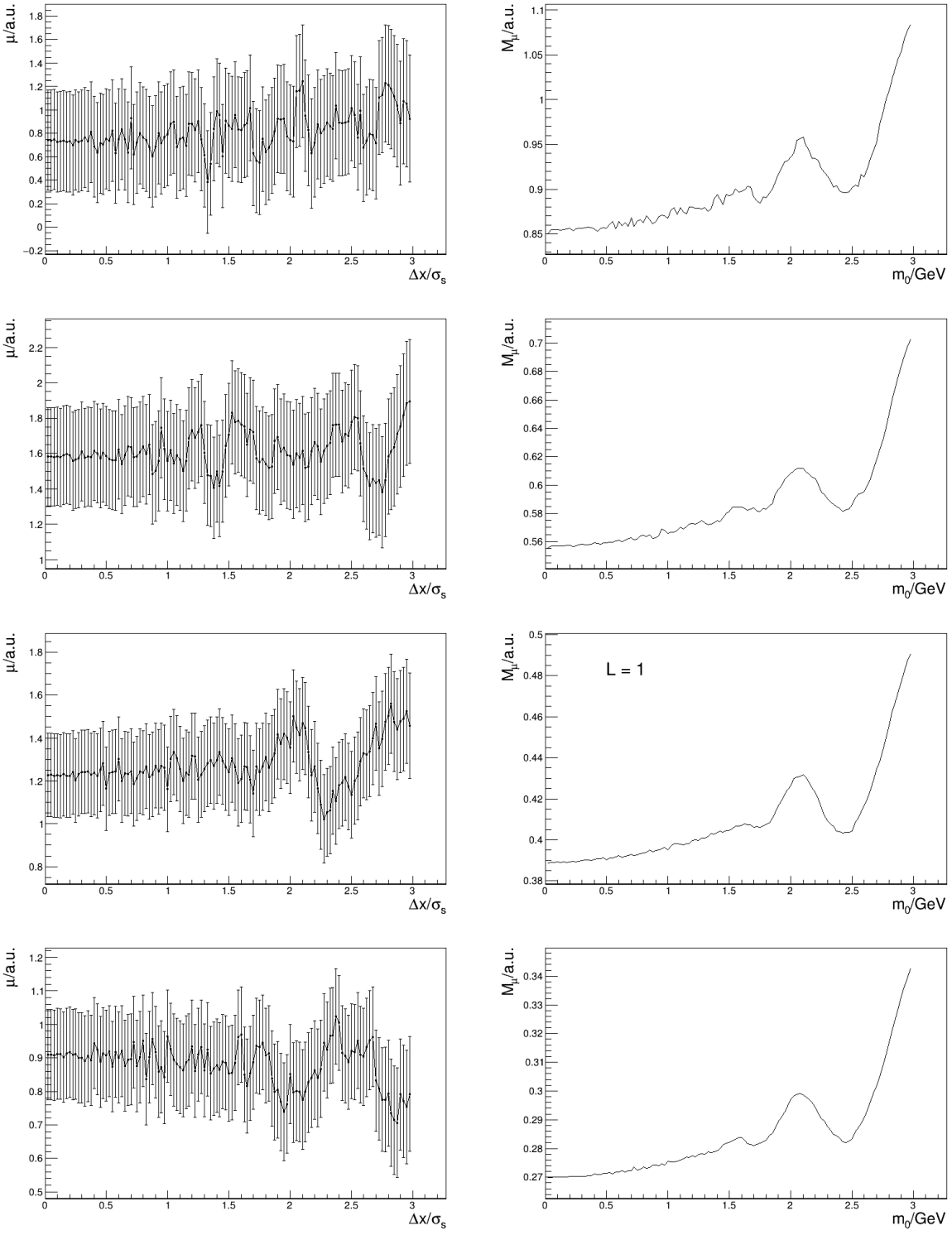
luminozitetima ima veći utjecaj na statističke fluktuacije nego na većim luminozitetima.

Možemo primijetiti da nam grafovi za greške, s desne strane slike 11, imaju isti oblik bez obzira na luminozitet, dok se najviše mijenja početna vrijednost. Pritom najmanju grešku uvijek imamo za početni histogram, te je ona gotovo konstantna do 0.5 σ_s, a onda raste sve do oko 1.6 σ_s, kada počinju oscilacije u vrijednostima. Općenito, što je veći luminozitet, to je greška manja, što smo i očekivali. Taj se rezultat slaže sa zaključkom da signifikantnost raste s luminozitetom za istu snagu signala, koji smo dobili u prethodnom potpoglavlju. Važan je i rezultat da utjecaj o širini razreda ne ovisi o luminozitetu. Dakle, odabir širine razreda ne treba mijenjati za promjene u luminozitetu.

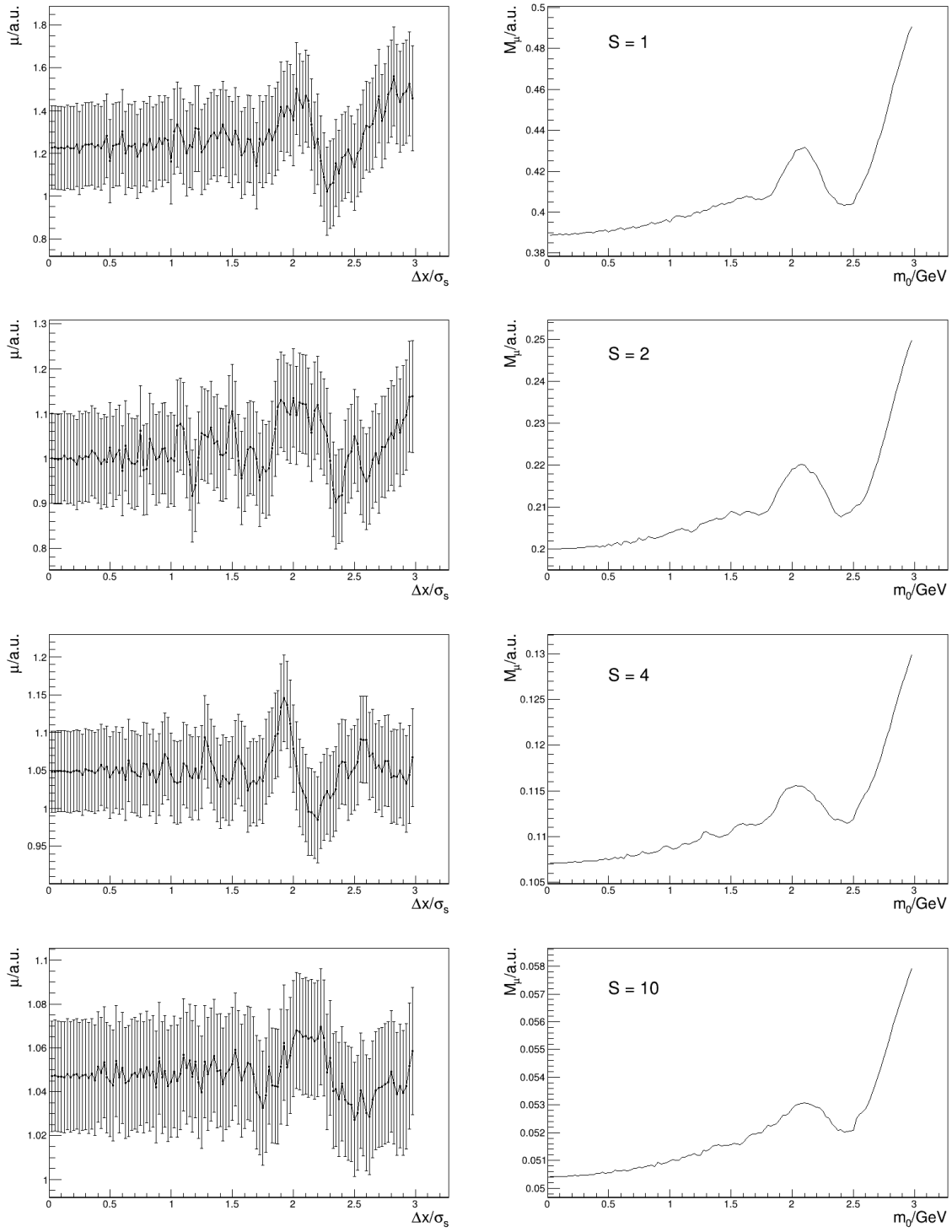
3.2.2 Variranje očekivane vrijednosti signala

Sljedeće što gledamo je utjecaj spajanja razreda kada je fiksiran početni luminozitet, a očekivanu vrijednost signala između pseudouzoraka mijenjamo. Na slici 12 vidimo rezultate takvog postupka. Pritom nam je stvarna vrijednost za snagu signala jednaka 1 u svakom slučaju jer uzimamo u obzir snagu signala s kojom je pseudouzorak generiran prilikom procjene. To nam pojednostavljuje račun jer izbjegavamo problem sa skaliranjem grešaka. Kao i u slučaju fiksirane početne očekivane vrijednosti signala, između pojedinih snaga signala postoji razlika u srednjim vrijednostima koja je posljedica generiranja pseudonasumične raspodjele. S lijeve strane slike 12, možemo vidjeti da nam srednja vrijednost u svim slučajevima ostaje otprilike konstantna sve do oko 1.7, izuzevši slučaj za 2 puta veću snagu signala od proizvoljne početne. Za razliku od prošlog slučaja, ovdje nam se "stabilnost" srednje vrijednosti pri spajanju razreda ne mijenja značajno jer je očekivana vrijednost signala znatno manja od očekivane vrijednosti pozadine, pa samim time i mijenjanje snage signala nema značajnog utjecaja na ukupnu očekivanu vrijednost.

Kao i u slučaju u kojem smo mijenjali početni luminozitet, grafovi za greške, prikazani na desnoj strani slike 12, imaju isti oblik, a početna se vrijednost mijenja. Opet, najmanju grešku uvijek imamo za početni graf, te ona raste sve do oko 1.7 σ_s, kada počinju oscilacije u vrijednostima. Općenito, što je signal jači, to je greška manja, što smo i očekivali. Taj se rezultat slaže s rezultatom koji smo dobili u prethodnom potpoglavlju, da nam je signifikantnost veća ako je signal snažniji. Kao i u slučaju s variranjem luminoziteta, utjecaj širine razreda ne ovisi o očekivanoj vrijednosti signala - odabir širine razreda ne treba prilagođavati jačini signala.



Slika 11: (Lijevo) Grafovi ovisnosti najveće izglednosti o širini razreda za luminozitete (od gore prema dolje) 0.2, 0.5, 1 i 2 proizvoljnog početnog luminoziteta. Na horizontalnoj osi prikazana je vrijednost omjera širine razreda naspram širini signala. (Desno) Odgovarajući grafovi ovisnosti grešaka najveće izglednosti o širini razreda.



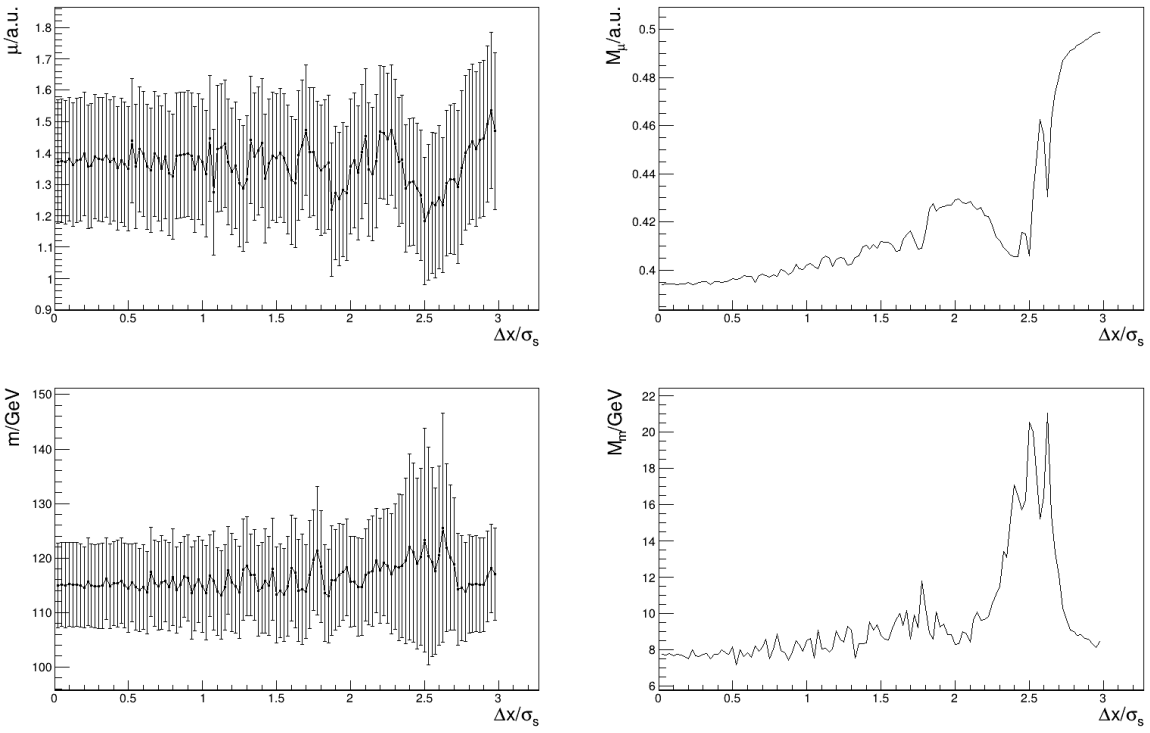
Slika 12: (Lijevo) Grafovi ovisnosti procjene snage signala i o širini razreda za očekivane vrijednosti signala (od gore prema dolje) 1, 2, 4 i 10 puta veće naspram proizvoljne početne očekivane vrijednosti signala za fiksirani luminozitet. Na horizontalnoj osi prikazana je vrijednost omjera širine razreda naspram širini signala. (Desno) Odgovarajući grafovi ovisnosti grešaka procjene snage signala o širini razreda.

3.2.3 Prilagodba mase na signal

Nakon što smo proučili kako nam spajanje razreda djeluje na procjenu snage signala u ovisnosti o luminozitetu i snazi signala, možemo uvesti i procjenu za masu, tj. srednju vrijednost signala. Stvarne vrijednosti parametara za snagu signala i masu iznose 1 i 120 GeV, respektivno, kao i što smo naveli početkom poglavlja. Na slici 13 prikazane su ovisnosti snage signala i mase te njihovih grešaka o širini razreda u jedinicama širine signala.

Na grafovima za snagu signala i njegovu grešku možemo primijetiti slično ponašanje kao i za slučajeve kada nismo prilagođavali masu na podatke. Snaga signala nam ostaje konstantna ili blago oscilira do širine razreda koja je oko 1.7 σ_s, nakon čega počinje znatno oscilirati. Greška na snagu signala lagano raste do 1.7 σ_s, kada počinje prvo oscilirati pa onda značajno rasti.

Za masu vidimo da ostaje konstantna ili blago oscilirati sve do širine razreda od oko 2 σ_s, te nakon toga počinje značajnije oscilirati. Pritom greške mase blago osciliraju, kao i sama masa, do širine razreda od oko 2 σ_s, nakon čega počinju značajnije rasti do otprilike 2.6 širina signala, kada počinju padati. Razlog zašto na tim vrijednostima greška značajno raste je u činjenici da se tada



Slika 13: (Lijevo) Graf ovisnosti procjene snage signala o širini razreda za snagu signala (gore) i masu (dolje) s pripadnim greškama. Na horizontalnoj osi prikazana je vrijednost omjera širine razreda naspram širini signala. (Desno) Odgovarajući grafovi ovisnosti grešaka snage signala o širini razreda.

praktički cijeli signal nalazi u jednom razredu koji je centriran blizu središta signala te pomicanjem lijevo i desno ne vidimo bitnu razliku u izglednosti. Kako se taj razred sve više izmiče iz centra signala, drugi razredi u histogramu počinju značajnije doprinositi izglednosti pa se i greška smanjuje. Ovaj efekt ćemo proučiti nešto kasnije.

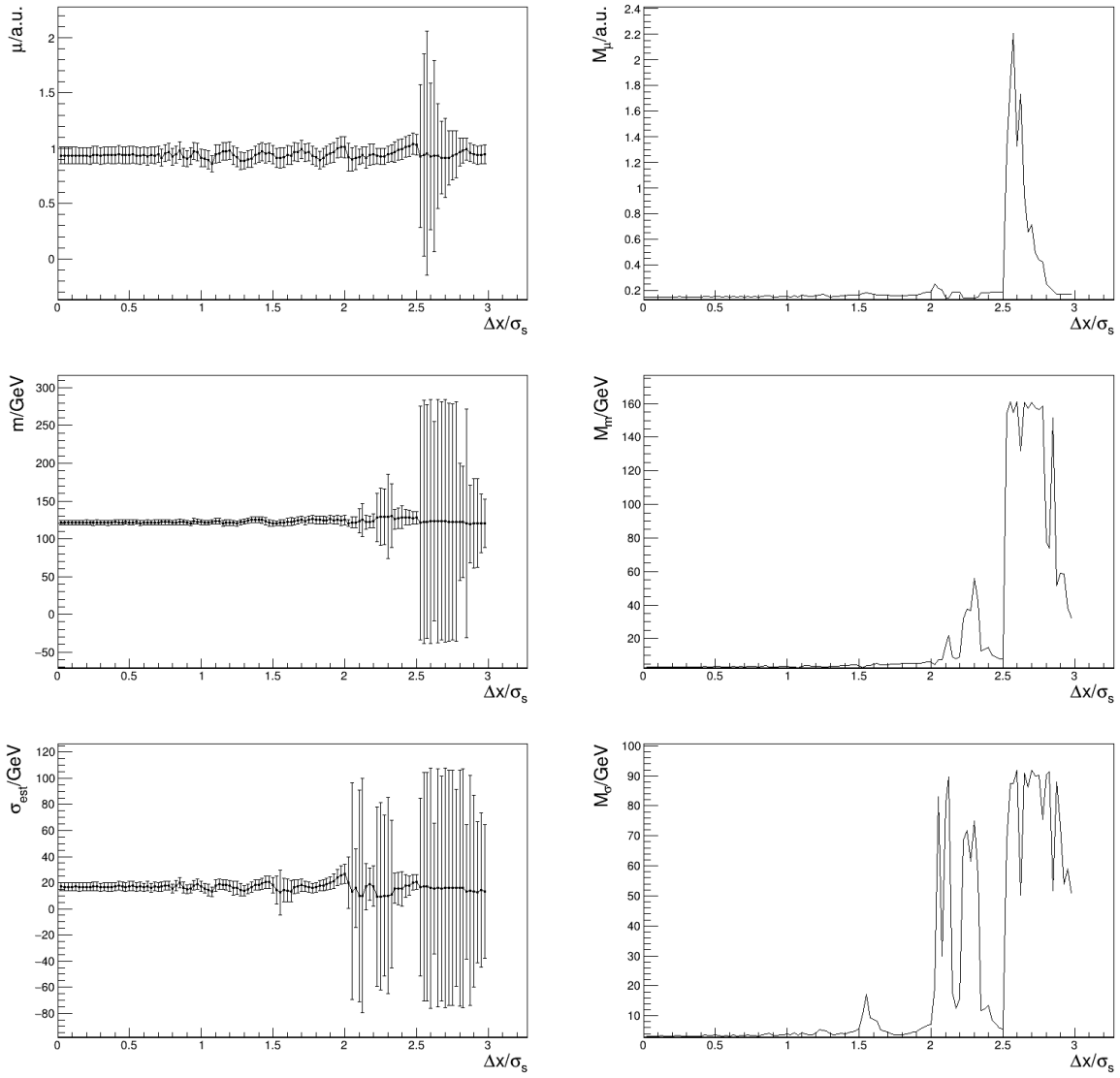
3.2.4 Prilagodba širine signala

Uz procjenu snage signala i mase, možemo promotriti i kako se procjena širine signala mijenja u ovisnosti o širini razreda u kombinaciji s drugim procjenama. Uz prethodno navedene snage signala od 1 i mase od 120 GeV, očekivana širina signala σ iznosi 20 GeV. Na slici 14 prikazani su grafovi izmjerenih vrijednosti snage signala, mase i širine signala u ovisnosti o širini razreda s lijeve strane i pripadni grafovi grešaka s desne. Svi grafovi za izmjerene vrijednosti su prilično stabilni do 2 σ_s, a grafovi za snagu i širinu signala i do 2.5 σ_s. Slična je situacija i s grafovima grešaka.

Ono što se događa na 2.5 σ_s je da širina razreda počinje biti znatno šira od samoga signala, s tim da razredi su centrirani oko mase od 120 GeV, pa cijeli signal bude u samo jednom razredu. Time nam ni pomicanje mase ni širenje signala nema velikog utjecaja na izglednost te greške zbog toga postaju vrlo velike. Posljedično, velike greške za masu i širenje signala, uzrokuje i velike greške na procjenu snage signala. Dakle, ako želimo precizno mjeriti i snagu signala i masu, ali i širinu signala, najbolje bi bilo uzeti da je širina razreda između 0.05 i 1 σ_s.

3.2.5 Utjecaj pozicije razreda

Kako je širina razreda u histogramu veća, tako nam je sve veći utjecaj pozicije razreda u području oko centra signala. Naime, izglednost nam neće isto izgledati raspodjeli li se signal na dva razreda podjednako ili većina signala ostane u samo jednom. Kako bi pogledali što se zbiva s procjenom snage signala i pripadnoj greški ako razrede pomičemo oko centra signala, generiramo raspodjelu s istom pozadinom i istim signalom, ali ne u području dosadašnjih 0 do 200 GeV, nego u području od 0 do 260 GeV. Zatim, uzimamo histogram koji ima raspon od 0 do 200 GeV, spajamo razrede tako da je širina pojedinog 20 GeV te gledamo procjenu snage signala i pripadnu grešku za dani histogram te taj postupak ponavljamo tako da uzmemo novi histogram, koji također ima raspon od 200 GeV, ali početak na energiji za 0.5 GeV većoj od prethodnog sve dok ne dođemo da nam je gornja granica 260 GeV. Time ne testiramo samo ponašanje pozicije razreda oko centra signala, nego i utjecaj rubova na izglednost signala za koju očekujemo da neće doprinositi jer su rubovi daleko od samoga signala. Očekujemo periodičko ponašanje za procjenu snage signala i greške jer će razredi oko centra signala

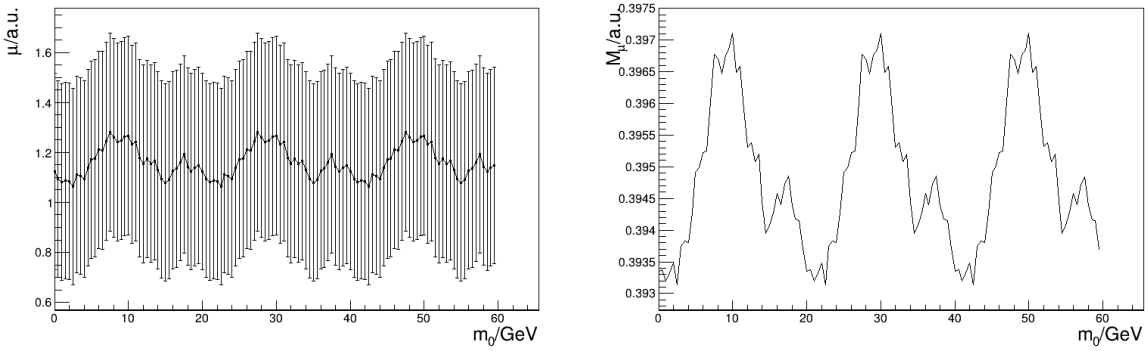


Slika 14: (Lijevo) Graf ovisnosti najveće izglednosti o širini razreda za snagu signala (gore), masu (u sredini) i širinu signala (dolje). Na horizontalnoj osi prikazana je vrijednost omjera širine razreda naspram širini signala. (Desno) Odgovarajući grafovi ovisnosti grešaka na snagu signala za vrijednost najveće izglednosti o širini razreda.

biti na isti pozicijama više puta tijekom ovog postupka.

Rezultati opisanog postupka prikazani su na slici 15. Kao što smo i očekivali, procijenjene greške snage signala, a i njegove greške, imaju periodičko ponašanje s periodom koji je točno jednak širini svakog razreda. Drugim riječima, rezultati su potpuno u skladu s očekivanjima. Pritom vidimo da utjecaj rubova, čak i ako postoji, nije vidljiv na prikazanom grafu, što smo također očekivali.

Na slici 16 prikazani su primjeri pseudouzorka i prilagodbe predloška signala na uzorak za minimum i maksimum greške prikazane na slici 15. Na lijevom grafu prikazani su pseudouzorak i

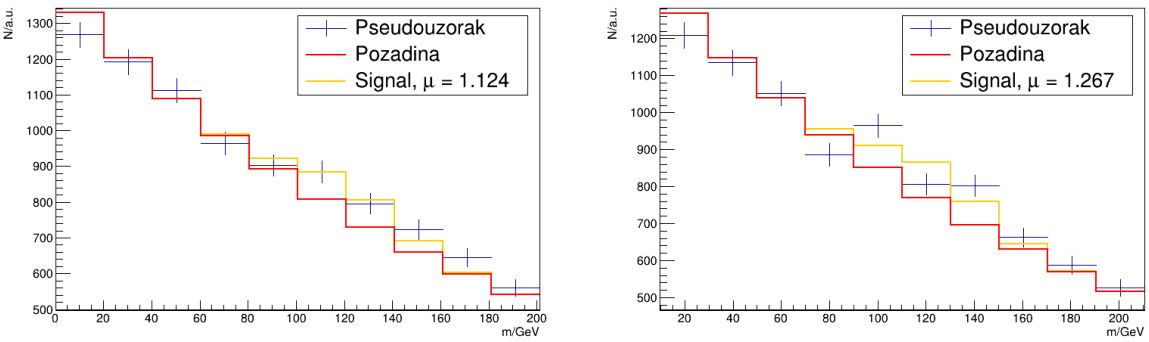


Slika 15: (Lijevo) Grafovi ovisnosti snage signala o širini razreda za snage signala za graf prikazan na slici 6. (Desno) Odgovarajući grafovi ovisnosti grešaka snage signala o širini razreda.

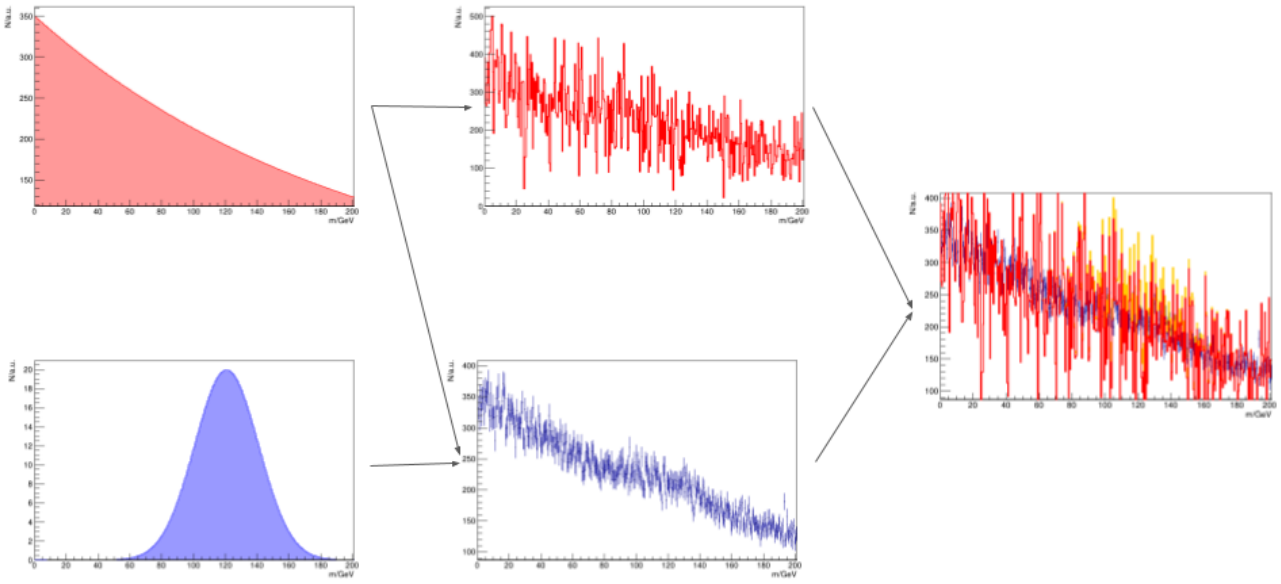
prilagodba koja odgovaraju minimumu greške, a na desnom prikazana su one koje odgovaraju maksimumu. Kao što vidimo, prilagodba je i vizualno bolja na lijevom grafu te možemo zaključiti da je procjena sigurnija ako je gotovo cijeli signal raspoređen podjednako u dva razreda nego dominantno u jednom. Taj zaključak slaže se s našim očekivanjima.

3.3 Utjecaj predložaka za pozadinu sa statističkim fluktuacijama

Kako našu pozadinu, kao i signal, prije svega simuliramo, njena krivulja neće biti savršeno glatka, nego će imati statističke fluktuacije od razreda do razreda. Naime, na raspolaganju imamo konačan broj simulacija. Ako promatramo događaje koji su na rubu faznog prostora, samo mali dio simulacija će upasti u njega pa nećemo dobiti punu sliku onoga što se u tom prostoru može dogoditi. Stoga, kako bi upotpunili naše istraživanje, trebamo uzeti i taj utjecaj u obzir.



Slika 16: Primjeri pseudouzorka i prilagodbe signala na podatke u slučajevima kada imamo: (Lijevo) signal ravnomjerno raspoređen u više razreda i (Desno) signal dominantno raspoređen u jednom razredu kojemu je centar jednak centru signala.



Slika 17: Skica procedure kojom generiramo predloške sa statističkim fluktuacijama. (Gore lijevo) Originalni predložak pozadine. (Dolje lijevo) Predložak signala. (Gore u sredini) Predložak pozadine sa statističkim fluktuacijama generiran iz originalnog predloška pozadine. (Dolje u sredini) Pseudouzorak generiran iz originalnih predložaka pozadine i signala. (Desno) Pseudouzorak i procjena snage signala pomoću predloška signala i pozadine sa statističkim fluktuacijama.

Općenita skica generiranja pseudonasumičnog uzorka i predloška pozadine sa statističkim fluktuacijama, kao i histograma na kojemu smo prilagodili signal na podatke, prikazani su shematski na slici 17. Krećemo od dva predloška na lijevoj strani, od kojih gornji prikazuje signal, a donji pozadinu. Pomoću oba predloška generiramo pseudonasumični uzorak, prikazan na gornjem histogramu u sredini. Na donjem histogramu u sredini prikazan je predložak pozadine sa statističkim fluktuacijama, generiran predloškom (savršeno glatke) pozadine pomoću algoritama koje ćemo u nastavku navesti. Naposljetku, radimo procjenu snage signala što je prikazano na desnom histogramu.

Kako bi simulirali takvu pozadinu, trebamo algoritam koji će generirati pozadinu sa željenim relativnim odstupanjima. Prvi algoritam s kojim smo generirali pozadinu je radio na sljedeći način: nakon što smo odredili kolika relativna odstupanja α pozadina treba imati, algoritam je generirao nasumični uzorak po Poissonovoj raspodjeli na očekivanoj vrijednosti λ_{gen} gdje su relativna odstupanja točno ona koja smo mi odredili:

$$\frac{1}{\sqrt{\lambda_{gen,1}}} = \alpha \tag{11}$$

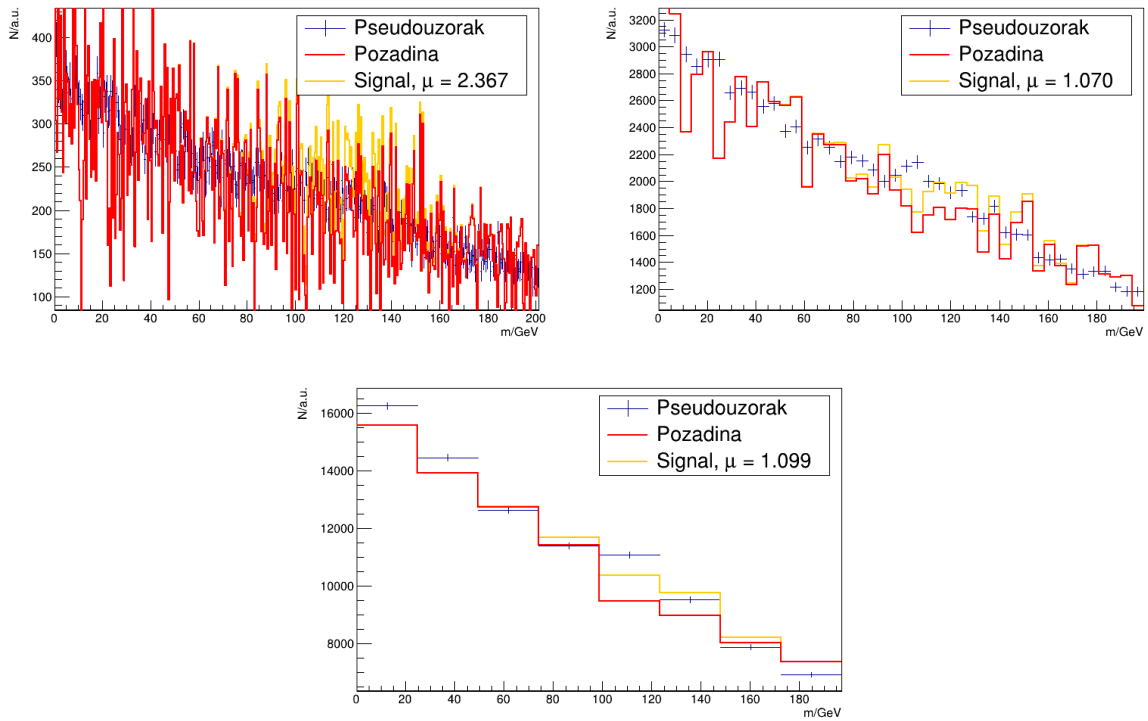
Zatim, nakon generiranja uzorka, algoritam je vrijednosti skalirao s omjerom očekivane vrijednosti pozadine $\lambda_{original}$ naspram očekivane vrijednosti Poissonove raspodjele na kojoj je generirao podatke

za svaki pojedini razred u histogramu:

$$n_i = \frac{\lambda_{original}}{\lambda_{gen,1}} n_{i,gen} \tag{12}$$

gdje je $n_{i,gen}$ generirana vrijednost u i-tom razredu, a n_i vrijednost u i-tom razredu dobivena skaliranjem. Tako dobivamo željenu pozadinu. Mana ovog algoritma je da u slučaju velikih relativnih odstupanja, očekivana vrijednost Poissonove raspodjele postaje niska te time postoji nezanemariva šansa da generira vrijednost nula. Skaliranje nule neće dati ništa drugo nego nulu, pa ako imamo pseudouzorak koja na tom mjestu nema nulu, šansa da se to opiše fluktuacijama pozadine, zbog svojstva Poissonove raspodjele, postaje nemoguća. To vodi do numeričkog problema jer se minimizira logaritam izglednosti, a logaritam nule divergira.

Drugi algoritam koji smo koristili taj problem zaobilazi jer se u njemu ne koristi množenje i dijeljenje, nego zbrajanje i oduzimanje. Ovaj algoritam, za razliku od prvog, ne uzima relativna odstupanja nego apsolutna u obzir. Nakon što odredimo relativna odstupanja koja želimo, algoritam prvo odredi apsolutna odstupanja koja trebaju biti za svaki pojedini razred histograma i generira



Slika 18: Primjer pseudouzorka i prilagodbe signala na podatke u slučaju kada pozadina ima statističke fluktuacije od 30%. (Gore lijevo) Početna raspodjela i pripadna prilagodba gdje je širina razreda 0.5 GeV. (Gore desno) Raspodjela i pripadna prilagodba nakon spajanja 20 susjednih razreda. (Dolje) Raspodjela i pripadna prilagodba nakon spajanja 50 susjednih razreda.

uzorak po Poissonovoj raspodjeli na očekivanoj vrijednosti koja ima upravo ta apsolutna odstupanja:

$$\sqrt{\lambda_{gen,2}} = \alpha\lambda_{original} \tag{13}$$

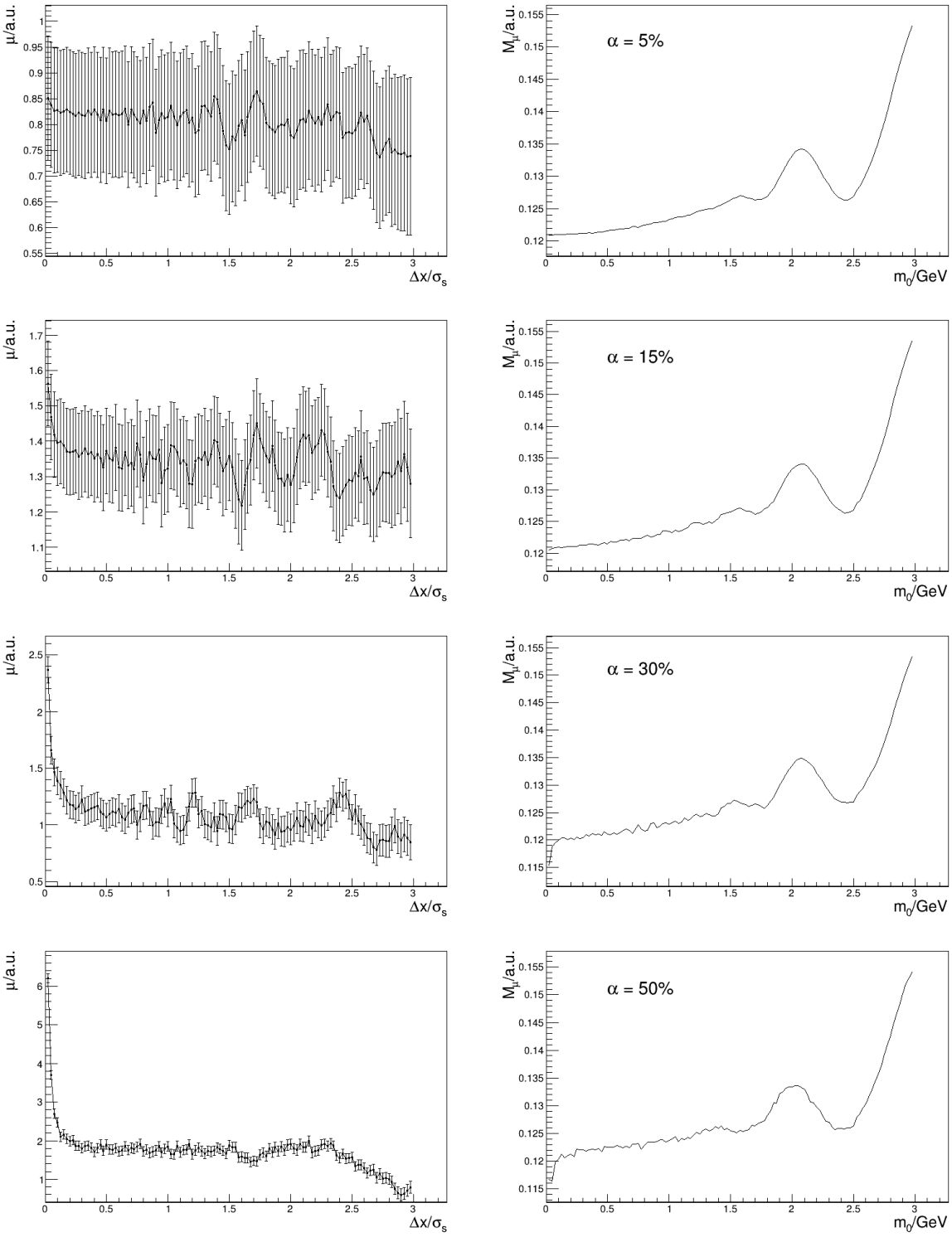
Nakon generiranja uzorka, algoritam te vrijednosti skalira tako da od svake vrijednosti oduzme razliku očekivanih vrijednosti Poissonove raspodjele po kojoj smo generirali i savršeno glatke krivulje za svaki razred:

$$n_i = n_{i,gen} - (\lambda_{gen,2} - \lambda_{original}) \tag{14}$$

Može se primijetiti da prilikom većih odstupanja, očekivana vrijednost Poissonove raspodjele po kojoj je se generira je veća od očekivane vrijednosti savršeno glatke krivulje pa je samim time i distribucija tih fluktuacija simetričnija nego u slučaju prvog algoritma, u kojem su očekivane vrijednosti Poissonove raspodjele, po kojoj su se generirali uzorci, bile male. No, u suštini, oba algoritma daju nam iste rezultate.

Na slici 18 prikazan je primjer generiranog pseudouzorka zajedno s pozadinom koja ima statističke fluktuacije od 30% i prilagođenim signalom te kako se slika mijenja spajanjem razreda. Pritom pseudouzorak generiramo po savršeno glatkoj pozadini, po kojoj generiramo i pozadinu sa statističkim fluktuacijama. Spajanjem razreda histograma, relativne statističke fluktuacije pozadine se smanjuju. To je posljedica Poissonove raspodjele jer zbroj Poissonovih raspodjela je Poissonova raspodjela, a što je veći uzorak manje su pripadne nepouzdanosti. najveće izglednosti.

Sljedeće što trebamo promotriti je kako nam snaga signala ovisi o spajanju razreda u slučaju kada pozadina ima statističku fluktuaciju i to promotriti za nekoliko različitih vrijednosti statističkih fluktuacija pozadine. Na slici 19 prikazane su ovisnosti izmjerene snage signala o širini razreda s lijeve strane i pripadni grafovi grešaka s desne. Kao što vidimo, prvih nekoliko točaka precjenjuju snagu signala, a pritom im je greška na najveću izglednost snage signala manja nego ostalima, dok se ostale točke ponašaju slično kao i u slučaju kada je pozadina savršeno glatka. Taj se efekt pojačava kako se i relativne statističke fluktuacije pozadine pojačavaju. Takav je efekt neočekivan i još ga ne razumijemo u potpunosti. Ono što bismo očekivali je da snage signala za nekoliko prvih točaka bude neprecizno, ali uz znatno veću grešku pa bi imali određenu vrijednost na kojoj bi bio minimum greške. Očito je da, koliko je god procjena tih točaka je izgledna, procjena okolnih točaka je još manje izgledna.



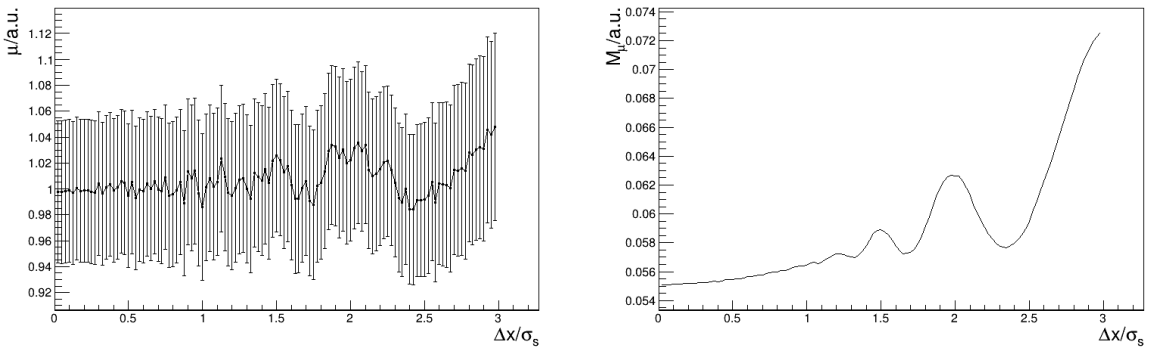
Slika 19: (Lijevo) Grafovi ovisnosti najveće izglednosti širine signala o širini razreda za pozadine s relativnim odstupanjima (od gore prema dolje) 5%, 15%, 30% i 50%. Na horizontalnoj osi prikazana je vrijednost omjera širine razreda naspram širini signala. (Desno) Odgovarajući grafovi ovisnosti grešaka na snagu signala za vrijednost najveće izglednosti o širini razreda.

3.4 Primjena na histogram invarijantne mase za $ZH \rightarrow llbb$

Naposljetku, trebamo primijeniti što smo naučili iz prethodnih potpoglavlja na graf (slika 6) koji nas je i motivirao za ovaj rad. U ovom slučaju, pseudonasumični uzorak nećemo generirati po savršeno glatkoj raspodjeli nego po raspodjeli dobivenoj iz simulacija. No to ne predstavlja preveliki problem, jer sama pozadina nema velike statističke fluktuacije što se vidi sa slike.

3.4.1 Utjecaj širine razreda

Prvo što gledamo je ponašanje snage signala u ovisnosti o širini razreda. Za to generiramo uzorak po podatcima koji dobivamo iz simulacija, kako smo i naveli. Možemo primijetiti da je ponašanje na grafovima na slici 20 vrlo slično kao u idealiziranim slučajevima koje smo promatrali u prethodnim potpoglavljima. Kako su fluktuacije pozadine male, nemamo efekt u kojem prvih nekoliko točaka ima bitno veće vrijednosti od ostalih koje je karakteristično za situacije kada imamo velike statističke fluktuacije. Graf grešaka ima gotovo isti oblik kao i u slučaju kada imamo idealiziranu pozadinu i signal i promatramo ovisnost najveće izglednosti snage signala o širini razreda. Greške na izmjerenu snagu signala su vrlo male, što je posljedica dovoljno visokog luminoziteta i relativno visoke očekivane vrijednosti signala naspram pozadine. Zaključak je ovog razmatranja, da spajanje razreda utječe na signifikantnost signala, pogotovo ako je širina razreda veća od snage signala i da signifikantnost slabi s većom širinom razreda. Optimalno je uzeti što uže razrede, ali nema prevelike razlike između različitih širina dok god uzimamo da je širina razreda između 0.5 i 20 GeV.

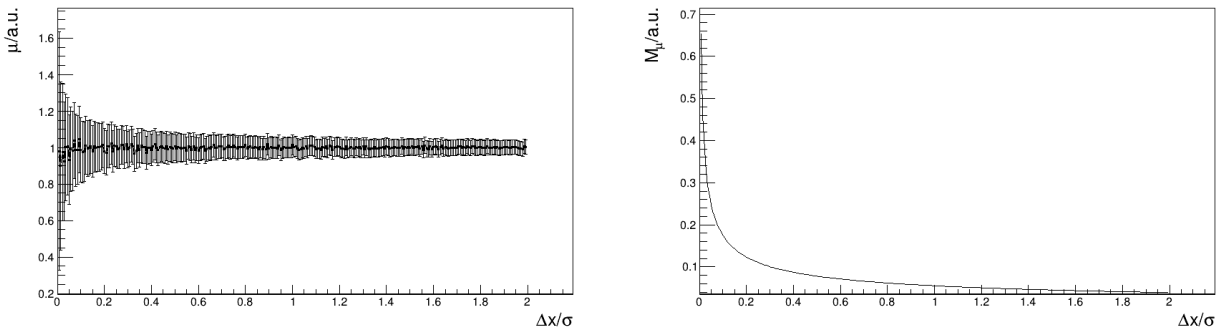


Slika 20: (Lijevo) Grafovi ovisnosti snage signala o širini razreda za snage signala za graf prikazan na slici 6. (Desno) Odgovarajući grafovi ovisnosti grešaka snage signala o širini razreda.

3.4.2 Ovisnost o luminozitetu

Korisno bi bilo promotriti i kako se snaga signala mijenja u ovisnosti o luminozitetu. Za to ćemo uzimati, donekle proizvoljno, raspodjelu u kojoj je širina razreda 8 GeV te prema njoj generiramo pseudouzorke na svakom luminozitetu. Na svakom luminozitetu generiramo 40 uzoraka te uzimamo njihov prosjek kako bi smanjili statističke fluktuacije koje su prisutne, pogotovo na nižim luminozitetima. Greške procjena se pritom ne mijenjaju znatno od uzorka do uzorka.

Na slici 21 prikazani su rezultati tog postupka. Vidimo da procjene ostaju konstantne na vrijednosti 1, što smo i očekivali, dok se greške smanjuju, što smo također očekivali. Greške bi se trebale opadati s korijenom ukupnog broja događaja, pa zbog toga i s korijenom luminoziteta. Kako bi to provjerili, na desni graf sa slike 21 prilagođavamo krivulju oblika $y = ax^{\alpha} + b$ i dobivamo za koeficijent α, koji nas najviše zanima, da iznosi $\alpha = 0.4997 \pm 0.002$ što je vrlo blizu očekivanoj vrijednosti od 0.5. Dakle, kako bismo bili što sigurniji u tvrdnju da je signal zaista prisutan, trebamo imati što veći luminozitet, ili, drugim riječima, trebamo što više proton-proton sudara.



Slika 21: (Lijevo) Graf ovisnosti izmjerene snage signala o luminozitetu s pripadnim greškama. Za izmjerenu snagu signala uzeta je prosječna vrijednost dobivena iz 40 pseudonasumičnih uzoraka. (Desno) Odgovarajući grafovi ovisnosti grešaka izmjerene snage signala o luminozitetu.

4 Zaključak

U ovome smo radu istraživali kako nam izbor razreda u histogramu utječe na snagu i signifikantnost signala. Pozadinu smo uzeli da ima oblik eksponencijalno padajuće funkcije, dok nam je signal bio u obliku gausijana s centrom na 120 GeV i standardnom devijacijom od 20 GeV. Prvo smo gledali kako nam statističke fluktuacije same pozadine mogu dovesti do krivog zaključka o prisutnosti signala tako što smo generirali veliki broj pseudonasumičnih uzoraka i mjerili koliko je puta integral ispod krivulje uzorka veći od integrala ispod krivulje za očekivane vrijednosti pozadine i signala. Mjerili smo nesigurnost u ovisnosti o luminozitetu, očekivanoj vrijednosti signala i intervalu integracije te dobili da nam za ovisnosti o luminozitetu i očekivanoj snazi signala signifikantnost raste, što smo i očekivali, a za ovisnost o intervalu integracije ima optimalnu širinu od oko 2.5 σ_s, što je također u skladu s očekivanjima jer signal doprinosi i izvan područja koju nazivamo njegovom širinom dok god pozadina u tom području nije prevelika.

Sljedeće što smo promatrali bio je utjecaj širine razreda na grešku procjene te gledali ga na različitim luminozitetima i očekivanim vrijednostima signala. Greške su pritom bile manje i za veće luminozitete i za veće očekivane vrijednosti signala, što smo i očekivali. U svim slučajevima greške su lagano rasle sa širinom razreda dok ne bi došli do dovoljno širokih razreda, koji su pokrivali gotovo cijeli signal, kada bi taj rast znatno ubrzao. Promotrili smo i kako nam se mijenja utjecaj širine razreda kada u procjenu uključimo i procjenu mase (ili centra signala) te širinu signala te zaključili da oni ne utječu znatno dok god širina razreda nije takva da obuhvati cijeli signal u jedan razred. Tada greške na procjenu mase i širine signala znatno narastu jer prilagodba predloška na podatke nema osjetljivost na male pomake u tim vrijednostima. Važno je bilo i promotriti kako nam pozicija centra razreda (u slučaju razreda širokih kao signal) utječe na grešku. Iz toga smo zaključili da nam je procjena sigurnija u slučaju da nam je signal ravnomjerno raspodijeljen u više razreda, nego ako imamo signal koji je dominantno raspodijeljen u jednom razredu.

Kako bi iz idealiziranog slučaja, u kojem imamo savršeno glatke krivulje, prešli na realniji slučaj, generirali smo predložak pozadine sa statističkim odstupanjima te ponovno pogledali koji je utjecaj širine razreda na greške. Došli smo do zaključka da se i snage signala i greške ponašaju slično kao i u slučaju savršeno glatkih krivulja izuzevši nekoliko najužih širina. Za te najuže širine dobivamo da je greška znatno manja, a snaga signala veća nego kod ostalih širina razreda te se taj efekt povećava kada je relativno odstupanje predloška pozadine veće. Iako smo očekivali da će za prvih nekoliko točaka biti odstupanja i kod greške i kod signala, očekivali smo da će te greške biti puno veće, a ne manje, što ukazuje da je procjena prilično osjetljiva na male pomake u snazi signala. Još uvijek ne

razumijemo zašto se taj efekt događa.

Na kraju, naučena smo znanja primijenili na histogram invarijantne mase za b kvark-antikvark raspade prikazan na slici 6. Kao i u slučajevima za savršeno glatku krivulju, dobili smo da greške lagano rastu sa širinom razreda sve dok širina razreda nije dovoljno velika da pokrije gotovo cijeli signal. Promotrili smo i kako nam snaga signala i pripadna greška ovise o luminozitetu te smo dobili da snaga signala ne ovisi o luminozitetu, dok greška pada s korijenom luminoziteta, što smo i očekivali.

5 Zahvale

Zahvalio bih se mentoru prof. dr. sc. Mirku Planiniću na podršci i preporuci ovoga rada, kao i savjetima, posebno onome koji me ponukao da pokucam na vrata Laboratorija za fiziku elementarnih čestica na Institutu Ruđer Bošković i izradim ovaj rad.

Posebno se zahvaljujem komentoru izv. prof. dr. sc. Vuki Brigljeviću bez čijeg vodstva istraživanje i izrada rada ne bi bili mogući. Osim stalnog poučavanja o području statistike, ali i fizike elementarnih čestica, na entuzijastičan način prenio je motivaciju rada te me uputio u širu problematiku eksperimentalne fizike elementarnih čestica. Bio je uvijek dostupan za sva pitanja, imala ona izravno veze s radom ili ne, te me definitivno motivirao na proučavanje statistike - nešto za što nikad ne bih očekivao da će se dogoditi - i daljnje istraživanje.

Naposljetku, zahvalio bi se kolegama Filipu Mirkoviću i Luki Matijeviću te posebno kolegici Noi Somun na davanju savjeta tijekom pisanja rada i ukazivanju na jezične pogreške, ali i na trpljenju mene tijekom svih godina studija.

Literatura

[1] Mark Thomson. *Modern particle physics*. New York: Cambridge University Press, 2013. ISBN: 978-1-107-03426-6.

[2] David Barney. „CMS Detector Slice". CMS Collection. Siječanj 2016. URL: http://cds.cern.ch/record/2120661.

[3] Luca Scodellaro. „b tagging in ATLAS and CMS". *5th Large Hadron Collider Physics Conference*. Rujan 2017. arXiv: 1709.01290 [hep-ex].

[4] ATLAS Collaboration. „Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment". *The European Physical Journal C* 76.1 (siječanj 2016.). DOI: 10.1140/epjc/s10052-015-3769-y. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-015-3769-y.

[5] P.A. Zyla i dr. „Review of Particle Physics". *PTEP* 2020.8 (2020.). and 2021 update, str. 083C01. DOI: 10.1093/ptep/ptaa104.

[6] Albert M Sirunyan i dr. „A measurement of the Higgs boson mass in the diphoton decay channel". *Phys. Lett. B* 805 (2020.), str. 135425. DOI: 10.1016/j.physletb.2020.135425. arXiv: 2002.06398 [hep-ex].

[7] D. de Florian i dr. „Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". 2/2017 (listopad 2016.). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[8] Y. Nambu i G. Jona-Lasinio. „Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. I". *Phys. Rev.* 122 (1 travanj 1961.), str. 345–358. DOI: 10.1103/PhysRev.122.345. URL: https://link.aps.org/doi/10.1103/PhysRev.122.345.

[9] Steven Weinberg. „A Model of Leptons". *Phys. Rev. Lett.* 19 (21 studeni 1967.), str. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264. URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[10] Albert M Sirunyan i dr. „Measurements of $t\bar{t}H$ Production and the CP Structure of the Yukawa Interaction between the Higgs Boson and Top Quark in the Diphoton Decay Channel". *Phys. Rev. Lett.* 125.6 (2020.), str. 061801. DOI: 10.1103/PhysRevLett.125.061801. arXiv: 2003.10866 [hep-ex].

[11] *Measurements of properties of the Higgs boson in the four-lepton final state at* $\sqrt{s} = 13$ TeV. Teh. izv. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2621419.

[12] Albert M Sirunyan i dr. „Evidence for the Higgs boson decay to a bottom quark–antiquark pair". *Phys. Lett. B* 780 (2018.), str. 501–532. DOI: 10.1016/j.physletb.2018.02.050. arXiv: 1709.07497 [hep-ex].

[13] M. Aaboud i dr. „Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector". *JHEP* 12 (2017.), str. 024. DOI: 10.1007/JHEP12(2017)024. arXiv: 1708.03299 [hep-ex].

[14] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik und ihrer Grenzgebiete. J. Springer, 1933. URL: https://books.google.hr/books?id=ob4rAAAAYAAJ.

[15] Luca Lista. *Statistical Methods for Data Analysis in Particle Physics*. Sv. 909. Springer, 2016. ISBN: 978-3-319-20175-7, 978-3-319-20176-4. DOI: 10.1007/978-3-319-20176-4.

[16] G. Cowan. *Statistical Data Analysis*. Oxford science publications. Clarendon Press, 1998. ISBN: 9780198501558. URL: https://books.google.hr/books?id=ff8ZyW0nlJAC.

[17] R.J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. Manchester Physics Series. Wiley, 1989. ISBN: 9780471922940. URL: https://books.google.hr/books?id=iEAtnwEACAAJ.

6 Sažetak

Autor: Filip Bilandžija

Naziv rada: Utjecaj izbora razreda u histogramu na signifikantnost signala

U fizici elementarnih čestica, novi fenomeni - procesi ili čestice - otkrivaju se u slučaju kada eksperimentalni podatci značajno odstupaju od predviđanja na temelju dosadašnjih saznanja. Pritom, tražeći nove fenomene, oni nam predstavljaju signal koji što jasnije želimo vidjeti, dok nam dosadašnja saznanja predstavljaju pozadinu. Podatci, i eksperimentalni i simulirani (teorijski), mogu se proučavati gledajući jednodimenzionalne raspodjele opservabli koje ih karakteriziraju. Te se raspodjele mogu prikazati u histogramima te se na temelju njih određuje je li eksperiment u skladu s dosad poznatom teorijom ili eksperimentalni podatci znatno odstupaju od teorijski očekivanih. Pitanje koje se postavlja je može li se izborom razreda to odstupanje, odnosno signifikantnost signala, maksimizirati.

Kako bi proučili kako izbor razreda utječe na signifikantnost signala, odabrali smo pozadinu u obliku padajuće eksponencijale te signal u obliku gausijana. Taj je izbor motivirani izgledom raspodjele invarijante mase za 2 b kvarka dobivene raspadom Higgsovog bozona iz procesa zajedničke tvorbe Higgsovog i Z bozona. Te smo histograme koristili kao predloške za generiranje pseudonasumičnih uzoraka Monte Carlo metodom. Kako bi dobili što bolje slaganje predložaka s podatcima, varirali smo parametre predloška signala i procjenjivali stupanj slaganja koristeći metodu najveće izglednosti. Najprije smo gledali procjenu snage signala u ovisnosti o širini razreda u histogramu. Razrede u histogramu smo dobivali spajajući određen broj susjednih razreda originalnog histograma. Kasnije smo dodali i procjene srednje vrijednosti signala i njegove širine. Pritom smo varirali luminozitet i omjer očekivane vrijednosti signala i pozadine. Kako bi uzeli u obzir da predložak pozadine, zbog konačnog broja simulacija, ima statističke nesigurnosti, uveli smo predložak koji ima značajne statističke fluktuacije. Na kraju, pogledali smo kako nam izbor razreda utječe na snagu signala kod procesa zajedničke tvorbe Higgsovog i Z bozona. Zaključili smo da je signifikantnost signala općenito veća što su luminozitet i očekivana vrijednost signala veći, te da je signifikantnost veća što je širina razreda manja.

Ključne riječi: histogram, procjena, izbor razreda u histogramu, Higgsov bozon, statistička signifikantnost

7 Summary

Author: Filip Bilandžija

Title: Signal significance dependency on binning choice

In modern particle physics, new phenomena - both particles and processes - are discovered when experimental data significantly deviate from predictions based on our present knowledge. When looking for new phenomena, that new phenomenon is considered as the signal, which we want to see as clearly as possible, while our present knowledge is treated as the background. Both experimental data and data simulated from theoretical models can be studied by looking one-dimensional distribution of observables characterizing them. That distribution can be graphically shown in form of a histogram. There it can be deduced whether experimental data matches simulated data or whether it significantly deviates from it. A question that naturally arises is whether we can maximize that deviation, i.e. signal significance, with our binning choice.

In order to investigate how binning choice affects signal significance, we chose a background with the shape of exponentially decreasing function and a signal of Gaussian shape. That choice was motivated by the shape the of invariant mass distribution of 2 b quarks from Higgs decay in the case of associated production of Higgs and Z bosons. We used those histograms (the background and the signal being separated) as templates for generating data samples with the Monte Carlo method. In order to obtain the optimal fit between data samples and templates, we varied the parameters of the signal template and estimated the "goodness" of the fit using the maximum likelihood method. First, we looked at the dependence of signal strength estimation in regard to the width of the bins in the histogram, by combining several adjacent bins into one for the background and the signal alike. Later, we included signal mean value and width estimations as well. In doing so, the luminosity and ratio of the expected value of the signal and the background have also been varied. To include the effect of the finite number of simulated events, which results in statistical uncertainties in the background template, we introduced the background template which has large statistical fluctuations. Finally, we investigated how binning choice affects the signal strength of the associated production of Higgs and Z boson. We concluded that signal significance is generally higher when luminosity and expected value of signal are higher and that signal significance is higher when bins in the histogram are narrower compared to the signal width.

Keywords: histogram, estimation, rebinning, Higgs boson, statistical significance