# 1. naloga
# Modeliranje 1-D porazdelitev: Razpadi Higgsovega bozona

Tadej Lozej 28242023

14. oktober 2025

Praktikum strojnega učenja v fiziki

Predavatelj: red. prof. dr. Borut Paul Kerševan
Asistent: Jan Gavranovič

# Kazalo

# 1 Uvod

Iskanje signala v veliki količini podatkov je velik izziv, saj so le-ti praviloma težko ločljivi od dominantnih procesov ozadja. Pomembno je, da kinematično porazdelitev ozadja čim bolj natančno opišemo, da ga lahko nato odštejemo od podatkov. Na ta način nam nato ostane le možen iskani signal.

Pri opisu procesov ozadja si pomagamo z regresijo ('fitom', parametrizacijo) kinematičnih porazdelitev, ki nas zanimajo. Postopek regresije ozadja izhaja ali iz simuliranih porazdelitev ali iz predpostavljenih funkcijskih oblik, ki jih poskušamo prilagoditi podatkom v kinematičnem območju porazdelitve, kjer signala ne pričakujemo in jih nato ekstrapoliramo v signalno območje.

Metode regresije so danes pomemben element metod strojnega učenja, kjer poskušamo iz omenjene količine podatkov izvleči čim več informacij o samem procesu.

V dani nalogi bomo obravnavali meritev eksperimenta ATLAS, kjer so iskali redek razpad Higgsovega bozona v dva miona ($H \to \mu^+\mu^-$). Ogledali si bomo kinematično porazdelitev dogodkov po rekonstruirani invariantni masi dveh izbranih mionov ($m_{\mu\mu}$). Glavni proces ozdaja je razpad šibkega bozona ($Z \to \mu^+\mu^-$), ki prispeva resonančno porazdelitev z vrhom pri $m_Z = 91\,\mathrm{GeV}$, kjer pa se rep invariantne mase razširi tudi precej nad maso Higgsovega bozona $m_H = 125\,\mathrm{GeV}$. Ta proces ozadja je za več velikostnih redov dogodkov večji, kot jih pričakujemo iz signalnega razpada Higgsovega bozona. K ozadju v manjši meri prispevajo tudi drugi procesi.

Za to nalogo so na razpolago simulirani in izmerjeni dogodki kolaboracije ATLAS. V primeru simulacij so dogodki uteženi tako, da je vsota uteži enaka pričakovani vrednosti števila izmerjenih dogodkov za vključene procese ($\sum wt = N_{proc}$). Celotna statistična napaka je $\sigma = \sqrt{\sum wt^2}$. To za simulacijo pomeni, da napaka ni Poissonova.

Signal lahko za potrebe določitve ozadja omejimo na primer na območje [120 Gev, 130 GeV]. Na voljo imamo cel šop metod za ustrezno določitev ozadja

- Polinom dovolj nizkega reda.
- Bolj sofisticirane funkcije (mBW)
- Support Vectom Machines (SVM) metode z različnimi jedri in regulizatorji
- Uporabimo gaussovske procese (Gaussian Process Regression, GPR) z različnimi jedri
- Uporabimo dekompozicijo na ortogonalne polinome

Za parametrizacijo signala se tipično uporabi Crystal Ball (CB) funkcija.

# 2 Navodila in usmeritve

1. Iz surovih podatkov zgeneriraj svoje histograme s pomočjo predpripravljene skripte `create_histograms.py`, pri kateri lahko spreminjaš število predalov in $m_{\mu\mu}$ interval, ki ga boš opazoval/-a. Histogrami (mejne in sredinske $x$ vrednosti predalov, vrednosti in napake) se shranijo v formatu `.npz`.

2. Ko imaš zgenerirane svoje histograme, jih lahko izrišeš s pomočjo skripte `visualize_data.py` (ustrezno s točko spremeni ime datotek, ki jih nalagaš).

3. Preveri, če imaš napake res pravilno upoštevane. Lahko jih namenoma pokvariš, in ponoviš prva dva koraka, da vidiš vpliv.

4. Da se spoznaš z osnovnim fitanjem, najprej zgladi histogram simuliranega ozadja s pomočjo preprostejših matematičnih funkcij in nadaljuj do različnih teoretično podkrepljenih nastavkov. Dobiš funkcijo $m(x_k)$.

5. Prilagodi funkcijo CB histogramu simuliranega signala, pri čemer upoštevaj še dodatni normalizacijski faktor. Dobiš funkcijo $s(x_k)$.

6. Ker simulacija ozadja ni vedno najboljša, se po navadi za oceno ozadja vzame izmerjene podatke, pri čemer pa je potrebno izključiti območje, kjer pričakujemo signal. Prilagodi torej funkcijo histogramu podatkov, da dobiš dobro oceno za ozadje in pri tem pazi, da pri fitu ne upoštevaš območja okrog mase Higgsovega bozona. Dobiš funkcijo $b(x_k)$.

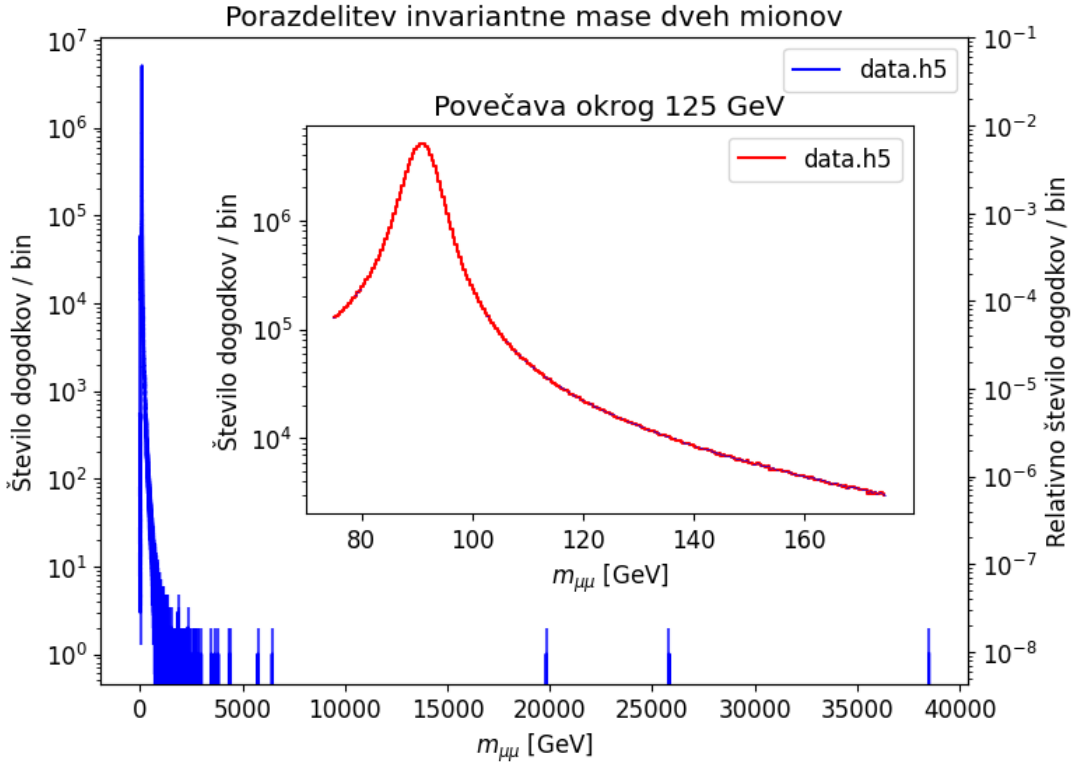
7. Od podatkov odštej čimbolj zglajeno ozadje, ki si ga dobil v prejšnji točki, da dobiš ekstrahiran signal $y(x_k)$.

8. Na ekstrahiran signal fitaj CB funkcijo s prostimi parametri, ki si jih dobil v točki 5. tako, da ji v resnici prilagodiš le nov normalizacijski faktor $\alpha_{norm} \cdot s(x_k)$.

9. Ker je izmerjenega signala še zelo malo, predlagamo, da postopek najprej narediš z umetno napihnjenim signalom - le tega množi s faktorjem $\gamma = 100$ in ga dodaj podatkom $d_{new}(x_k) = d(x_k) + \gamma s(x_k)$. Ker bo ta signal na ta način lepo izstopal iz ozadja, ga boš lažje izluščil.

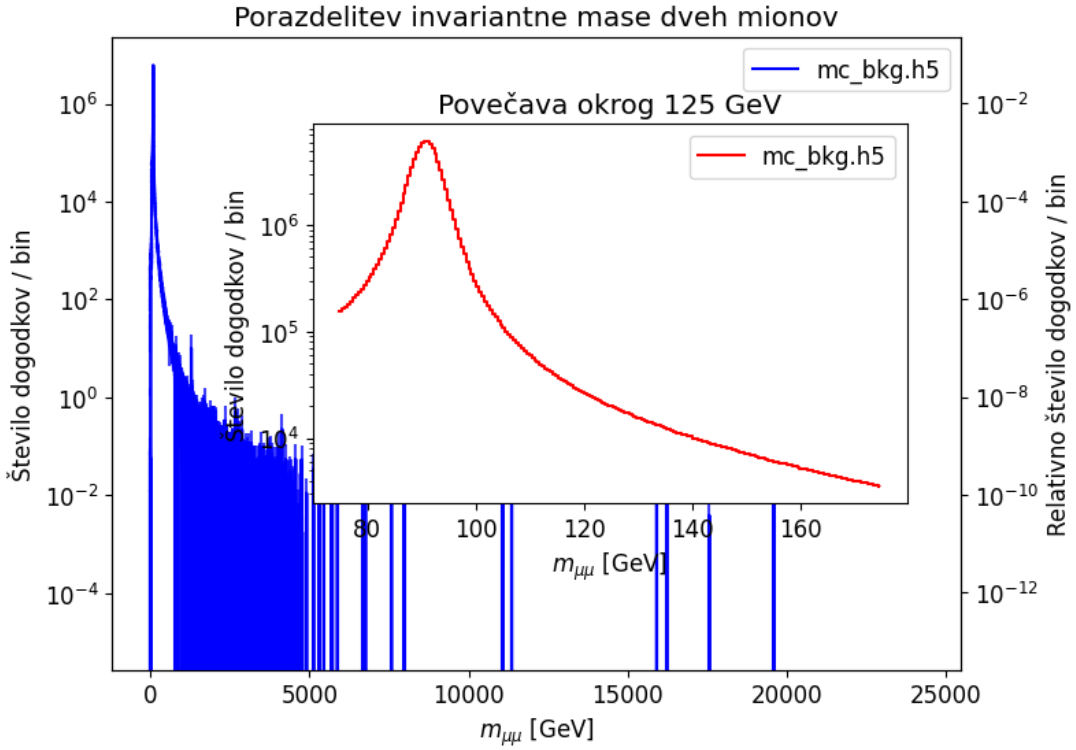
## 3 Podatki

Za začetek si oglejmo realne podatke, ki jih imamo v datoteki `data.h5`. V tej datoteki imamo na voljo 106148612 dogodkov. Na sliki 1 imamo histogram z 100 000 predalčki med minimalno in maksimalno vrednostjo invariantne mase dogodkov. Na levi osi vidimo absolutno število dogodkov pri določenem predalčku invariantne mase na desni osi pa relativno število dogodkov. Na manjše je prikazan tuki histogram v široki okolici mase Higgsovega bozona $m_H = 125\,\text{GeV}$. Vidimo visok vrh pri razpadu šibkega bozona pri invariantni masi $m_Z = 91\,\text{GeV}$. Vrh Higgsovega bozona na histogramu ni niti viden.

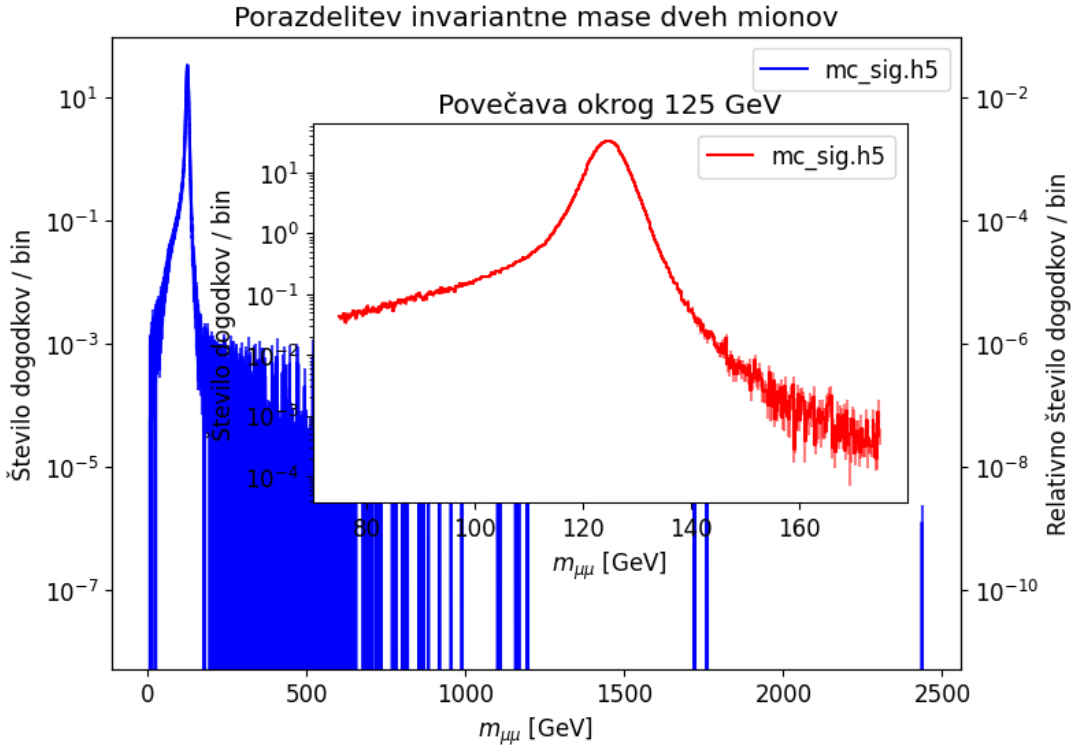



Slika 1: Histogam porazdelitve invariantne mase dveh mionov iz datoteke `data.h5`. Na levi osi je absolutna skala na desni pa relativna. Na manjše je prikazan histogram v okolici invariantne mase Higgsovega bozona.

Na sliki 2 je prikazana porazdelitev invariantne mase dveh mionov iz datoteke `mc_bkg.h5`. Na voljo imamo 311281902 dogodkov. Razporedil sem jih v 50 000 enako velikih predalčkov med minimalno in maksimalno invariantno maso. Te podatke so pridobili s pomočjo Monte Carlo simulacije. Simulirali so zgolj ozadje procesa. Vsak dogodek je utežen in zato smo pri risanju histograma morali biti na to pozorni in




Slika 2: Histogam porazdelitve invariantne mase dveh mionov iz datoteke `mc_bkg.h5`. Na levi osi je absolutna skala na desni pa relativna. Na manjše je prikazan histogram v okolici invariantne mase Higgsovega bozona.



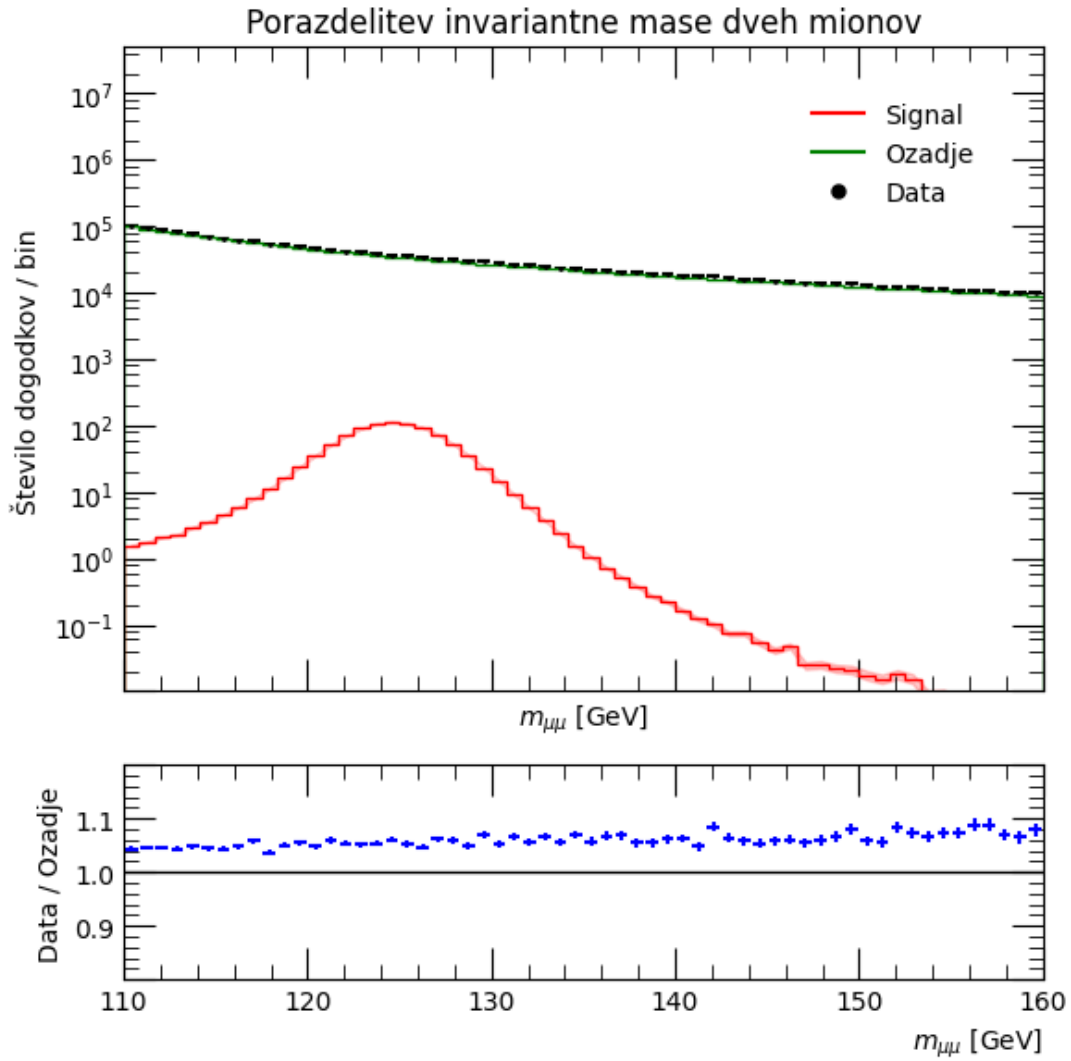

Slika 3: Histogam porazdelitve invariantne mase dveh mionov iz datoteke `mc_sig.h5`. Na levi osi je absolutna skala na desni pa relativna. Na manjše je prikazan histogram v okolici invariantne mase Higgsovega bozona.

dogodek šteti sladno z njegovo utežjo. Kot na prejšnjem grafu sta tudi tukaj leva in desna os grafa absolutno ter relativno število dogodkov. Vidimo, da je histogram na prvi pogled precej podoben. Na manjše imamo

kot prej prikazan histogram v širši okolici invariantne mase Higgsovega bozona.

Na sliki 3 lahko vidimo porazdelitev invariantne mase dveh mionov iz datoteke `mc_sig.h5`. Na voljo imamo 5498243 dogodkov. Razporedil sem jih v 10 000 enako velikih predalčkov med minimalno in maksimalno invariantno maso. Te podatke so pridobili s pomočjo Monte Carlo simulacije. Simulirali so zgolj signal procesa t. j. razpad Higgsovega bozona. Enako kot prej imamo na levi ter desni absolutno ter relativno preštete dogodke v predalčkih invariantnih mas. Vidimo, da je tovrstnih dogodkov precej malo. Na manjšem grafu z rdečo vidimo histogram v širši okolici invariantne mase Higgsovega bozona.

Na sliki 4 je prikazan histogram generiran s funkcijo `make_histograms` in narisan s funkcijo `visualize_histograms`. Vidimo porazdelitev invariantne mase dveh mionov v merjenih podatkih (Data), simuliranem ozadju (Ozadje) ter simuliranem signalu (Signal). Na spodnjem histogramu lahko vidimo razmerje med številom merjenih dogodkov ter simuliranih dogodkov v ozadju.




Slika 4: Histogam porazdelitve invariantne mase dveh mionov izvseh treh datotek. Spodaj je prikazano razmerje med merjenimi podatki ter simuliranim ozadjem.

## 4 Glajenje ozadja z analitičnimi funkcijami

Cilj naloge je odšteti ozadje meritvam in izluščiti signal. Za ta proces bo potrebno dobro modelirati ozadje. V tem delu naloge bomo ozadje poskušali modelirati z analitičnimi funkcijami

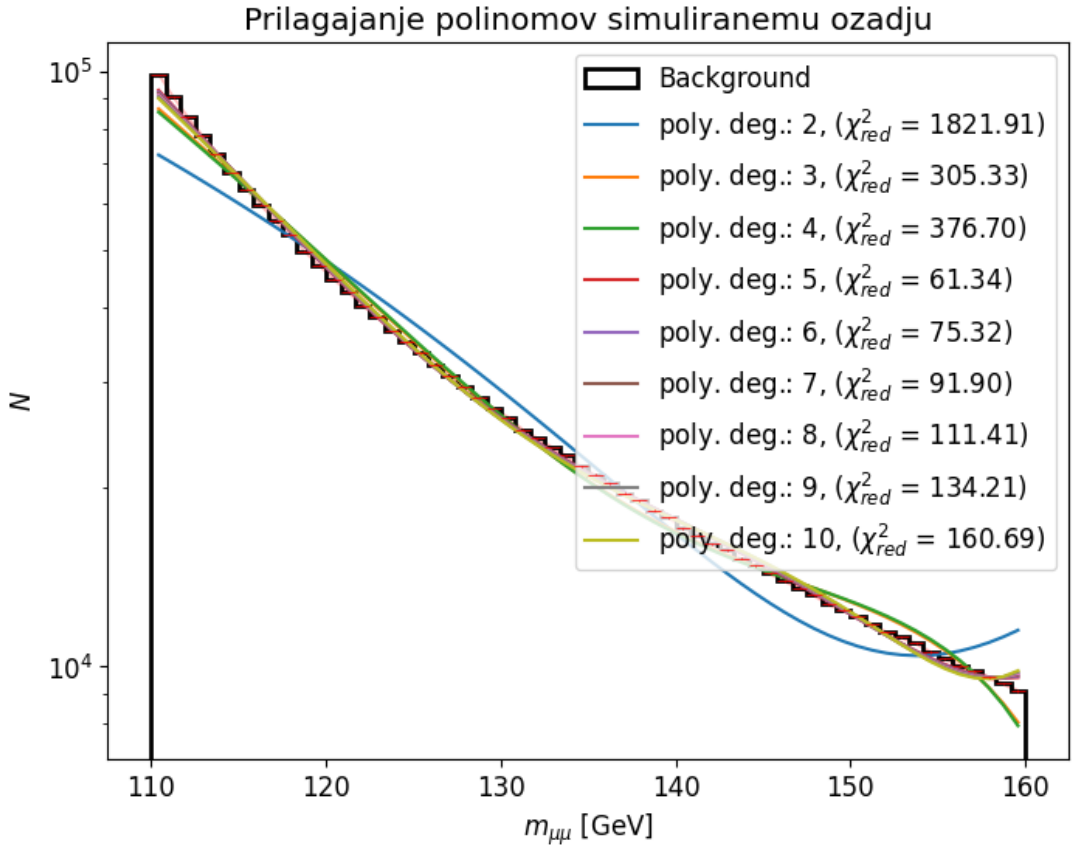
- Polinomi različnih stopenj: $a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ...$

- Potenčna funkcija: $ax^b$
- Eksponentna funkcija: $a_1 \exp(-b_1 x) + a_2 \exp(-b_2 x)$
- Dijet funkcija: $a(1-x^d)^c x^{b_1+b_2 \ln(x)}$
- Eksponent polinoma: $a \exp(b_1 x + b_2 x^2 + b_3 x^3)$
- CMS funkcija: $\exp(a_2 m + a_3 m^2)/((m-m_Z)^{a_1} + (0.5 g_Z)^{a_1})$,

kjer je $x$ število dogodkov, ki se pojavijo v histogramu, $m_Z$ in $g_Z$ pa lastnosti Z bozona. Vse ostalo so prosti parametri.

### 4.1 Simulirano ozadje

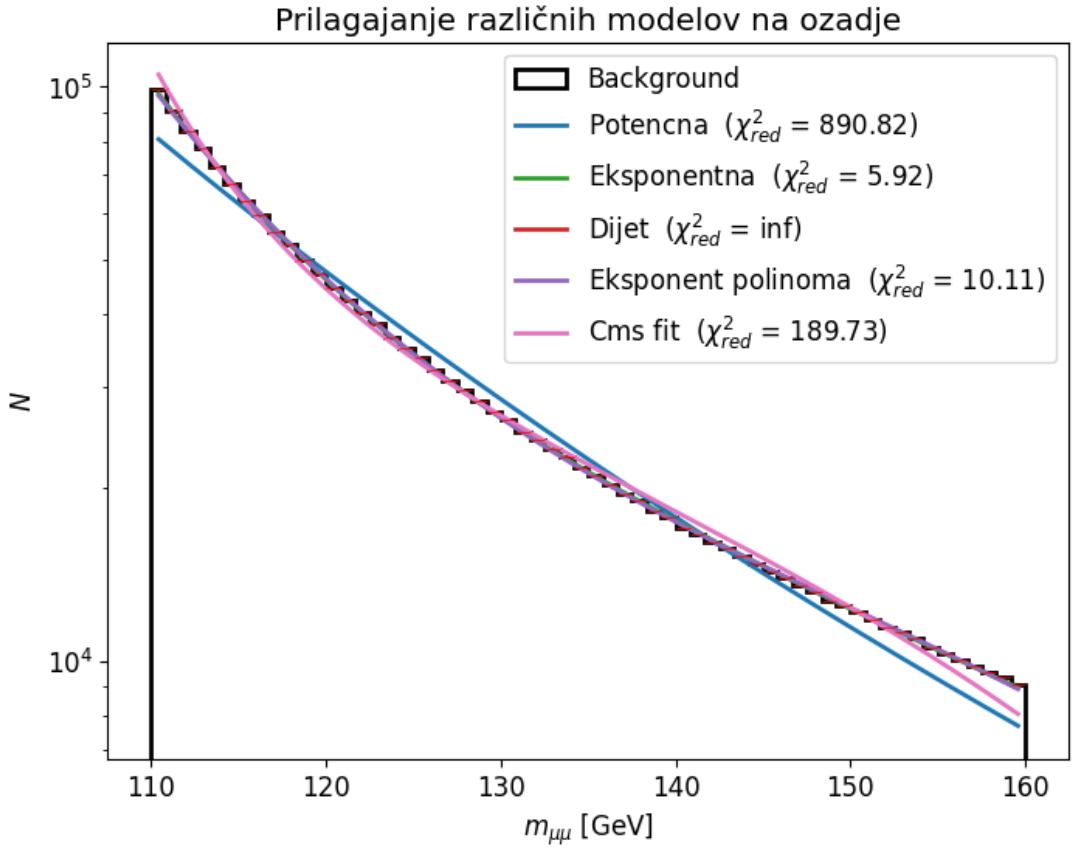
Osredotočimo se na simulirano ozadje iz datoteke `mc_sig.h5`. Z različnimi funkcijami bomo poskusili čimbolje opisati ozadje na intervalu prikazanem na sliki 4. Za začetek si na sliki 5 oglejmo različne fite polinomov na simulirano ozadje. Z različnimi barvami so prikazane različne stopnje polinomov. V legendi poleg tega piše tudi vrednost reduciranega chi kvadrata fita $\chi^2_{red}$. To je navaden $\chi^2$ deljen s prostorskimi stopnjami modela $\nu = N - p$, kjer je $N$ število izmerjenih točk ter $p$ število prostih parametrov modela.To vrednost povezujemo z kvaliteto modela. Če je vrednost enaka 1 pomeni, da model precej dobro opiše obnašanje in tudi ni preveč prilagojen na konkreten primer. Če je vrednost manjša od 1 je bodisi naš model preveč natančen ali pa so napake meritev podcenjene. Če je vrednost večja kot 1 pa bodisi naš model ni preveč dober bodisi so napake precenjene. Vidimo, da v primeru polinomov najmanjšo vrednost zavzame polinom s stopnjo 5.



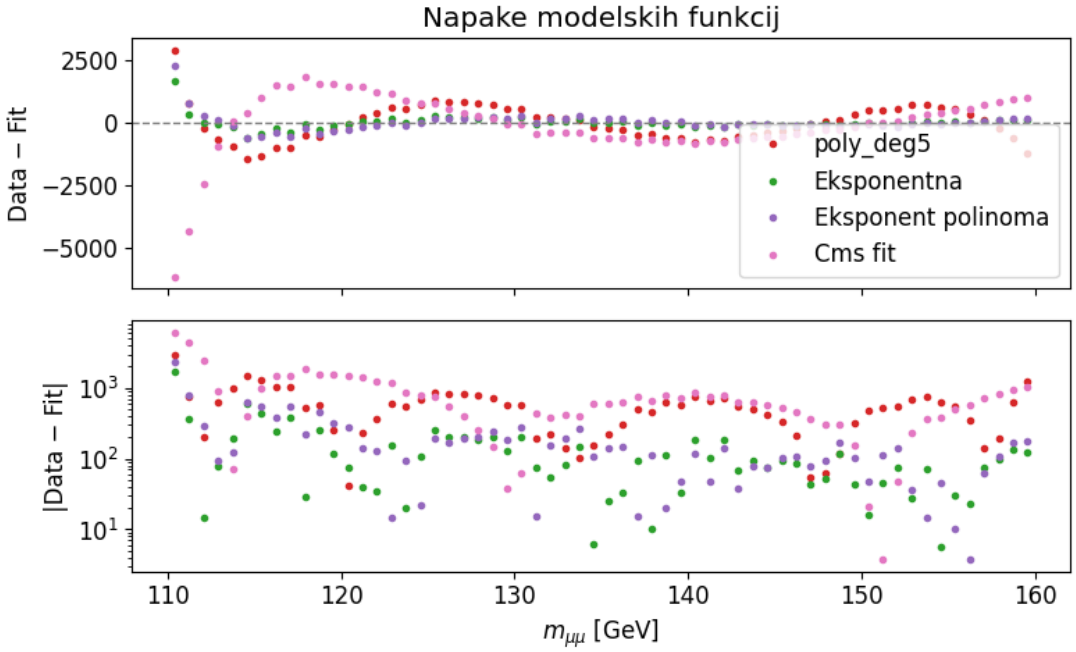

Slika 5: Polinomi različnih stopenj prilagojeni simuliranemu ozadju. V legendi imamo za vsak model podano tudi vrednost reduciranega $\chi^2_{red}$.

Naprej si poglejmo še preostale analitične funkcije. Na sliki 6 lahko vidimo preostale omenjene modelske funkcije. Eksponentna in eksponent polinoma imata zelo majhno vrednost $\chi^2_{red}$. Tudi bolj teoretično

podkovana funkcija CMS se precej lepo prilagaja ozadju. Program je imel težave pri ocenjevanju napak parametrov Dijet funkcije in zato tudi ni znal izračunati pripadajoče vrednosti $\chi^2_{red}$.




Slika 6: Različne modelske funkcije prilagojene simuliranemu ozadju. V legendi imamo za vsak model podano tudi vrednost reduciranega $\chi^2_{red}$.

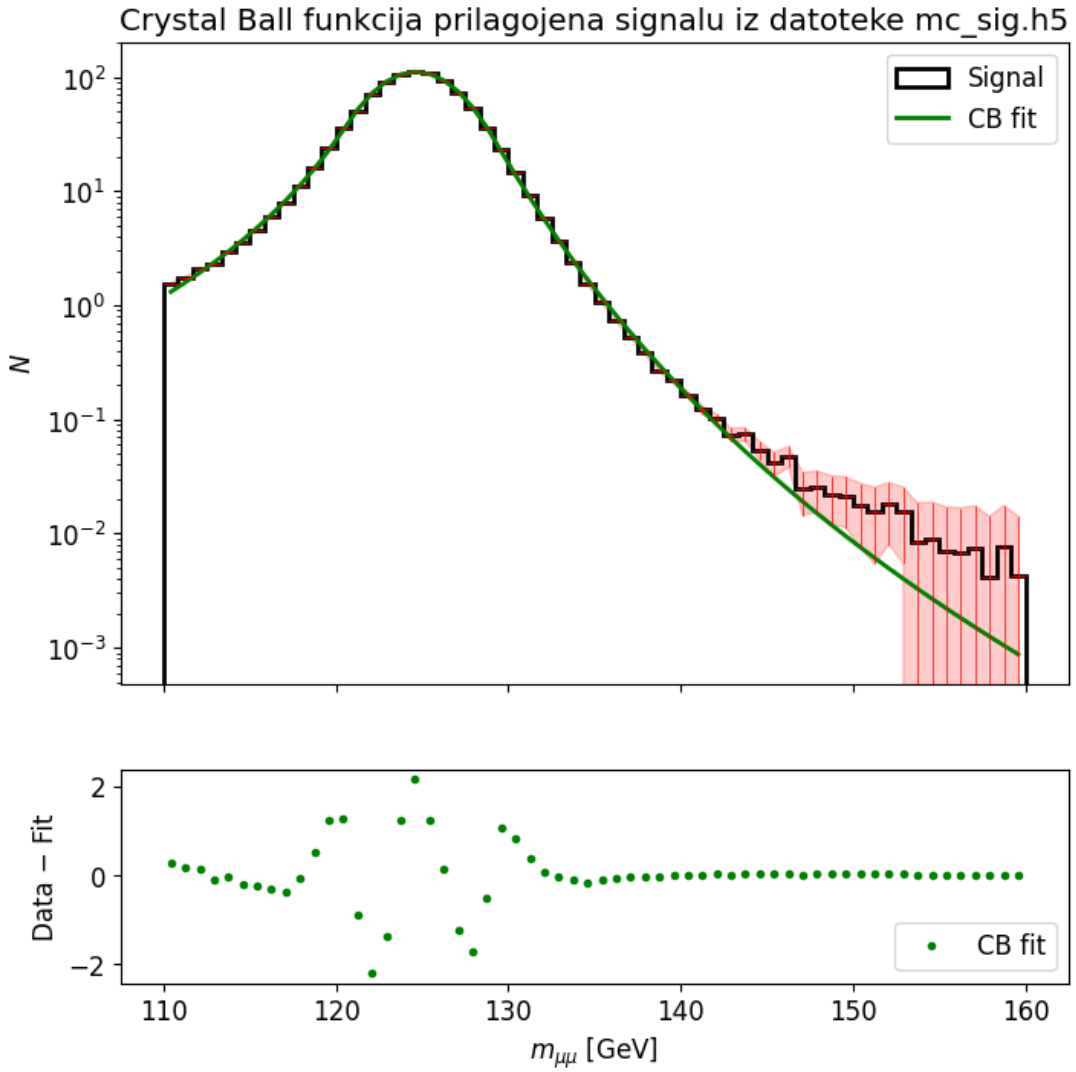



Slika 7: Napake najboljših modelskih funkcij. Zgoraj v linearni skali spodaj pa v logaritemski.

Na sliki 7 lahko vidimo napake najboljših modelov. Vidimo, da se v območju mase Higgsovega bozona najbolje prilegata eksponent in eksponent polinoma. Nevem pa, če bi temu lahko rekli dobro prileganje...

### 4.2 Simuliran signal

Za vajo lahko poskusimo prilagoditi Crystal Ball funkcijo simuliranemu signalu iz datoteke `mc_sig.h5`. Na sliki 8 lahko vidimo Crystal Ball funkcijo prilagojeno simuliranemu signalu. Spodaj lahko vidimo tudi njeno odstopanje. Prilagojena funkcija se precej dobro ujema s podatki.

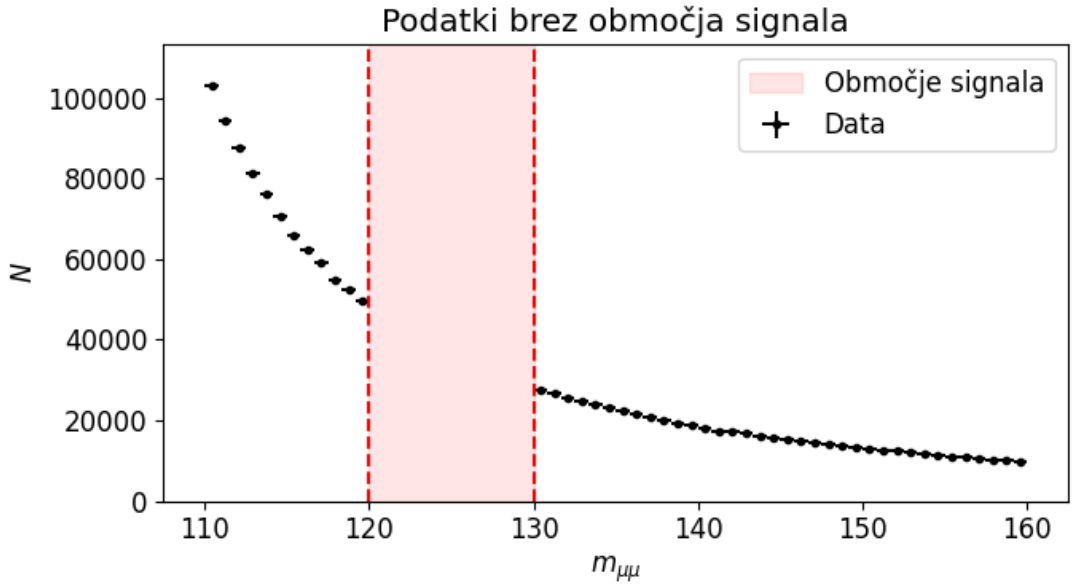



Slika 8: Crystal Ball funkcija prilagojena simuliranemu signalu. Sporaj je prikazano odstopanje fita od podatkov.

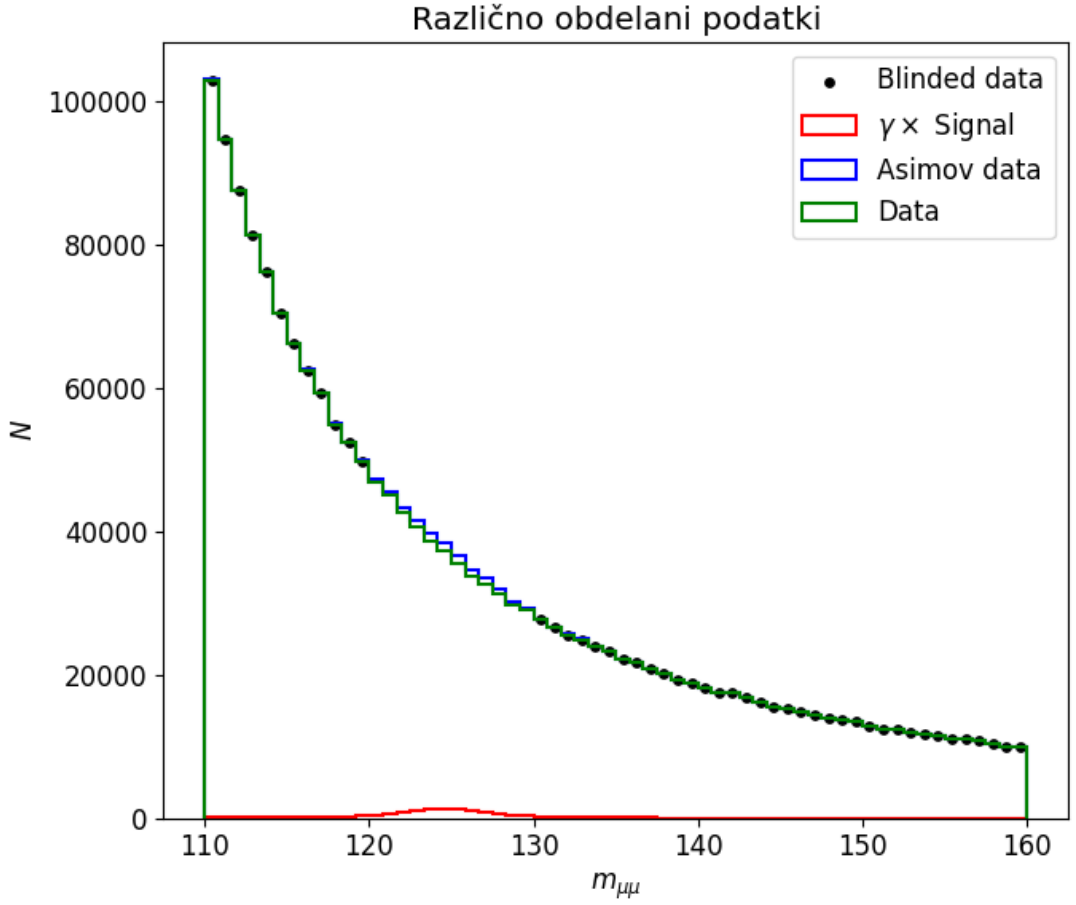
### 4.3 Dejanski in napihnjeni podatki

Poglejmo si sedaj še prilagajanje funkcij dejanskih podatkom iz datoteke `data.h5`. Pazizi moramo, da ne upoštevamo območja signala. Na sliki 9 lahko vidimo podatke z izrezanim območjem v okolici vrha signala. Območje smo odrezali med invariantno maso 120 ter 130 GeV.

Na sliki 10 imamo prikazane različne podatke. Z črnimi pikami vidimo podatke iz katerih smo odvzeli signalno območje. Tem bomo prilagajali naše modelske funkcije. Zelen histogram prikazuje vse podatke. Rdeč histogram prikazuje podatke iz simulacije signala pomnožene s faktorjem $\gamma = 10$. Moder histogram prikazuje podatke katerim smo prišteli umetno napihnjen simuliran signal. Tako je moder histogram v resnici zeles s prištetim rdečim. Tem podatkom bomo rekli Asimov podatki.



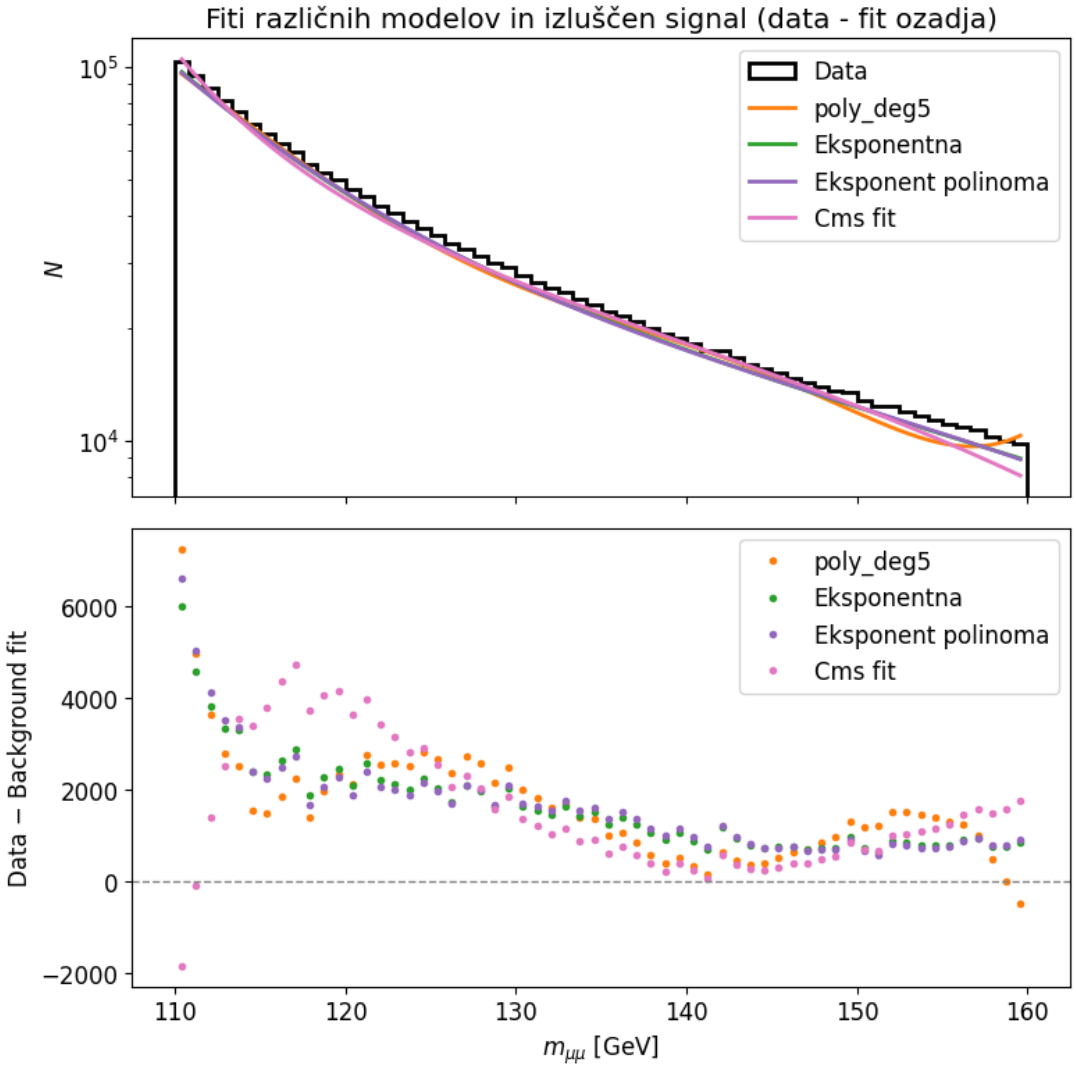

Slika 9: Dejanski podatki brez območja signala.



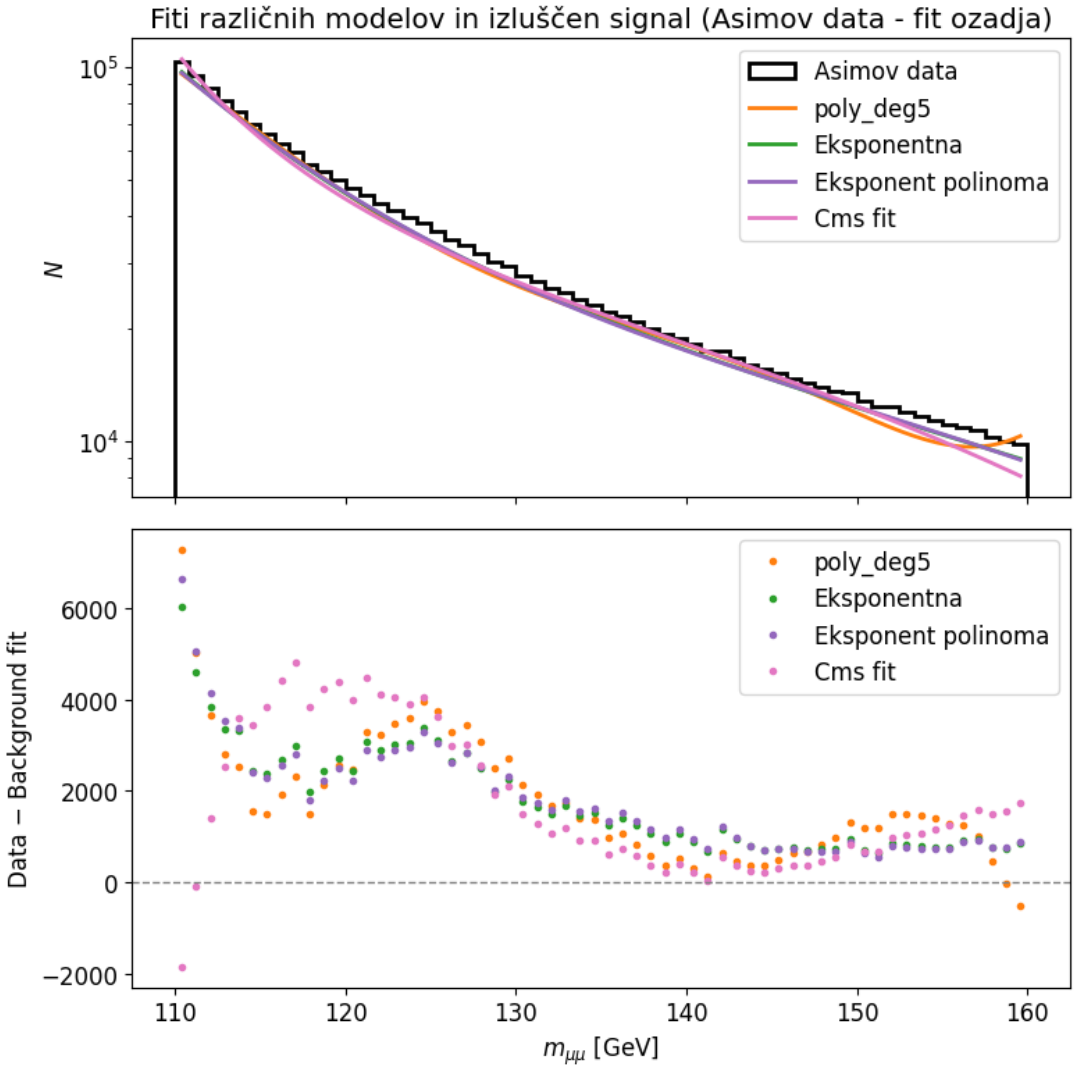

Slika 10: Različni podatki. Z črnimi pikami so prikazani podatki brez območja signala. Tem bomo prilagajali funkcije. Nato so pa po vrsti Napihnjen signal, Asimov podatki ter podatki.

Na naslednjih dveh slikah 11 in 12 lahko vidimo funkcije ozadja narisane čez podatke ter napšihnjene podatke ter izluščen signal iz podatkov ter napihnjenih podatkov. Kot pričakovano smo več signala dobili iz podatkov z napihnjenim signalom. Izluščenim signalom sem se trudil prilagoditi CrystallBall funkcijo

ampak žal neuspešno.




Slika 11: Funkcije prilagojene ozadju narisena čez podatke zgoraj ter izluščen signal spodaj.




Slika 12: Funkcije prilagojene ozadju narisena čez napihnjene podatke zgoraj ter izluščen signal spodaj.

# 5 Glajenje ozadja z Gaussovskimi procesi

Ozadje lahko modeliramo tudi z Gaussovskimi procesi (GPR – Gaussian Process Regression). Gre za metodo strojnega učenja, ki je posebej uporabna, kadar imamo opravka s podatki, katerih odvisnost od spremenljivk ni znana ali jo je težko eksplicitno opisati. Metoda temelji na uporabi jedrskih funkcij (kernels), ki določajo stopnjo podobnosti med podatkovnimi točkami in s tem vplivajo na gladkost ter obliko napovedane funkcije. Z ustrezno izbiro jedra lahko izboljšamo napoved. Preizkusil sem več jeder a smiselne rezultate sem dobil le z RBF, Matern, RQ ter ExpSineSquared jedri.

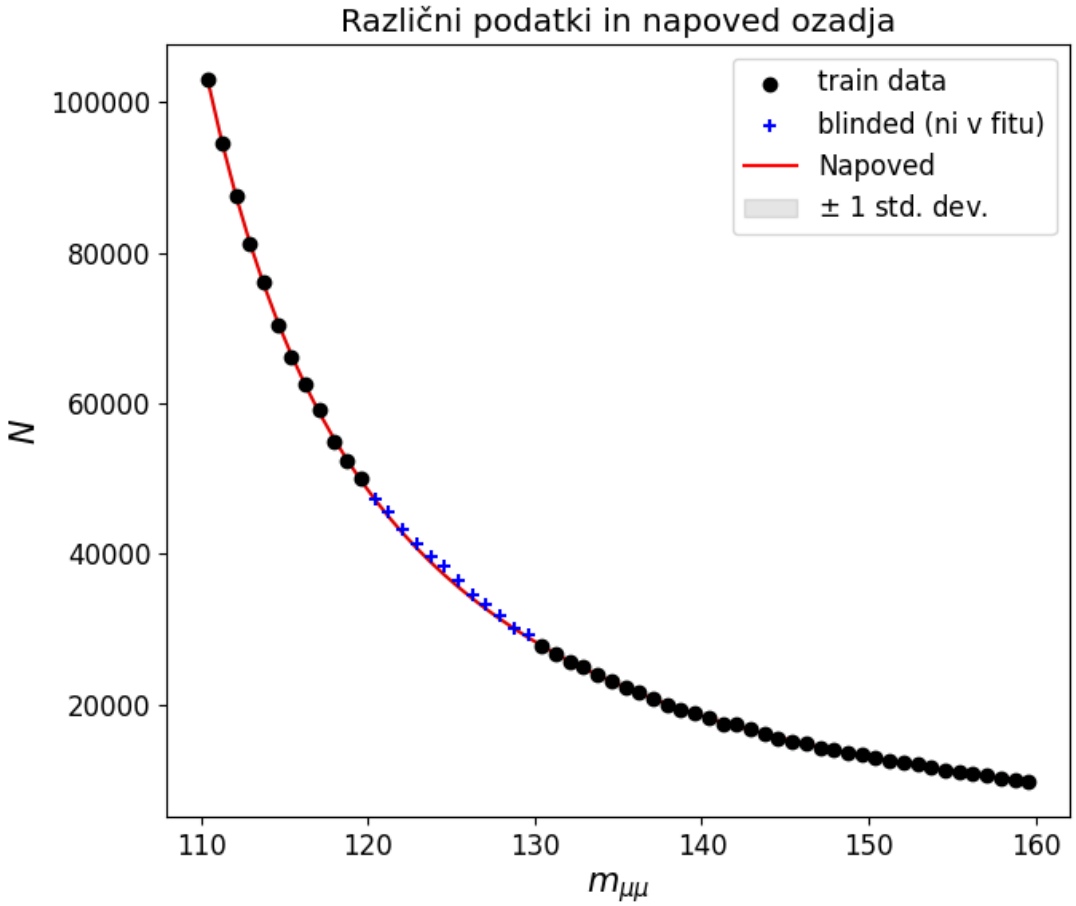
Prav tako moramo vhodne podatke ustrezno normalizirati, saj tako vplivamo na stabilnost in učinkovitost učenja modela. V analizi sem poreizkusil tri vrste normalizacij: standardno, min-max ter logaritemsko. S standardno normalizacijo žal iz nekega razloga nisem dobil smiselnih rezultatov, zato bom podal le rezultate dobljene z min-max ter logaritemsko normalizacijo.

## 5.1 Luščenje signala iz Asimov podakov

V tem podpoglavju bomo poiskusili modelirati ozadje Asimov podatkov. To so podatki, katerim umetno prištejemo signal.

Enako kot pri prilagajanju analitičnih funkcij podatkom moramo iz podatkov odstraniti območje okrog invariantne mase Higgsovega bozona. Na sliki 13 lahko vidimo podake uporabljene za trening, podatke, ki v

treningu niso bili uporabljeni ter predikcijo našega modela. Za primere v tem podpodpoglavju sem uporabil logaritemsko normalizacijo ter RGF jedro - enako kot na vajah.



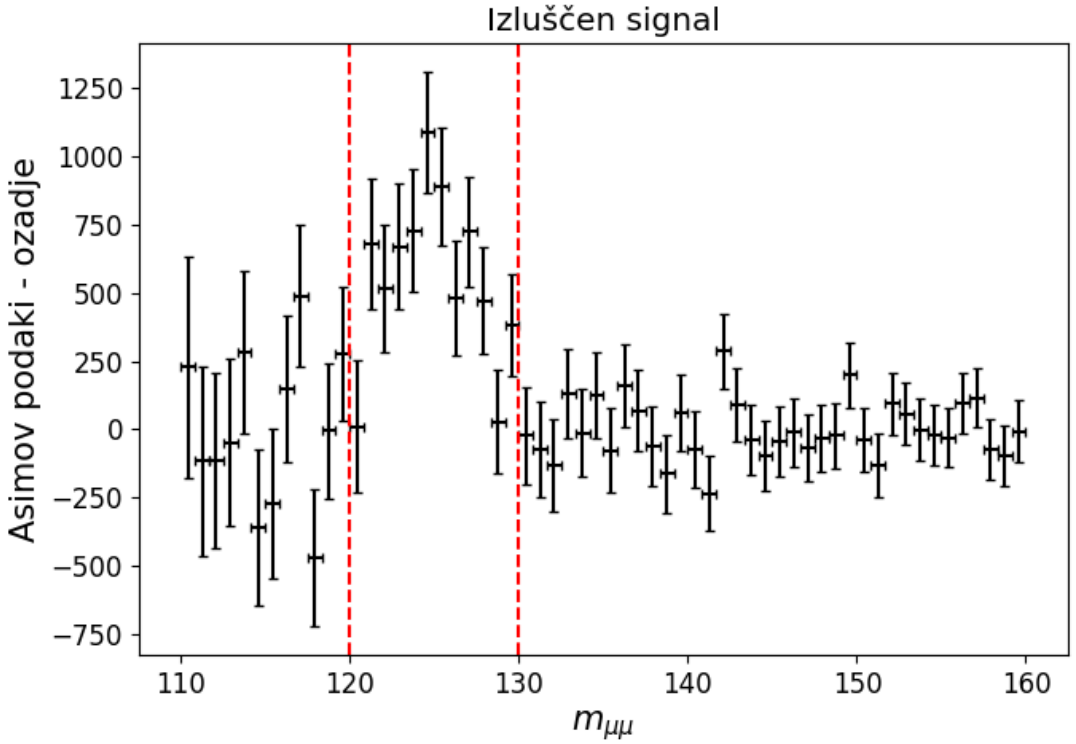

Slika 13: S črno so prikazani podatki uporabljeni za trening, z modro so prikazani podatki, ki so iz treninga izpuščeni in z rdečo je prikazana napoved našega modela. Uporabil sem logaritemsko normalizacijo ter RBF jedro.

Če Asimov podatkom odštejemo modelirano ozadje dobimo izluščen signal prikazan na sliki 14. Oblika signala spominja na CrystallBall funkcijo.

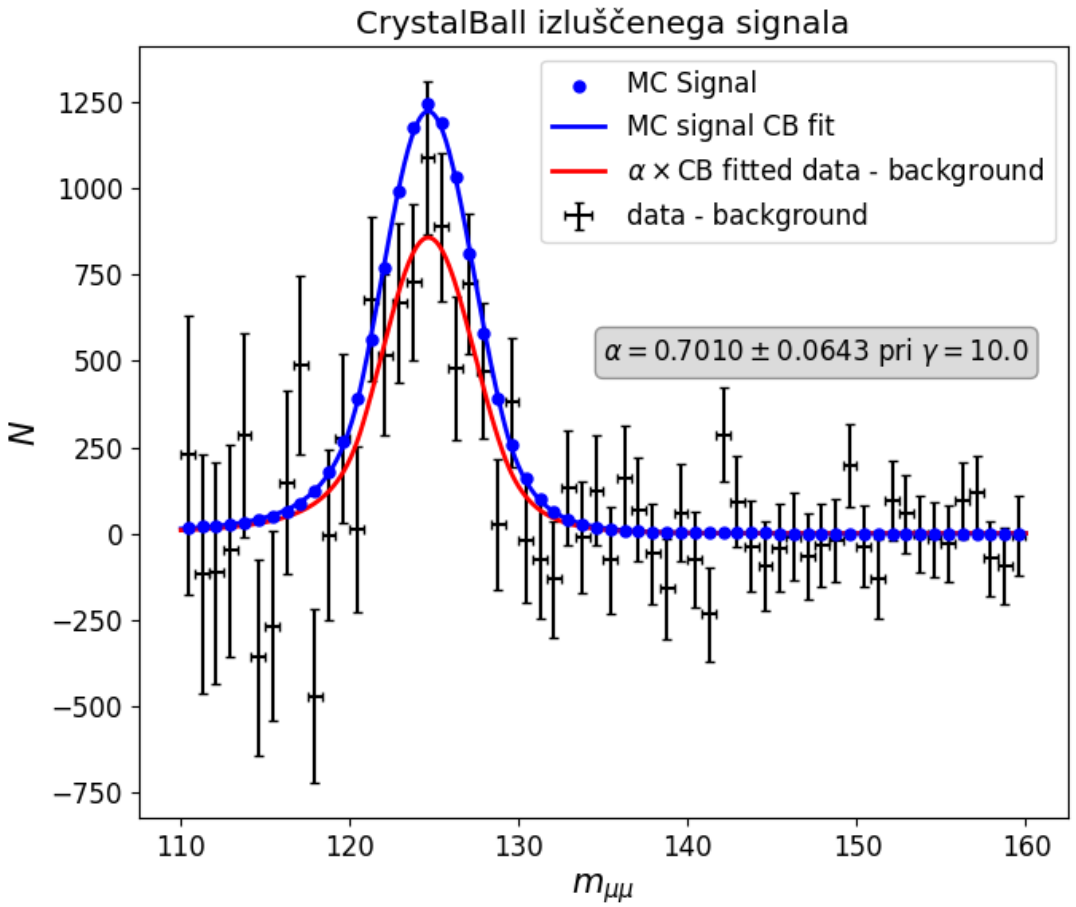
Na sliki 15 lahko z rdečo vidimo CristallBall funkcijo prilagojeno našemu izluščenemu signalu. Z modro vidimo CrystallBall funkcijo, ki je prilagojena zgolj simuliranemu signalu. S parametrom $\alpha$ lahko vidimo, kako dobro smo izluščli signal iz podatkov. Ta nam pove razmerje med izluščenim signalom in simuliranim signalom.

Na sliki 16 lahko vidimo vrednosti parametra alpha za različna uporabljena jedra ter za dve različni normalizaciji podatkov. Vidimo, da se dejanskemu signalu najbolj približamo z uporabo RBF jedra ter logaritemske normalizacije ter ExpSineSquared jedra z uporabo logaritemske normalizacije. V splošnem smo z logaritemsko normalizacijo bolje izluščli signal kot z min - max noralizacijo.

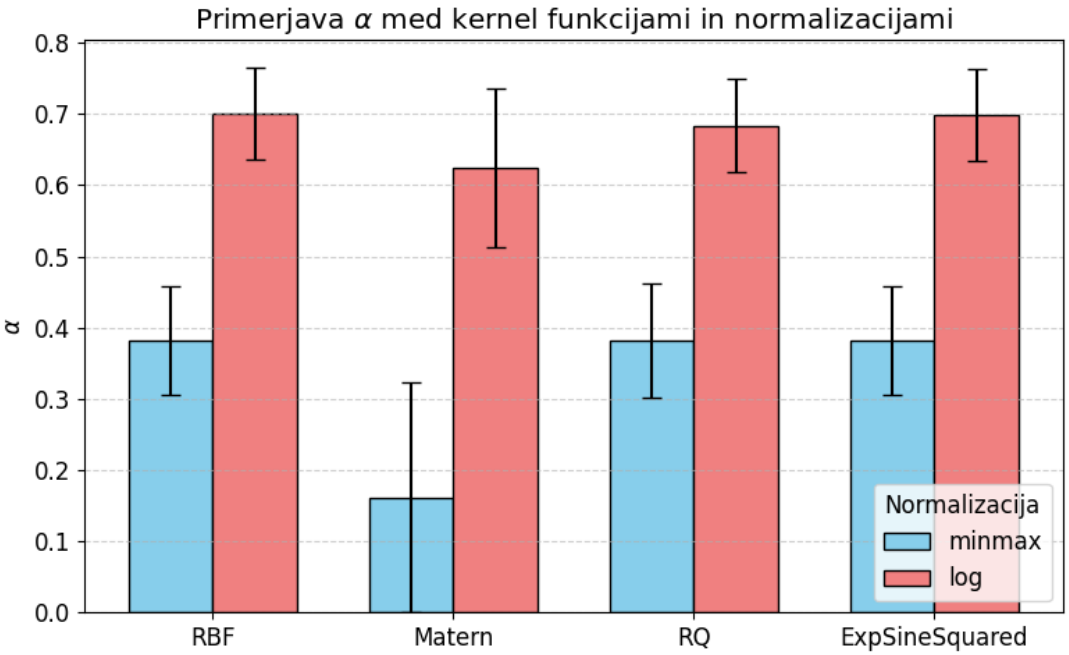
## 5.2 Luščenje signala iz podatkov




Slika 14: Izluščen signal iz Asimov podatkov.




Slika 15: CrystallBall izluščenega signala.




Slika 16: CrystallBall izluščenega signala.