DOMAĆI 4-Analiza masa crnih rupa

Tatjana Novaković

August 8, 2025

Uvod

Analiziramo skup masa crnih rupa (`bhm.npy`):

1. računamo osnovne statistike (**standardna devijacija**, **medijana**, **asimetrija**) i procjenjujemo njihove neodređenosti metodama **Bootstrap** i **Jackknife (leave-two-out)**;
2. procjenjujemo uticaja dodavanja *ekstremnih vrijednosti*;
3. pronalazak *najviših pikova* raspodjele masa;
4. fitovanje **Gaussian Mixture Modela (GMM)** za $N = 1, \dots, 10$, izbor najboljeg modela prema AIC, interpretaciju komponenti i ispitivanje stabilnosti parametara $(\alpha_j, \mu_j, \sigma_j)$.

1 Podaci i priprema

Koristimo niz masa $m_1, \dots, m_n$ (u jedinicama $M_\odot$). Za homogenost i replikabilnost uzorkovali smo $n = 10{,}000$ podataka sa ponavljanjem iz izvornog niza. Osnovne procjene:

$$\hat{\sigma} = \mathrm{std}(m; \texttt{ddof} = 1), \qquad \widehat{\mathrm{med}} = \mathrm{median}(m), \qquad \widehat{\mathrm{skew}} = \mathrm{skew}(m).$$

Koristimo `ddof=1` radi nepristrasnosti procjene σ za uzorak.

2 Bootstrap: metod i rezultati

Ideja i implementacija

Bootstrap aproksimira distribuciju statistike resampliranjem sa vraćanjem. Generiše se $B = 10{,}000$ bootstrap uzoraka $m^{*(b)}$ (svaki iste veličine n), pa se računa

$$\theta^{*(b)} = \theta\Big(m^{*(b)}\Big), \qquad b = 1, \dots, B,$$

za $\theta \in \{\sigma,\ \text{medijana},\ \text{skew}\}$. Procjena standardne greške je

$$\widehat{\mathrm{SE}}_{\mathrm{boot}}(\theta) = \mathrm{sd}\Big(\theta^{*(1)}, \dots, \theta^{*(B)}\Big).$$

Prednosti: nema pretpostavki o obliku distribucije; hvata varijabilnost usljed punog resampliranja. **Mane:** računski intenzivan (ali lako paralelizovan).

3 Jackknife (leave-two-out): zašto traje duže

Metod i formula

U *leave-two-out* (L2O) Jackknife-u brišemo po dva podatka i računamo statistiku na preostalom skupu:

$$\theta_{(-i,-j)} = \theta\Big(m \setminus \{m_i, m_j\}\Big).$$

Egzaktno bi zahtijevalo $\binom{n}{2} = \mathcal{O}(n^2)$ reuzoraka ($\sim$ 50 miliona za $n = 10^4$) — neizvodivo. Zato koristimo **randomizovani L2O** sa tačno $K = 10{,}000$ nasumičnih parova (i, j). Procjene iz replikacija $\{\theta_k\}_{k=1}^{K}$:

$$\bar{\theta} = \frac{1}{K}\sum_{k=1}^{K}\theta_k, \qquad \widehat{\mathrm{SE}}_{\mathrm{L2O}}(\theta) = \sqrt{\frac{4}{K}\sum_{k=1}^{K}\Big(\theta_k - \bar{\theta}\Big)^2}.$$

Faktor 4 dolazi iz teorije skaliranja varijanse kod L2O Jackknife-a.

Zašto je sporiji?

- Egzaktni L2O je kvadratne složenosti po n i uključuje ogromno brisanje/kreiranje poduzoraka.
- I sa random $K = 10{,}000$ postoji overhead zbog *alokacija* i *kopiranja* nizova pri `np.delete`; ipak svodi se na par sekundi umjesto desetina minuta.

4 Numerički rezultati i značenje

Konačni brojevi:

Table 1: Osnovne statistike i procijenjene neodređenosti (Bootstrap i Jackknife L2O, $K = 10{,}000$).

Statistika	**Vrijednost**	**Bootstrap ±**	**Jackknife ±**
Standardna devijacija	7.066748	0.046463	0.001284
Medijana	26.517859	0.119159	0.000617
Asimetrija	-0.182896	0.023416	0.000627

Tumačenje. Medijana ima najmanju grešku (robustna na repove). Bootstrap daje veće SE jer resamplira *čitav* skup i osjetljiv je na repove; Jackknife je lokalniji, pa su SE značajno manje — to ne znači da su "bolje", već da procjenjuju drugačiji aspekt nesigurnosti.

5 Uticaj ekstremnih vrijednosti i pikovi

Dodavanje deset ekstremnih vrijednosti ($m = 1000\, M_\odot$)

- σ snažno raste (varijansa je vrlo osjetljiva na outliere),
- medijana se mijenja minimalno (robustna),
- asimetrija postaje pozitivna i veća (desni rep dominira).

Zaključak: i mali broj outliera *drastično* utiče na momente višeg reda.

Najviši pikovi u raspodjeli

Pikove detektujemo iz glatke procjene PDF-a (ukupni GMM) traženjem lokalnih maksimuma. Oni označavaju najvjerovatnije mase i tipično upućuju na *različite populacije* / kanale formiranja.

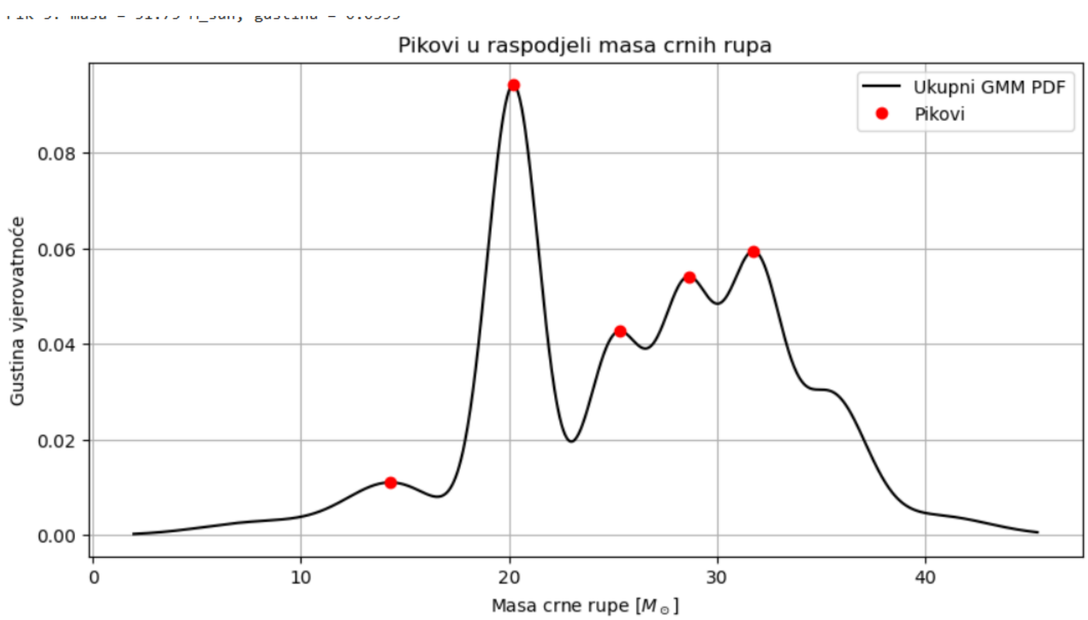

Figure 1: **Pikovi u raspodjeli masa crnih rupa.** Crna puna linija je ukupni GMM PDF; crvene tačke su lokalni maksimumi. Najviši pik oko $\sim 20\, M_\odot$ odgovara glavnoj populaciji; pikovi oko ~ 28–$32\, M_\odot$ su sekundarne grupe; mali pik oko ~ 14–$15\, M_\odot$ je slabija subpopulacija. Fizički, različiti pikovi impliciraju više kanala formiranja (različite progenitorske mase, spajanja, selekcione efekte).

6 Gaussian Mixture Model (GMM): izbor N, komponente i stabilnost

Pretpostavljamo mješavinu N Gaussovih komponenti

$$p(m) = \sum_{j=1}^{N} \alpha_j \mathcal{N}(m \mid \mu_j, \sigma_j^2), \qquad \sum_j \alpha_j = 1,\ \alpha_j \geq 0,$$

fitovanih EM algoritmom. Broj komponenti biramo minimizacijom **AIC**:

$$\mathrm{AIC} = 2k - 2\ln \hat{L}.$$

Izbor broja komponenti (AIC)

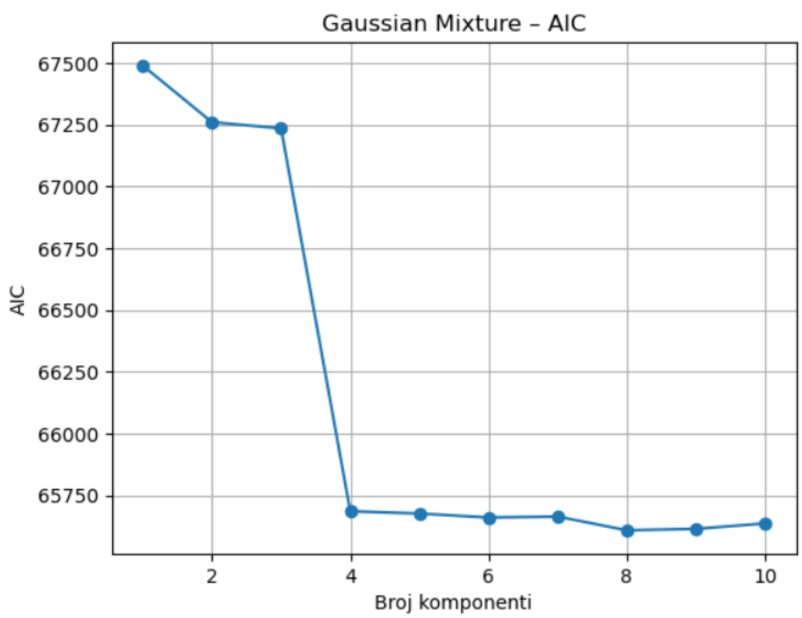

Figure 2: **Gaussian Mixture – AIC.** AIC opada kako N raste, potom se stabilizuje; minimum (ovdje oko $N \approx 8$) označava optimalni kompromis složenosti i prilagođenosti. Rani dramatični pad (npr. sa $N = 3 \to 4$) znači da dodatna komponenta hvata stvarni mod u podacima, dok kasniji mali dobici mogu biti “fino podešavanje”.

Komponente i ukupni PDF

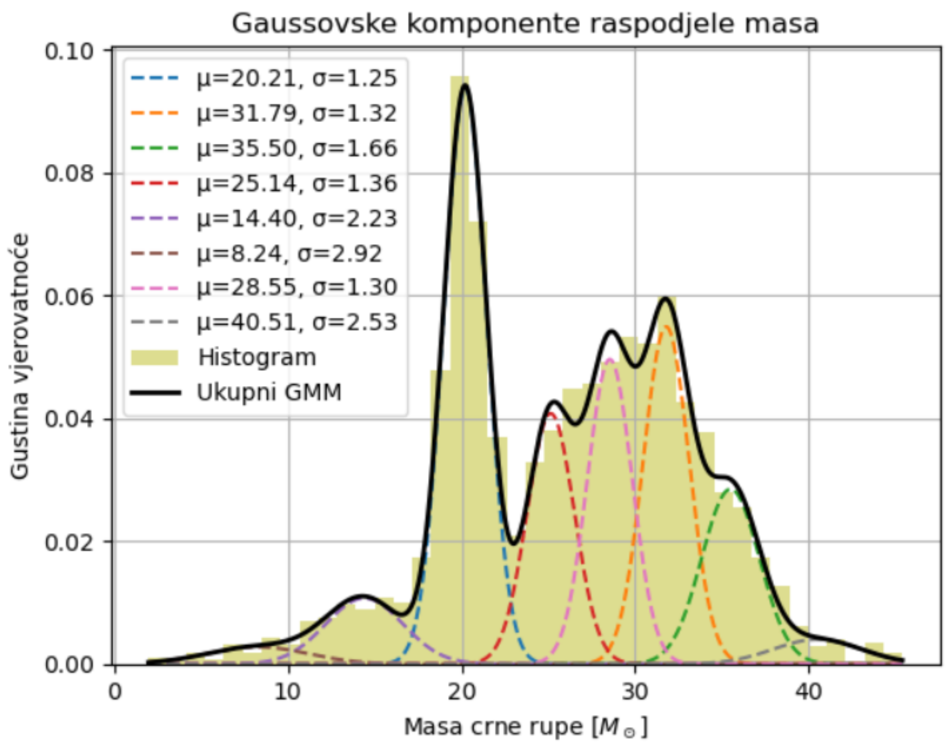

Figure 3: **Gaussovske komponente raspodjele masa.** Žuti histogram je normalizovan, isprekidane krive su pojedinačne komponente sa procijenjenim (μ_j, σ_j), crna puna je njihov zbir. U tvom fitu glavna uska komponenta je oko $\mu \approx 20\,M_\odot$, dok komponente kod $\sim$ 28–32 $M_\odot$ imaju veće σ i manje težine α_j.

Stabilnost parametara u zavisnosti od N

Stabilnost procjenjujemo kroz niz fitova za $N = 1, \ldots, 10$. Robusni modovi se *ponavljaju* kroz različite N sa sličnim (μ_j, σ_j); "novi" modovi koji se pojavljuju tek za velika N i imaju male α_j često su znak *overfitting*-a.

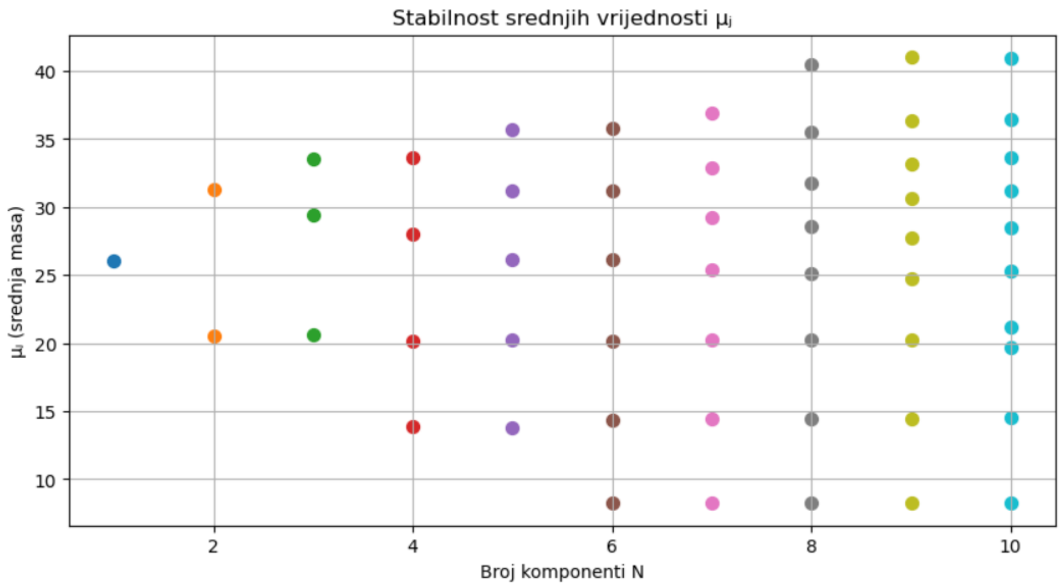

Figure 4: **Stabilnost srednjih vrijednosti μ_j.** Tačke su μ_j po komponentama za svako N. Grupa oko $\sim 20\,M_\odot$ se pojavljuje uporno (robustan glavni mod). Skupine oko ~ 28–$33\,M_\odot$ su sekundarni modovi. Veoma niski μ kod velikih N su obično artefakti *overfitting*-a (mini-komponente koje hvataju šum).

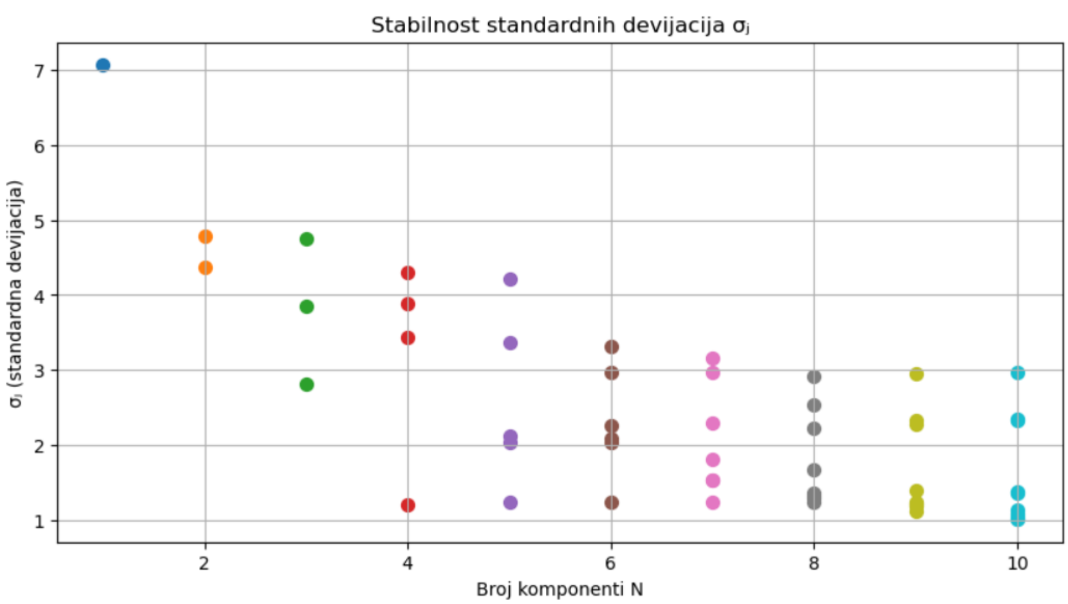

Figure 5: **Stabilnost standardnih devijacija σ_j.** Širina glavnog moda (oko $20\,M_\odot$) ostaje mala–umjerena, što znači oštar, dobro definisan pik. Sekundarni modovi imaju veće σ_j (šire repove), što je u skladu sa mješavinom različitih potpopulacija ili spajanja.

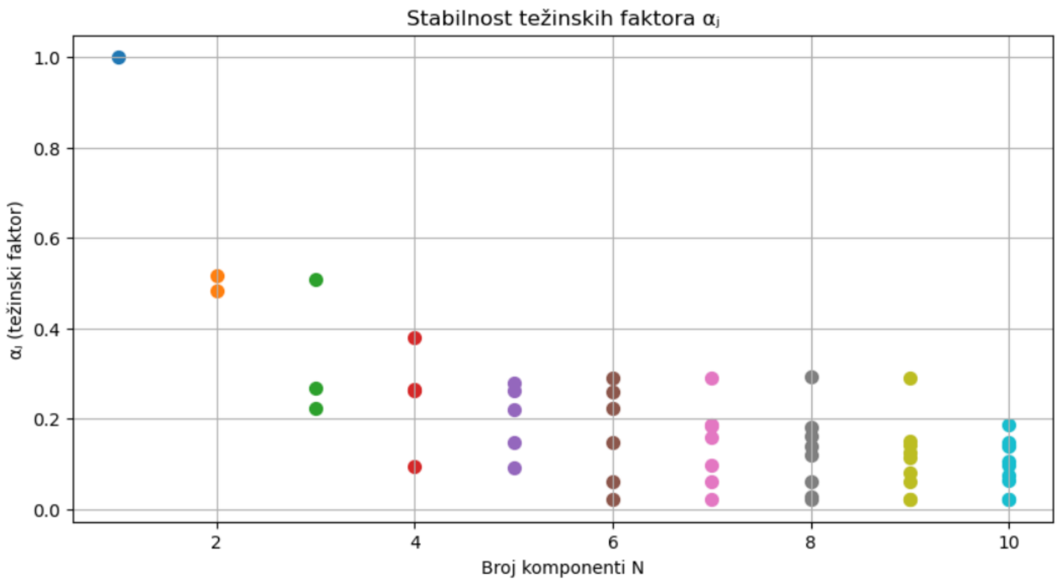

Figure 6: **Stabilnost težinskih faktora α_j.** Za mala N najveći dio mase nosi 1–2 komponente (dominantne populacije). Sa većim N, pojavljuju se komponente sa vrlo malim α_j — često *regularizovani* overfitting (hvatanje lokalnih fluktuacija bez jasnog fizičkog značenja).

Fizičko tumačenje.

- Stabilni pik oko $\sim 20\,M_\odot$: vjerovatno najčešći *kanal formiranja* (kolaps masivnih zvijezda u binarnim sistemima).
- Pikovi $\sim$28–32 $M_\odot$: potencijalno *spajanja* (mergers) ili drugačiji progenitorski scenariji; veće σ_j ukazuju na veću heterogenost.
- Slabi niski/visoki modovi koji se pojavljuju tek za $N \gg 4$ i imaju male α_j vjerovatno nisu zasebne populacije već statističke fluktuacije.

7 Zaključak

- **Bootstrap** i **Jackknife L2O** daju konzistentne poruke uz različite SE: Bootstrap hvata varijaciju kompletnog resampliranja i daje veće greške; Jackknife (random L2O) je lokalniji i brži uz K=10,000, ali sa manjim SE.
- **Ekstremne vrijednosti** snažno utiču na σ i *skewness*, dok je medijana stabilna — korisno za robustnu deskripciju distribucije masa.
- **GMM** detektuje višemodalnu strukturu: stabilni pikovi (naročito oko $\sim 20\,M_\odot$ i $\sim 30\,M_\odot$) ukazuju na više *kanala formiranja* i/ili post-procesne efekte (spajanja). AIC vodi ka razumnom N i izbjegava overfitting.