Domača naloga 5

Neža Kržan, Tom Rupnik Medjedovič

1 Cilj naloge

Imava dano pogojno porazdelitev oz. gostoto porazdelitve, iz katere ne moreva vzorčiti z običajnimi metodami. V ta namen bova uporabila algoritem Metropolis-Hastings, s pomočjo katerega bova vzorčila iz dane porazdelitve (z uporabo gostote). V najnem primeru je potrebno generirati koordinate točk, torej pare (x_i, y_i). Na začetku si izberemo neki začetni vrednosti x_0 in y_0, za kateri mora veljati, da je gostota večja od 0 (je možen izid). Nato na vsakem koraku s pomočjo gostote porazdelitve $g(X_p|X_i = x_i)$ predlagamo novo vrednost (x_p) pogojno na predhodnjo vrednost (x_i). Vendar pa še ne vemo ali predlagano vrednost (x_p) zares sprejmemo. Zato izračunamo verjetnost $\alpha = min\left(\frac{f(x_p)g(x_i|x_p)}{f(x_i)g(x_p|x_i)}, 1\right)$, ki nam pove verjetnost sprejema nove vrednosti, $(1-\alpha)$ pa verjetnost za ohranitev predhodnje na sledeč način

$$x_{i+1} = \begin{cases} x_p, \text{ z verjetnostjo } \alpha \\ x_i, \text{ z verjetnostjo } 1-\alpha \end{cases}$$

Nato bova še preverila kakšna je verjetnost, da velja $(x_i, y_i) \leq (1, 1)$, za vzorce velikosti 100 in izračunala pokritost 95% intervala zaupanja za to verjetnost.

2 Generiranje vrednosti

Z uporabo algoritma Metropolis-Hastings bova generirala vrednosti iz porazdelitve, ki ima gostoto proporcionalno

$$f(x,y) = \begin{cases} x^2y^2e^{-x}e^{-y}e^{-xy}, \text{ kjer } x > 0 \text{ in } y > 0 \\ 0, \text{ sicer} \end{cases}$$

Po že zgoraj opisanem postopku sledimo korakom algoritma. Pri tem sva za predlaganje novih vrednosti izbrala

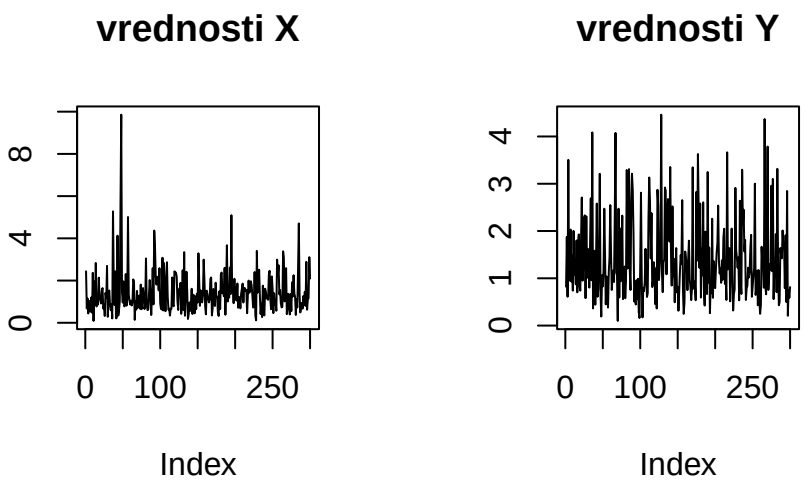
$$x_p = N(x_i, 1)$$
$$y_p = N(y_i, 1)$$

ter za gostoto porazdelitve

$$g((x_p, y_p)|(x_i, y_i)) = f_{N(x_i,1)}(x_p) \cdot f_{N(y_i,1)}(y_p),$$

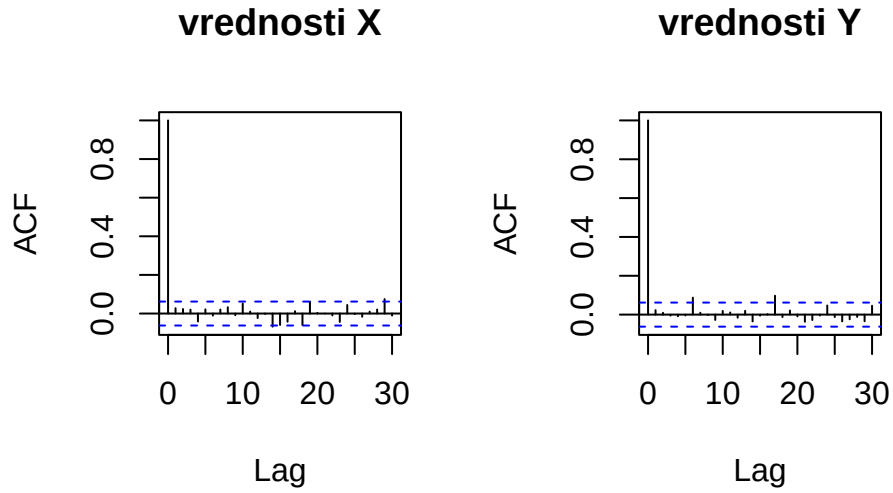
kjer sta $f_{N(x_i,1)}$ in $f_{N(x_i,1)}$ gostoti $X_i = x_i$ in $Y_i = y_i$ ter s standardnim odklonom 1.

Pri generiranju podatkov moramo nasatviti še vrednosti parametrov `burn in` in `step`. Parameter `burn in` nam določi koliko začetnih vrednosti izpustimo iz vzorca (jih ne vključimo). Vrednost tega sva nastavila na 100, saj vrednost dokaj hitro skonvergira. To lahko preverimo tudi grafično. Prikazala sva gibanje vrednosti za prvih 300 vrednosti iz vzorca.



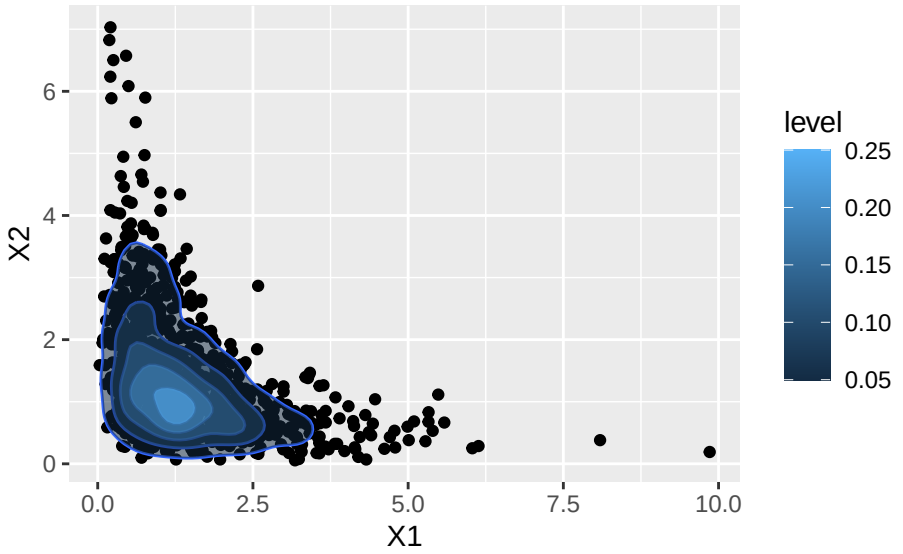
Res lahko vidimo, da se vrednosti gibljejo znotraj pričakovanega območja (ni prisotnega naraščanja oz. padanja).

Paramater `step` nam določi koliko zaporednih vrednosti ne vključimo v vzorec. S spreminjanjem te vrednosti želimo izločiti avtokorelacijo med zaporednimi elementi v vzorcu. To vrednost sva nastavila na 100, saj je bila v vzorcu prisotna visoka avtokorelacija. Tudi to lahko preverimo grafično z avtokorelogramom.



Res so skoraj vse vrednosti znotraj 95% intervala.

2.1 Prikaz vrednosti vzorca



Vidimo, da je večina vrednosti tako za X kot tudi Y zgočenih na intervalu $[0, 3]$. To je tudi nekako pričakovano, saj gostota porazdelitve nekoliko spominja na gostoto eksponentne porazdelitve in so temu primerno razporejene tudi točke.

3 Verjetnost

Želimo oceniti verjetnost, da sta obe vrednosti (X in Y) manjši od 1. Na dovolj velikem vzorcu lahko to vrednost ocenimo kot delež točk, ki se nahajajo znotraj območja $[0, 1] \times [0, 1]$. Ker je generiranje velikih vzorcev časnovno zahtevni proces, sva zgenerirala vzorec velikosti 1000000.

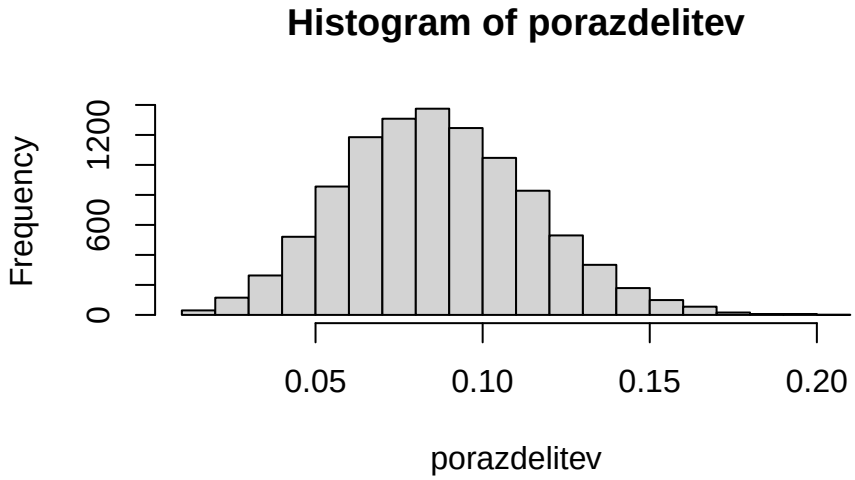
Na tem (velikem) vzorcu, sva izračunala željeni delež (obe vrednosti sta manjši od 1) in dobila vrednost 0.091. Ker se nam ta vrednost zdi dokaj majhna glede na zgornji graf vrednosti, si oglejmo naslednjo tabelo.

	Y<1	1<Y<2	Y>2
X<1	91475	162263	157342
1<X<2	162526	155607	54260
X>2	156955	54113	5459

Vidimo, da je res velik delež vrednosti, ko je vsaj ena od vrednosti (x ali y) nekoliko višja od 1.

3.1 Vzorci velikosti 100

Poglejmo si kakšna je porazdelitev zgornjega deleža v primeru, da imamo vzorce velikosti le 100. Generiramo veliko število vzorcev (npr. 10000) velikosti 100, na vsakem izračunamo verjetnost da sta tako X kot Y manjša od 1 in narišemo histogram.



Vidimo lahko da se vrednosti porazdeljujejo po normalni porazdelitvi. V našem primeru imata parametra vrednosti μ=0.0916 in σ=0.0285.

Da se prepričamo o pravilnosti vrednosti parametrov postopek ponovimo 10-krat in rezultate prikazemo v tabeli.

Vidimo, da so si vrednosti med seboj zelo podobne, zato smo s tem zadovoljni.

3.1.1 Pokritost

Izračunajmo še pokritost 95% intervala zaupanja za to vrednost. Interval zaupanja za povprečje porazdelitve oziroma verjetnosti, da sta tako X kot Y manjša od 1, bomo izračunali po običajni formuli $\left[\hat{\mu} - 1.96\frac{\hat{s}}{\sqrt{n}},\ \hat{\mu} + 1.96\frac{\hat{s}}{\sqrt{n}}\right]$. Vendar pa nastopi problem, saj ne poznamo "prave" vrednosti verjetnosti, saj ne poznamo parametrov porazdelitve oziroma vrednosti populacije. To lahko rešimo tako, da za "pravo" vrednost vzamemo verjetnosti delež izračunan na velikem vzorcu (v našem primeru ima 1000000 enot).

Prava vrednost deleža je enaka 0.09148.

V ta namen bomo naredili simulacijo kjer določimo število ponovitev (korakov simulacije) in števila vzorcev velikosti 100. V najnem primeru sva izbrala 1000 kot vrednost obeh parametrov. Po izvedeni simulaciji je vrednost pokritja 95% intervala zaupanja enaka 0.957

4 Zaključek