

Doc. dr. Safet Hamedović

Van. prof. dr. Hermina Alajbegović

ASIMPTOTSKA RASPODJELA MM PROCJENITELJA U JEDNOM MODELU SA ADITIVNOM GREŠKOM

Sažetak

U radu razmatramo simetrični model uniformne raspodjele sa aditivnom greškom. Uz egzistenciju dovoljnog broja momenata greške, izvedena je asimptotska raspodjela MM procjenitelja poluširine uniformne raspodjele. Eksplicitno je data i njena asimptotska varijansa.

Ključne riječi: uniformna raspodjela, aditivna greška, MM procjenitelj, asimptotska raspodjela

Uvod

Osnovni problem u statistici možemo opisati na sljedeći način: na osnovu opserviranih podataka (realizovanog uzorka) $\mathbf{x} = (x_1, x_2, \dots, x_n)$ želimo izvesti određene zaključke o mehanizmu koji generiše te podatke. Opservirani podaci su realizacija slučajnog vektora $\mathbf{X} = (X_1, X_2, \dots, X_n)$. Uobičajene pretpostavke (često opravdane u primjenama) su da su komponente $X_1, X_2, \dots, X_n$ slučajnog vektora $\mathbf{X}$ nezavisne i sa istom raspodjelom (*eng. independent and identically distributed* - i.i.d.) kao i slučajna promjenljiva X . Zadatak se onda svodi na procjenu zajedničke funkcije raspodjele F_X ili funkcije gustine f_X , ako je X neprekidna. Stvarna funkcija raspodjele F_X obično nije poznata, već za nju pretpostavljamo da pripada jednoj dopustivoj porodici raspodjela $\mathcal{F}$ koju zovemo statistički model. Ukoliko raspodjele iz modela imaju poznatu formu, određenu konačno dimenzionalnim parametrom $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, $k \in \mathbb{N}$, model je parametarski. Inače, modele čije raspodjele ne dopuštaju indeksiranje konačno dimenzionalnim parametrom zovemo neparametarskim modelima. Međutim, u primjenama nam ponekad nisu direktno dostupne opservacije promjenljive X . Jednu takvu situaciju opisujemo modelom sa aditivnom greškom

$$Y = X + \varepsilon,$$

gdje je X ciljna promjenljiva, dok promjenljivu ε tretiramo kao grešku u mjerenju. Za promjenljive X i ε pretpostavljamo da su nezavisne. Zadatak je procijeniti raspodjelu promjenljive X na osnovu i.i.d. uzorka $Y_1, Y_2, \dots, Y_n$ iz Y . Drugim riječima, umjesto $X_1, X_2, \dots, X_n$ iz X , raspolažemo samo kontaminiranim podacima $Y_1, Y_2, \dots, Y_n$ iz Y . Ovaj problem je poznat kao dekonvolucijski problem (vidi [10]). Potreba za modelima sa aditivnom greškom se javlja u raznim oblastima, kao što su: biologija, astronomija, ekonomija, javno zdravstvo, itd..

Klasične metode dekonvolucije su uglavnom neparametarske. Obično se pretpostavlja da su sabirci X i ε apsolutno neprekidni, pri čemu je gustina greške f_ε poznata. Među neparametarskim procjeniteljima gustine f_X , najpoznatiji je dekonvolucijski procjenitelj jezgrom (*eng. deconvolution kernel density estimator*), vidjeti [14]. Opis ove metode, kao i drugih neparametarskih metoda, možemo naći u [10].

Ukoliko je za ciljnu promjenljivu X predložen adekvatan parametarski model, procjenu odgovarajućih parametara (barem načelno) možemo obaviti klasičnim metodama: metodom maksimalne vjerodostojnosti (*eng. maximum likelihood* - ML) ili metodom momenata (*eng. method of moments* - MM). Možemo reći da smo na taj način izvršili parametarsku dekonvoluciju.

Primjere parametarskog pristupa nalazimo, recimo, u [13] (konvolucija Poisson-ove i binomne raspodjele), [8] (konvolucija normalne i lognormalne raspodjele), [1] i [12] (konvolucija uniformne i normalne raspodjele), itd..

U radu ćemo izvesti asimptotske osobine MM procjenitelja širine nosača uniformne raspodjele U u simetričnom modelu (detaljniji opis takvog modela ćemo dati u narednoj sekciji)

$$X = U + \varepsilon,$$

na osnovu i.i.d. uzorka iz X , pri čemu je greška ε apsolutno neprekidna i nezavisna od U . Uočimo da je ovdje U ciljna promjenljiva, koja je neopservabilna, te da je X opservabilna. Navedeni model sa normalnom greškom je analiziran u [1], dok je u [2] tretiran model sa Laplaceovom greškom. Modeli sa opštijim tipom grešaka su analizirani u [7]. Pokazalo se da bi ovakvi modeli mogli biti korisni u procjeni veličine objekata u zamagljenim slikama (vidjeti [3], [6], [7]).

Model

Model koji razmatramo je oblika

$$X = U + \sigma\varepsilon, \quad (1)$$

gdje slučajna promjenljiva U ima uniformnu raspodjelu na intervalu $[-a, a]$ (u oznaci $U \sim \mathcal{U}(-a, a)$), $a > 0$. Dakle, $f_U(x) = \frac{1}{2a} I_{[-a, a]}(x)$, gdje je I_A indikator funkcija skupa A . Greška ε je apsolutno neprekidna sa parnom funkcijom gustine f , te funkcijom raspodjele F . Uobičajeno su familije raspodjela kojima modeliramo grešku skalirane, pa smo radi udobnosti izvukli parametar skaliranja σ ($\sigma > 0$) ispred. Za parametar σ se pretpostavlja da je poznat. U ovakvom zapisu ćemo smatrati da je ε standardni predstavnik familije (parametar skaliranja u ε je jednak jedinici). Ako je varijansa od ε konačna, po potrebi ćemo uzimati standardizovane predstavnike ($\text{Var}(\varepsilon) = 1$). Za promjenljive U i ε ćemo pretpostaviti da su nezavisne. Zadatak je procijeniti parametar a , tj. poluširinu nosača uniformne, na osnovu i.i.d. uzorka iz X .

Korištenjem konvolucijske formule, gustina promjenljive X je

$$f(x; a, \sigma) = \int_{-\infty}^{\infty} f_U(t) f_{\sigma\varepsilon}(x - t) dt = \frac{1}{2a} \left(F\left(\frac{x+a}{\sigma}\right) - F\left(\frac{x-a}{\sigma}\right) \right).$$

Poznato je da model bez aditivne greške, $X \sim \mathcal{U}(-a, a)$, nije regularan jer tada nosač od X zavisi od nepoznatog parametra. Iako dodavanje aditivne greške komplikuje procjenu parametra a , ono nam omogućava proširenje nosača na čitav $\mathbb{R}$, tj. dobijanje regularnog modela.

Model (1) sa normalnom greškom (tj. $\varepsilon \sim \mathcal{N}(0, 1)$) i poznatim parametrom σ je analiziran u [1]. Tamo je izveden MM procjenitelj i pokazano je da je model regularan, tj. da je ML procjenitelj parametra a asimptotski efikasan. Simetrični model sa Laplaceovom greškom ($\varepsilon \sim \mathcal{L}(0, 1)$) je opisan u [2]. I ovdje je izveden MM procjenitelj, te utvrđena regularnost modela. U oba slučaja je, prema očekivanju, potvrđeno da je ML procjenitelj efikasniji od MM procjenitelja. Za takve modele se ispostavilo da mogu biti korisni u procjeni dimenzija objekata u nepreciznim slikama, kao i u procjeni granica kružnih ili elipsastih objekata ([2], [3], [6], [7]). U te svrhe je razvijen i R-paket „LeArEst“ ([4]) koji je dostupan na CRAN repozitoriju ([5], [11]). Između ostalog, paket omogućava učitavanje slike i procjenu veličine ciljnog objekta na temelju teorijskih rezultata u modelu uniformne sa aditivnom greškom.

U [7] je tretiran model sa opštijom formom greške. Namjera je bila dati prihvatljive uslove regularnosti u koje bi se uklopili razni tipovi simetričnih grešaka. Ispostavilo se da te uslove zadovoljavaju i neke raspodjele sa težim repovima, poput Studentovih raspodjela sa nižim brojevima stepeni slobode. Takve raspodjele su interesantne sa stanovišta robusnosti (tj.

„otpornosti“ na veći broj stršćih vrijednosti, tzv. outliers), što može biti korisno u pomenutim primjenama (procjena veličine objekata u zamagljenim slikama).

MM procjenitelj i njegove osobine

Nepoznati parametar a u modelu (1) možemo procijeniti metodom momenata, ukoliko postoji dovoljan broj momenata promjenljive ε . Iako je, generalno, ML procjenitelj efikasniji od MM procjenitelja, MM procjenitelji su najčešće dati u zatvorenoj formi što omogućuje jednostavno dobijanje početnih tačaka u optimizacijskim procedurama za dobijanje ML procjenitelja.

Pretpostavimo da u modelu (1) greška ε ima konačan moment 4. reda, te da je ε standardizovani predstavnik (svoje) familije. Imajući u vidu i simetričnost greške, tada je

$$E\varepsilon = 0, \quad E\varepsilon^2 = 1 \quad \text{ i } \quad \kappa_\varepsilon = \frac{E(\varepsilon - E\varepsilon)^4}{(E(\varepsilon - E\varepsilon)^2)^2} = E\varepsilon^4$$

gdje je κ_ε tzv. kurtosis¹ promjenljive ε . Privremeno tretiranje modela sa standardizovanim predstavnikom greške olakšava izvođenje MM procjenitelja. MM procjenitelj u modelu sa standardnim predstavnikom greške će se onda dobiti jednostavnom modifikacijom (reskaliranjem). Osim toga, kurtosis je invarijantan na skaliranje raspodjele². S obzirom na nezavisnost promjenljivih U i ε , te činjenice da je

$$EU^k = \begin{cases} \frac{a^k}{k+1}, & \text{ ako je } k \text{ paran} \\ 0, & \text{ ako je } k \text{ neparan} \end{cases},$$

lako dobijamo momente drugog i četvrtog reda promjenljive X :

$$EX^2 = E(U + \sigma\varepsilon)^2 = \frac{a^2}{3} + \sigma^2 \quad \text{ i } \quad EX^4 = E(U + \sigma\varepsilon)^4 = \frac{a^4}{5} + 2a^2\sigma^2 + \sigma^4\kappa_\varepsilon$$

Ako je parametar σ poznat, MM procjenitelj parametra a dobijamo izjednačavanjem teorijskog i uzoračkog momenta drugog reda $m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$

$$\hat{a}_{MM} = \sqrt{3(m_2 - \sigma^2)}.$$

¹ Kurtosis se u našoj literaturi uglavnom pogrešno interpretira kao mjera šiljatosti ili spljoštenosti raspodjele. Otud i naziv koeficijent spljoštenosti. Međutim, kurtosis je prvenstveno određen ponašanjem repova (vidi [15]), pa bi adekvatniji naziv bio koeficijent težine repova.

² Napomenimo da je u [7] dat osvrt na mogućnost korištenja kurtosisa, kao mjere težine repova, u odabiru potencijalnog modela greške. Međutim, analiza je pokazala da κ_X zavisi ne samo od κ_ε , već i od količnika a/σ .

Ukoliko je u modelu (1) greška ε data u standardnom obliku, pri čemu je $\text{Var}(\varepsilon) = \sigma_0^2$, MM procjenitelj parametra a je onda

$$\hat{a}_{MM} = \sqrt{3(m_2 - \sigma_0^2 \sigma^2)} \quad (2)$$

Sada prelazimo na glavni rezultat rada - izvođenje asimptotske raspodjele MM procjenitelja. U tu svrhu koristimo proceduru poznatu kao Δ -metod.

Teorema 1 (Δ -metod, [9], Teorema 5.2.9, p. 335)) Neka je $\{X_n\}$ niz slučajnih promjenljivih takav da važi

$$\sqrt{n}(X_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

Pretpostavimo da je funkcija $g(x)$ diferencijabilna u θ i da važi $g'(\theta) \neq 0$. Tada je

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2(g'(\theta))^2\right).$$

Teorema 2 (asimptotska raspodjela MM procjenitelja) Neka je $(X_1, X_2, \dots, X_n)$ i.i.d. uzorak iz promjenljive X date sa (1), $E\varepsilon^4 < \infty$ i $E\varepsilon^2 = \sigma_0^2$. Asimptotska raspodjela MM procjenitelja (2) je data sa

$$\sqrt{n}(\hat{a}_{MM} - a) \xrightarrow{D} \mathcal{N}(0, V_{MM}),$$

gdje je

$$V_{MM} = \frac{a^2}{5} + 3\sigma^2\sigma_0^2 + \frac{9(\kappa_\varepsilon - 1)\sigma^4\sigma_0^4}{4a^2}.$$

Dokaz. Prema centralnoj graničnoj teoremi, asimptotska raspodjela uzoračkog momenta drugog reda $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ je

$$\sqrt{n}(m_2 - \mu_2) \xrightarrow{D} \mathcal{N}(0, \mu_4 - (\mu_2)^2),$$

gdje je

$$\mu_2 = EX^2 = \frac{a^2}{3} + \sigma^2\sigma_0^2 \quad \text{ i } \quad \mu_4 = EX^4 = \frac{a^4}{5} + 2a^2\sigma^2\sigma_0^2 + \sigma^4\sigma_0^4\kappa_\varepsilon.$$

Procjenitelj $\hat{a}_{MM}$ je funkcija statistike m_2 :

$$\hat{a}_{MM} = \sqrt{3(m_2 - \sigma_0^2\sigma^2)} = g(m_2).$$

Funkcija $g(x) = \sqrt{3(x - \sigma_0^2\sigma^2)}$ je diferencijabilna i važi $g(\mu_2) = \frac{3}{2\sqrt{a^2}}$. Primjenom Δ -metoda, imamo

$$\sqrt{n}(g(m_2) - g(\mu_2)) \xrightarrow{D} \mathcal{N}\left(0, (\mu_4 - (\mu_2)^2)(g'(\mu_2))^2\right),$$

odnosno

$$\sqrt{n}(\hat{a}_{MM} - a) \xrightarrow{D} \mathcal{N}\left(0, \frac{a^2}{5} + 3\sigma^2\sigma_0^2 + \frac{9(\kappa_\varepsilon - 1)\sigma^4\sigma_0^4}{4a^2}\right). \quad \blacksquare$$

Ukoliko je i parametar σ takođe nepoznat, moguće je naći MM procjenitelje oba parametra, a i σ . Dobijamo ih izjednačavanjem momenata drugog i četvrtog reda sa odgovarajućim uzoračkim momentima³

$$m_2 = \frac{a^2}{3} + \sigma^2,$$

$$m_4 = \frac{a^4}{5} + 2a^2\sigma^2 + \sigma^4\kappa_\varepsilon.$$

Gornji sistem generalno ima 8 rješenja u $\mathbb{C}^2$ (računajući i višestrukost). Znajući prirodu parametara a i σ , zanimaju nas samo realna i pozitivna rješenja. U [7] su navedena rješenja sistema, tj. procjenitelji $\hat{a}_{MM}$ i $\hat{\sigma}_{MM}$, u zavisnosti od κ_ε . U pomenutom radu nije diskutovan slučaj $1 < \kappa_\varepsilon < 3$. Ispostavlja se da tada sistem može imati dva para rješenja. Međutim, sa stanovišta primjene to i nije problem imajući u vidu da se tada radi o raspodjelama sa jako tankim repovima (kurtosis normalne raspodjele je jednak 3).

Zaključak

Model uniformne raspodjele sa aditivnom greškom može biti koristan u različitim primjenama. U ranijim istraživanjima je uglavnom korištena normalna greška. Nedavno je uočena i moguća primjena takvih modela u procjeni dimenzija objekata u nepreciznim slikama. U tu svrhu analizirane su još neke familije grešaka. U radu je dokazana asimptotska normalnost MM procjenitelja poluširine uniformne raspodjele, te je eksplicitno data njegova asimptotska varijansa.

Literatura

- [1] Benšić, M., Sabo, K. (2007). Estimating the width of a uniform distribution when data are measured with additive normal errors with known variance, *Computational Statistics and Data Analysis*, 51, 4731-4741.
- [2] Benšić, M., Sabo, K. (2016). Uniform distribution width estimation from data observed with Laplace additive error, *Journal of the Korean Statistical Society* 45, 505-517.

³ I ovdje je udobnije u računanju koristiti model sa standardizovanom greškom. $\hat{\sigma}_{MM}$ u modelu sa standardnom greškom dobijamo reskaliranjem sa faktorom $1/\sigma_0$.

- [3] Benšić, M., Taler, P., Hamedović, S., Nyarko E.K. and Sabo, K. (2017). LeArEst: The Length and Area Estimation from Data Measured with Additive Error, *R Journal* 9:2, 461-473.
- [4] Benšić, M., Hamedović, S., Sabo, K. and Taler, P. (2018). LeArEst: Border and Area Estimation of Data Measured with Additive Error, R package version 1.0.0, <https://cran.r-project.org/package=LeArEst>
- [5] The Comprehensive R Archive Network, <https://cloud.r-project.org/>.
- [6] Hamedović, S., Benšić, M., Sabo, K., Taler, P. (2018). Estimating the size of an object captured with error, *CEJOR*, 26, 771-781.
- [7] Hamedović, S., Benšić, M., Sabo, K. (2020). Estimating the width of a uniform distribution under symmetric measurement errors, *Journal of the Korean Statistical Society*, 49, 822-840.
- [8] Hawkins, D., 1991. The convolution of the normal and lognormal distributions. *Journal of South African Statistics*, 25, 99-128.
- [9] Hogg, R.V., McKean, J.W., Craig, A.T., (2019). *Introduction to mathematical statistics (8th Ed)*. Pearson, Boston.
- [10] Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*, Springer-Verlag, New York.
- [11] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [12] Schneeweiss, H. (2004). Estimating the endpoint of a uniform distribution under measurement errors. *CEJOR* 12, 221-231.
- [13] Sprott, D.A. (1983), Estimating the Parameters of a Convolution by Maximum Likelihood, *Journal of the American Statistical Association*, 78, 457-460.
- [14] Stefanski, L., Carroll, R.J. (1990). Deconvoluting kernel density estimators, *Statistics*, 21, 169-184.
- [15] Westfall, P.H. (2014). Kurtosis as Peakedness, 1905 - 2014. R.I.P.. *Am Stat.* 68 (3): 191-195.

THE ASYMPTOTIC DISTRIBUTION OF THE MM ESTIMATOR IN AN ADDITIVE ERROR MODEL

Abstract

In the paper, we consider symmetric uniform distribution model with an additive error. The asymptotic distribution of the MM estimator of the half-width of the uniform distribution is derived, assuming the existence of sufficient moments of the error term. Its asymptotic variance is also given explicitly.

Keywords: *uniform distribution, additive error, MM estimator, asymptotic distribution*