Лабораторна робота №4
з дисципліни ”комп’ютерна статистика”
Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп'ютерна статистика"

23 листопада 2020 р.

1 Вступ.

У даній роботі вказана інформація про отримані результати під час виконання роботи №4: знайдена оцінка методу максимальної вірогідності для невідомої дисперсії однієї з компонент суміші двох нормальних розподілів. Ця оцінка була реалізована в середовищі мови програмування R. Перевірено умови регулярності та консистентності для функції вірогідності та ОМВ відповідно. В кінці наведені дані про ефективність ОМВ та моментної оцінки, отриманої в минулій лабораторній роботі.

2 Функція вірогідності. Умови регулярності.

Нагадаємо щільність двокомпонентної гауссової суміші:

$$\psi(t) = \frac{p}{\theta}\varphi\left(\frac{t-\mu_1}{\theta}\right) + \frac{1-p}{\sigma}\varphi\left(\frac{t-\mu_2}{\sigma}\right),\ p \in (0,1),\ \mu_j \in \mathbb{R},\ \theta, \sigma > 0,$$

де φ - щільність стандартного нормального розподілу. Тоді функція вірогідності одноелементної вибірки ватиме вигляд:

$$L(\xi_1, \theta) = \frac{p}{\theta}\varphi\left(\frac{\xi_1-\mu_1}{\theta}\right) + \frac{1-p}{\sigma}\varphi\left(\frac{\xi_1-\mu_2}{\sigma}\right)$$

Перевіримо виконання умов регулярності на задану функцію вірогідності. Зауважимо, що для будь-якого $\theta \in \Theta = (0, \infty)$ функція вірогідності $L(\xi_1, \theta) > 0$, оскільки $\varphi(\cdot) > 0$. Звідси неважко побачити, що за θ функція $1/L(\xi_1, \theta)$ буде неперервною. Розглянемо похідну по параметру:

$$\frac{\partial}{\partial\theta}L(\xi_1, \theta) = \frac{\partial}{\partial\theta}\left(\frac{p}{\theta}\varphi\left(\frac{\xi_1-\mu_1}{\theta}\right)\right) = p\varphi\left(\frac{\xi_1-\mu_1}{\theta}\right)\left(-\frac{1}{\theta^2} + \frac{1}{\theta^4}(\xi_1-\mu_1)^2\right)$$

Функція впливу $U(\xi_1, \theta) = \left(\frac{\partial}{\partial\theta}L(\xi_1, \theta)\right)/L(\xi_1, \theta)$ ненульова та інтегровна в квадраті. За теоремою про диференціювання під знаком інтеграла Лебега можна довести останню умову регулярності.

Оскільки умови регулярності виконуються, тому математичне сподівання від функції впливу $U(\xi_1, \theta)$ дорівнює нулю. Тому інформація за Фішером за одним спостереженням можна обчислити як другий теоретичний момент величини $U(\xi_1, \theta)$, тобто

$$i(\theta) = \mathbb{M}\left[U(\xi_1, \theta)\right] = \int\limits_{\mathbb{R}} U^2(t, \theta)\psi(t)dt = \int\limits_{\mathbb{R}} \left(p\varphi\left(\frac{t-\mu_1}{\theta}\right)\left(-\frac{1}{\theta^2} + \frac{1}{\theta^4}(t-\mu_1)^2\right)\right)^2 \Big/ \psi(t)dt$$

Інтеграл у правій частині останнього виразу обчислимо за допомогою чисельних методів в R. Застосуємо функцю *integrate*, підставивши конкретні значення параметрів (як і в третій роботі, ми покладемо $\theta = 0.05$, $\mu_1 = 1$, $\mu_2 = 0$, $\sigma = 0.75$, $p = 0.6$).

```
given.mu.1 <- 1
given.mu.2 <- 0
given.sigma.2 <- 0.75
given.p <- 0.6
true.theta <- 0.05

# (U(x, theta))^2 * psi(x), x in R
h <- function(t, m.1, m.2, s.1, s.2, p)
{
  # Похідна за невідомим параметром від функції вірогідності одного спостереження
  u <- exp(-0.5 * (t - m.1)^2/s.1^2)
  g <- p/sqrt(2*pi) * u * ((t - m.1)^2 * s.1^(-4) - s.1^(-2))
  # Щільність гауссової суміші
  l <- d.gaussmixt(s.1, s.2, m.1, m.2, p, t)
  g^2/l
}

# Інформація за Фішером для одного спостереження
# Інтеграл береться по скінченному проміжку, який
# містить носій функції (U(x, theta))^2 * psi(x)
fisher.info <- function(m.1, m.2, s.1, s.2, p, a = -2, b = -a)
{
  integration.result <- integrate(
    function(x)
    {
      h(x, m.1, m.2, s.1, s.2, p)
    },
    a, b
  )
  integration.result
}

# Практичне застосування
fisher.info(given.mu.1, given.mu.2, true.theta, given.sigma.2, given.p)$value
# 353.5854
```

Після перевірки умов регулярності знайденої функції вірогідності, переходимо до ключового моменту – знаходження оцінки невідомої дисперсії компоненти суміші.

3 Знаходження ОМВ.

Відомо, що знаходження ОМВ для невідомих параметрів гауссового розподілу не викликало труднощів внаслідок простої явної форми логарифма функції вірогідності. У випадку оцінювання за ОМВ невідомих дисперсій компонент суміші, ситуація ускладнюється неможливістю виразити оцінку параметра θ аналітично, дивлячись на вигляд функції вірогідності вибірки довільного обсягу:

$$L(X,\theta) = \prod_{j=1}^{n} \left(\frac{p}{\theta}\varphi\left(\frac{\xi_j - \mu_1}{\theta}\right) + \frac{1-p}{\sigma}\varphi\left(\frac{\xi_j - \mu_2}{\sigma}\right)\right)$$

Слушною думкою буде застосування чисельних методів оптимізації функціоналу. У рамках цієї роботи, формулюється наступна екстремальна задача:

$$\ln L(X,\theta) = \sum_{j=1}^{n} \ln \left(\frac{p}{\theta}\varphi\left(\frac{\xi_j - \mu_1}{\theta}\right) + \frac{1-p}{\sigma}\varphi\left(\frac{\xi_j - \mu_2}{\sigma}\right)\right) \to \max_{\theta > 0}$$

Для її розв'язання застосуємо модифікований метод Ньютона для оптимізації. реалізований в функції nlm, тому нам залишається використати її:

```
# Реалізація моментної оцінки, отриманої в минулій самостійній роботі
est.mm <- function(x, m.1, m.2, s.2, p)
{
  est.r <- mean(x^2)/p - (1-p)/p * (s.2^2 + m.2^2) - m.1^2
  sqrt(abs(est.r))
}
# Щільність гауссової суміші
d.gaussmixt <- function(s.1, s.2, m.1, m.2, p, x)
{
  p*dnorm(x, m.1, s.1) + (1-p)*dnorm(x, m.2, s.2)
}
# Логарифмічна функція вірогідності вибірки
ll.mxt <- function(s.1, s.2, m.1, m.2, p, x)
{
  sum(log(d.gaussmixt(s.1, s.2, m.1, m.2, p, x)))
}
# Реалізація оцінки максимальної вірогідності
est.ml <- function(x, m.1, m.2, s.2, p, x.0 = est.mm(x, m.1, m.2, s.2, p))
{
  calc <- nlm(
    function(t){
      -ll.mxt(t, s.2, m.1, m.2, p, x) # ll.mxt -> max <=> ll.mxt -> min
    }, x.0 # - це початкове набляження розв'язку, беремо в якості
          # такого значення оцінки методу моментів за вибіркою
    )
  calc$estimate
}
```

Позначимо через $\hat{\theta}_{MLE,n}$ розв'язок оптимізаційної задачі, який буде шуканою оцінкою найбільшої вірогідності.

Спробуємо оцінити невідомий параметр на вибірці з 20 спостережень:

$$X = \{1.00783442,\ 1.05639144,\ 0.59559056,\ -1.48290781,\ -0.47177821,\ 0.99357730,\ 0.15523127,\\ -0.31360251,\ 0.99375741,\ -0.17947315,\ 0.95849106,\ -0.13911929,\ 0.87597920,\ -0.08290412,\\ 1.09222355,\ 0.98143038,\ 1.04635710,\ 0.98137625,\ -0.40099368,\ 1.03208599\}$$

За цією вибіркою отримали $\hat{\theta}_{MLE,20} \approx 0.05115761$. На думку автора звіту, значення досить хороше, зважаючи на невелику кількість спостережень. Далі перевіримо умови консистентності, щоб робити певні висновки про властивості отриманої оцінки.

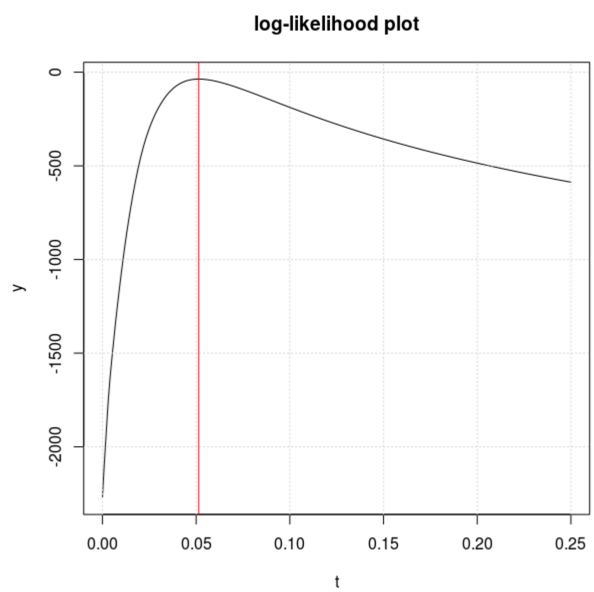

Рис. 1: Графік логарифмічної функції вірогідності за вибіркою X. Вертикальна пряма побудована з точки $\hat{\theta}_{MLE,20}$.

4 Властивості ОМВ.

Раніше було показано, що для функції вірогідності однієї вибірки $L(\xi_1, \theta)$ виконуються умови регулярності. Знову розглянемо похідну $U(\xi_1, \theta) = \frac{\partial}{\partial \theta} L(\xi_1, \theta)$. Далі, $\forall \theta > 0$:

$$\mathbb{M}\left[|U(\xi_1, \theta)|\right] = \int\limits_{\mathbb{R}} \left|\frac{\partial}{\partial \theta}\psi_\theta(t)\right| dt = \int\limits_{\mathbb{R}} \left| p\varphi\left(\frac{t-\mu_1}{\theta}\right)\left(-\frac{1}{\theta^2} + \frac{1}{\theta^4}(t-\mu_1)^2\right)\right| dt \leqslant$$

$$\leqslant p\left(\frac{1}{\theta^2}\int\limits_{\mathbb{R}} \varphi\left(\frac{t-\mu_1}{\theta}\right) dt + \frac{1}{\theta^4}\int\limits_{\mathbb{R}} \varphi\left(\frac{t-\mu_1}{\theta}\right)(t-\mu_1)^2 dt\right) =$$

$$= p\left(\frac{1}{\theta} + \frac{1}{\theta}\right) = \frac{2p}{\theta} < \infty$$

Тому за наслідком теореми 2 про консистентність ОМВ (Боровков А.А., "Математическая статистика", параграф №16) шукана оцінка максимальної вірогідності буде строго консистентною оцінкою невідомого параметра компоненти суміші.

Оцінка є асимптотично нормальною з граничною дисперсією рівній $V_{MLE} = 1/i(\theta) \approx 0.002828171$. Якщо побудувати великий масив таких оцінок, тоді можна спостерігати збіжність до відповідного розподілу та граничних числових характеристик. При $n = 1000$ маємо:

```
I.d <- fisher.info(given.mu.1, given.mu.2, true.theta, given.sigma.2, given.p)
I <- I.d$value
V <- 1/I

n <- 1000
B <- 1000

# generate.estimates береться з минулої самостійної роботи, замінивши ОММ на ОМВ
UU <- generate.estimates(
true.theta, given.mu.1, given.mu.2, given.sigma.2, given.p, n, B
)

# Зміщення
sqrt(n)*(mean(UU) - true.theta)
# -0.001578291

# Вибіркова дисперсія
n*var(UU)
# 0.002980108
```

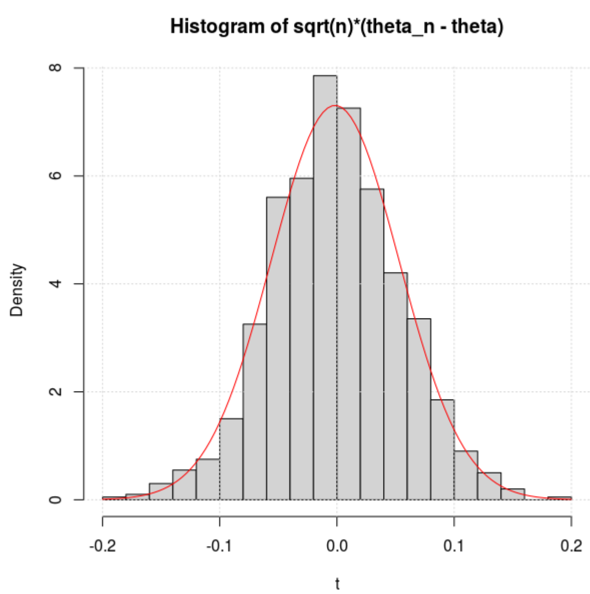

Рис. 2: Гістограма оцінок після відповідного нормування.

Нагадаємо моментну оцінку з минулої роботи та асимптотичну дисперсію для неї:

$$\hat{\theta}_{MM,n} = \sqrt{\left|\frac{5}{3} \cdot \hat{\mu}_{2,n} - \frac{11}{8}\right|}, h(t) = t^2, H(\theta) = p(\mu_1^2 + \theta^2) + (1-p)(\mu_2^2 + \sigma^2) \Rightarrow H'(\theta) = 2p\theta, \theta > 0$$

$$V_{MM}(\theta) = \frac{\mathbb{D}\left[\zeta^2\right]}{(H(\theta)')^2} = \frac{p(3\theta^4 + 6\theta^2\mu_1^2 + \mu_1^4) + q(3\sigma^4 + 6\sigma^2\mu_2^2 + \mu_2^4) - \left(p\left(\mu_1^2 + \theta^2\right) + q\left(\mu_2^2 + \sigma^2\right)\right)^2}{4p^2\theta^2}$$

$$\Rightarrow V_{MM}(0.05) \approx 84.8879$$

Тому відносна асимптотична ефективність оцінок УММ та ОМВ дорівнює

$$\frac{V_{MM}}{V_{MLE}} \approx 30015.13$$

Звідси маємо наступну інтерпретацію. Для забезпечення такої ж точності моментної оцінки, яку маємо за оцінкою найбільшої вірогідності, обсяг вибірки необхідно збільшити у 30000 разів.

5 Висновки.

Побудована оцінка найбільшої вірогідності невідомої дисперсії однієї з двох компонент гаусової суміші та досліджені її властивості. Показали, що така оцінка є більш ефективною, ніж момента оцінка, яка отримана у самостійній роботі №3.