Лабораторна робота №3
з непараметричної статистики

Горбунов Даніел Денисович
1 курс магістратури
група "Прикладна та теоретична статистика"
Варіант №4

22 березня 2022 р.

Вступ.

У даній роботі вказані результати підбору оптимальної ядерної оцінки вигляду:

$$\hat{f}_n(t) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{t-\xi_j}{h}\right) \tag{1}$$

для невідомої щільності розподілу f, що відповідає моделі двокомпонентної гауссової суміші з середніми $\mu_1 = -1$, $\mu_2 = 1$, одиничними дисперсіями та ймовірністю змішування $p = 1/2$. У формулі (1) фігурують: $h > 0$ – параметр згладжування, ξ_j – j-те спостереження кратної вибірки обсягу n: $\xi = (\xi_1, \ldots, \xi_n)$, $K(t) = \frac{3}{4}(1-t^2)\mathbb{1}\{|t| < 1\}$ – ядро Єпанечнікова. Позначимо $\mathrm{ISE}(\hat{f}_n) = ||f - \hat{f}_n||^2_{L_2(\mathbb{R})}$, $\mathrm{MISE}(\hat{f}_n) = \mathbb{E}\left[\mathrm{ISE}(\hat{f}_n)\right]$, $\mathrm{aMISE}(\hat{f}_n) = \frac{1}{4}\varphi D^2 H^4 + \frac{d^2}{H}$ – головна частина MISE при $h = Hn^{-1/5}$. Визначення h здійснено трьома підходами:

1. **Правило Сільвермана.** Припускаючи, що щільність є гауссовою, то оптимальний параметр згладжування h адаптивного методу визначається за правилом:

$$h_{silv} = \left(\frac{d^2 8\sqrt{\pi}}{3nD^2}\right)^{1/5}\hat{S},$$

   де $d^2 = \int\limits_{-\infty}^{\infty} K^2(t)dt$, $D = \int\limits_{-\infty}^{\infty} t^2 K(t)dt$. Якщо покласти $\hat{S} := \sqrt{S_0^2(\xi)}$, де $S_0^2(\xi)$ – виправлена вибіркова дисперсія за вибіркою ξ, маємо просте правило Сільвермана. Інакше, якщо взяти $\hat{S} = \min(\sqrt{S_0^2(\xi)}, \mathrm{IQR}(\xi)/1.34)$, де $\mathrm{IQR}(\xi)$ – інтерквартильний розмах за вибіркою ξ, отримаємо поліпшене правило Сільвермана.

2. **Непараметричний підхід.** Використовуємо адаптивний метод без додаткових припущень про відповідність щільності деякій моделі розподілу. Спочатку береться пілотна ядерна оцінка щільності $\tilde{f}_n$ за деяким параметром згладжування. За цією оцінкою, обчислюється оцінка для φ:

$$\hat{\varphi} = \int\limits_{-\infty}^{\infty} (\tilde{f}_n''(t))^2 dt$$

Використовуючи $\hat{\varphi}$, можна оцінити параметр згладжування, що мінімізує головну частину MISE:

$$\hat{h} = \left(\frac{d^2}{nD^2\hat{\varphi}}\right)^{1/5}$$

Остаточно будується ядерна оцінка $\hat{f}_n$ з параметром згладжування $\hat{h}$.

3. **Техніка кросс-валідації.** Параметр згладжування обирається як точка мінімуму функціонала

$$\mathrm{CV}(h) = \int_{-\infty}^{\infty} (\hat{f}_n(t))^2 dt - 2\hat{J}(h),\ \hat{J}(h) = \frac{1}{n}\sum_{j=1}^{n} \hat{f}_n^{j-}(\xi_j),$$

де $\hat{f}_n(x) = \hat{f}_n(x,h)$, $\hat{f}_n^{j-}$ – ядерна оцінка щільності, побудована за спостереженнями з ξ без j-го спостереження.

Підготовча робота.

Інтеграли в один рядок.

Обчислимо d^2, D, бо потрібно переконатися, що ці величини є скінченними, інакше використати вищевказану теорію буде неможливо.

$$d^2 = \int_{-\infty}^{\infty} K^2(t)dt = \frac{9}{16}\int_{-1}^{1}(1-t^2)^2 dt = \frac{9}{8}\int_{0}^{1}(1-t^2)^2 dt = \frac{9}{8}\int_{0}^{1}(1-2t^2+t^4)dt = \frac{9}{8}\left(1-\frac{2}{3}+\frac{1}{5}\right) = \frac{3}{5} < +\infty$$

$$D = \int_{-\infty}^{\infty} t^2K(t)dt = \frac{3}{4}\int_{-1}^{1} t^2(1-t^2)dt = \frac{3}{2}\left(\frac{1}{3}-\frac{1}{5}\right) = \frac{1}{5} < +\infty$$

Величини справді скінченні – теорія застосовна до нашого випадку.

Функція підрахунку ядерної оцінки щільності.

Функція, що повертає іншу функцію для підрахунку значень ядерної оцінки щільності:

```
dens.estim <- function(x, h, K)
{
  n <- length(x)
  g <- function(t) { sum(K((t - x) / h)) / (n * h) }
  gv <- function(t) { sapply(t, g) }; gv
}
```

Програмна реалізація підрахунку ядра Єпанєчнікова:

```
epan.kernel <- function(t) { 3/4 * (1 - t^2) * (abs(t) < 1) }

# Ті самі значення інтегралів, що рахували на початку
ep.d.sq <- 3/5
ep.D <- 1/5
```

Гауссова двокомпонентна суміш: генерування послідовностей та обчислення характеристик.

Опишемо програмну реалізацію генерування псевдовипадкових послідовностей із заданим розподілом з використанням стандартного генератора в R:

```
r.norm.mixt <- function(n, m1, m2, s1, s2, p)
{
  ind <- sample(c(1, 2), n, replace = T, prob = c(p, 1 - p))
  rnorm(n, mean = c(m1, m2)[ind], sd = c(s1, s2)[ind])
}
```

Реалізуємо підрахунок щільності заданого розподілу:

```
d.norm.mixt <- function(t, m1, m2, s1, s2, p)
{
  p * dnorm(t, m1, s1) + (1 - p) * dnorm(t, m2, s2)
}
```

Додатково опишемо обчислення функції розподілу та квантилів – це знадобиться для побудови графіків:

```
p.norm.mixt <- function(t, m1, m2, s1, s2, p)
{
  p * pnorm(t, m1, s1) + (1 - p) * pnorm(t, m2, s2)
}

q.norm.mixt <- function(alpha, m1, m2, s1, s2, p, q0 = 1)
{
  P <- function(q) { p.norm.mixt(q, m1, m2, s1, s2, p) - alpha }
  nleqslv::nleqslv(q0, P)$x
}
```

Параметричний підхід. Правило Сільвермана.

Спочатку опишемо функцію, де можна підставити довільну оцінку для розкиду, а потім розглянемо часткові випадки (про які згадувалося раніше):

```
# Правило Сільвермана: загальна формула
h.silv.rule <- function(d.sq, D, n, sigma.est)
{ sigma.est * ((8 * sqrt(pi) * d.sq) / (3 * D^2 * n))^(1/5) }

# Просте правило Сільвермана
h.silv.simple <- function(d.sq, D, x)
{ h.silv.rule(d.sq, D, length(x), sd(x)) }

# Поліпшене правило Сільвермана
h.silv.improved <- function(d.sq, D, x)
{ h.silv.rule(d.sq, D, length(x), min(sd(x), IQR(x)/1.34)) }
```

Теоретично оптимальний параметр згладжування.

Обчислимо теоретично оптимальний параметр згладжування h_{opt} за формулою

$$h_{opt} = \left(\frac{d^2}{nD^2\varphi}\right)^{1/5},\ \varphi = \int\limits_{-\infty}^{+\infty} (f''(t))^2 dt$$

де $f(t) = \frac{1}{2\sqrt{2\pi}}\left(e^{-(t-1)^2/2} + e^{-(t+1)^2/2}\right)$ – щільність, яка досліднику невідома. Неважко показати, що

$$f''(t) = \frac{1}{2\sqrt{2\pi}}\left(((t-1)^2-1)e^{-(t-1)^2/2} + ((t+1)^2-1)e^{-(t+1)^2/2}\right)$$

А тому квадрат від другої похідної матиме вигляд:

$$(f''(t))^2 = \frac{1}{8\pi}\left(((t-1)^2-1)e^{-(t-1)^2/2} + ((t+1)^2-1)e^{-(t+1)^2/2}\right)^2 =$$
$$= \frac{1}{8\pi}\left(((t-1)^2-1)e^{-(t-1)^2/2} + ((t+1)^2-1)e^{-(t+1)^2/2}\right)^2 =$$
$$= \frac{1}{8\pi}\left(((t-1)^2-1)^2e^{-(t-1)^2} + 2((t-1)^2-1)((t+1)^2-1)e^{-(t+1)^2/2-(t-1)^2/2} + ((t+1)^2-1)^2e^{-(t+1)^2}\right)$$

Обчислимо інтеграли:

$$\int\limits_{-\infty}^{+\infty} ((t-1)^2-1)^2e^{-(t-1)^2}dt = \left|z = t-1\right| = \int\limits_{-\infty}^{+\infty}(z^2-1)^2e^{-z^2}dz = \left|z = u/\sqrt{2}\right| =$$
$$= \sqrt{\pi}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(\frac{u^2}{2}-1\right)^2 e^{-\frac{u^2}{2}}du = \sqrt{\pi}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(\frac{u^4}{4}-u^2+1\right)e^{-\frac{u^2}{2}}du =$$
$$= \sqrt{\pi}\left(\frac{3}{4}-1+1\right) = \sqrt{\pi}\cdot\frac{3}{4}$$

Аналогічно можна показати, що

$$\int\limits_{-\infty}^{+\infty} ((t+1)^2-1)^2e^{-(t+1)^2}dt = \sqrt{\pi}\cdot\frac{3}{4}$$

Залишається обчислити другий інтеграл:

$$\int\limits_{-\infty}^{+\infty} ((t-1)^2-1)((t+1)^2-1)e^{-t^2-1}dt = \left|t = z/\sqrt{2}\right| =$$
$$= \frac{\sqrt{\pi}}{e}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(\left(\frac{z}{\sqrt{2}}-1\right)^2-1\right)\left(\left(\frac{z}{\sqrt{2}}+1\right)^2-1\right)e^{-\frac{z^2}{2}}dz =$$
$$= \frac{\sqrt{\pi}}{e}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(\frac{z^4}{4}-2z^2\right)e^{-\frac{z^2}{2}}dz = \frac{\sqrt{\pi}}{e}\left(\frac{3}{4}-2\right) = -\frac{\sqrt{\pi}}{e}\cdot\frac{5}{4}$$

Отже $\varphi = \frac{1}{8\sqrt{\pi}} \cdot \left(\frac{3}{2} - \frac{5}{2e}\right) \approx 0.040925$, звідки

$$h_{opt} = n^{-1/5} \left(15 \cdot \left(\frac{1}{8\sqrt{\pi}} \cdot \left(\frac{3}{2} - \frac{5}{2e}\right)\right)^{-1}\right)^{1/5} \approx n^{-1/5} \cdot 3.257021$$

Програмна реалізація очевидна, тому не наводиться.

Непараметричний підхід.

Ядерна функція має простий вигляд, можна "погратися" з спрощенням обчислень. Перші дві похідні $K(t)$ мають вигляд:

$$K'(t) = -\frac{3}{2} \cdot t \cdot \mathbb{1}\{|t| < 1\},\ K''(t) = -\frac{3}{2} \cdot \mathbb{1}\{|t| < 1\}$$

Можна собі дозволити спрощення запису для $\hat{f}_n''(t)$:

$$\hat{f}_n''(t) = \frac{1}{nh^3} \sum_{j=1}^{n} K'' \left(\frac{t - \xi_j}{h}\right) = -\frac{3}{2} \cdot \frac{1}{nh^3} \sum_{j=1}^{n} \mathbb{1}\{|t - \xi_j| < h\}$$

Тоді, припускаючи що $\xi_i \le \xi_j$ при $i < j$:

$$\left(\hat{f}_n''(t)\right)^2 = \frac{9}{4} \cdot \frac{1}{n^2h^6} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\{\xi_i - h < t < \xi_i + h\} \mathbb{1}\{\xi_j - h < t < \xi_j + h\} =$$

$$= \frac{9}{4} \cdot \frac{1}{n^2h^6} \left(\sum_{j=1}^{n} \mathbb{1}\{\xi_j - h < t < \xi_j + h\} + 2 \sum_{i=1}^{n} \sum_{j<i} \mathbb{1}\{\xi_i - h < t < \xi_i + h\} \mathbb{1}\{\xi_j - h < t < \xi_j + h\}\right)$$

Тому можна спробувати спростити запис для $\hat{\varphi}$:

$$\hat{\varphi} = \int_{-\infty}^{+\infty} \left(\hat{f}_n''(t)\right)^2 dt = \frac{9}{2} \cdot \frac{1}{n^2h^6} \left(nh + \sum_{i=1}^{n} \sum_{j<i} \int_{-\infty}^{+\infty} \mathbb{1}\{\xi_i - h < t < \xi_i + h\} \mathbb{1}\{\xi_j - h < t < \xi_j + h\} dt\right) =$$

$$= \frac{9}{2} \cdot \frac{1}{n^2h^6} \left(nh + \sum_{i=1}^{n} \sum_{j<i} (\xi_j - \xi_i + 2h) \cdot \mathbb{1}\{\xi_j - \xi_i + 2h > 0\}\right)$$

А тому можна не звертатися до чисельних методів інтегрування і все порахувати "в лоб":

```
phi.estim.epan <- function(x, h.pilot)
{
  n <- length(x); idx <- 1:n; sx <- sort(x)
  s <- n * h.pilot + sum(sapply(idx[-1], function(i) {
    deltai <- sx[idx < i] - sx[i] + 2 * h.pilot
    sum(deltai * (deltai > 0))
  }))
  s * 9/2 * 1/(n^2 * h.pilot^6)
}

h.nonparam.epan <- function(x, h.pilot)
{
  phi.estim <- phi.estim.epan(x, h.pilot)
  (ep.d.sq / (length(x) * ep.D^2 * phi.estim))^(1/5)
}
```

Техніка кросс-валідації.

Можна знову повеселитися з спрощеннями. Доведемо, що CV(h) можна подати у більш зручній формі для обчислення:

$$\mathrm{CV}(h) = \int\limits_{-\infty}^{\infty} (\hat{f}_n(t))^2 dt - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_n^{i-}(\xi_i) =$$

$$= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int\limits_{-\infty}^{+\infty} K\left(\frac{t-\xi_i}{h}\right) K\left(\frac{t-\xi_j}{h}\right) dt - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K\left(\frac{\xi_i-\xi_j}{h}\right)$$

Легко бачити, що інтеграл у подвійній сумі – це згортка ядер (врахувавши парність ядра):

$$\int\limits_{-\infty}^{+\infty} K\left(\frac{t-\xi_i}{h}\right) K\left(\frac{t-\xi_j}{h}\right) dt = \left| z = \frac{t-\xi_j}{h} \right| = h\int\limits_{-\infty}^{+\infty} K\left(z\right) K\left(z - \frac{\xi_i-\xi_j}{h}\right) dz = h(K{*}K)\left(\frac{\xi_i-\xi_j}{h}\right)$$

Отже CV - функціонал спрощується до вигляду:

$$\mathrm{CV}(h) = \frac{1}{n^2h}\left(nd^2 + 2\sum_{i=1}^{n}\sum_{j<i}(K*K)\left(\frac{\xi_i-\xi_j}{h}\right)\right) - \frac{4}{n(n-1)h}\sum_{i=1}^{n}\sum_{j<i} K\left(\frac{\xi_i-\xi_j}{h}\right) =$$

$$= \frac{1}{nh}\left(d^2 + 2\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j<i}(K*K)\left(\frac{\xi_i-\xi_j}{h}\right) - \frac{2}{n-1}\sum_{j<i} K\left(\frac{\xi_i-\xi_j}{h}\right)\right)\right)$$

Програмна реалізація матиме вигляд (де згортка була попередньо обчислена на папері):

```
f.a <- function(t, a) { ((1 + t)^a - (-1)^a) / a }

epan.kernel.conv <- function(t)
{
  ifelse(abs(t) >= 2, 0, {
    z <- -abs(t)
    (1-z^2) * f.a(z,1) + 2*z*(f.a(z,2) - f.a(z,4)) + (z^2-2)*f.a(z,3) + f.a(z,5)
  })
}

CV.h <- function(h, x)
{
  n <- length(x); idx <- 1:n
  double.sum <- sum(sapply(idx, function(j) {
    delta <- (x[idx < j] - x[j])/h
    A <- sum(epan.kernel.conv(delta)) * 9/16; B <- sum(epan.kernel(delta))
    A / n - 2 * B / (n - 1)
  }))
  (ep.d.sq + 2 * double.sum) / (n * h)
}

h.crossvalid <- function(x, h.min, h.max)
{ optimize(function(h) { CV.h(h, x) }, c(h.min, h.max))$minimum }
```

Емпіричний аналог MISE. Обчислення aMISE.

На кожній повторній вибірці рахуємо ISE, далі – беремо середнє арифметичне.

```
l2.sq <- function(f, g, a = -Inf, b = +Inf, n.subdiv = 10^4)
{
  i <- integrate(function(t) { (f(t) - g(t))^2 },
                 a, b, subdivisions = n.subdiv, stop.on.error = F)
  i$value
}

# Побудова графіку емпіричного MISE як функції від h
plot.l2.sq <- function(f, K, num.gen, N, seed.val = 0, n.sub = 10^3,
                       h.min = 0.01, h.max = 2, h.n = 200, B = 100)
{
  set.seed(seed.val)
  h.vals <- ((h.n-1):0)/(h.n-1) * h.min + (0:(h.n-1))/(h.n-1) * h.max
  l2.val <- rep(0, h.n)
  for(b in 1:B)
  {
    x.boot <- r.gen(N)
    for(j in 1:h.n)
    {
      l2.val[j] <- l2.val[j] + l2.sq(f, dens.estim(x.boot, h.vals[j], K))
    }
  }
  l2.val <- l2.val / B
  plot(h.vals, l2.val, type = "l",
       xlab = "smoothing", ylab = "l2-squared",
       main = paste("distance plot, N =", N, ", B =", B))
  grid()
}
```

Програмна реалізація далека від хорошої, обчислення інтегралів не оптимізовано. Тому, наприклад, навіть при $N = 1000$ функція plot.l2.sq виконувалася близько п'яти хвилин. От для обчислення aMISE все просто реалізується (будемо чесними, все більш-менш нескладно реалізувати):

```
true.phi.val <- (3/2 - 5/(2*exp(1))) / (8 * sqrt(pi))

aMISE <- function(H)
{
  1/4 * true.phi.val * ep.D^2 * H^4 + ep.d.sq / H
}
```

Моделювання, застосування.

Далі змоделюємо три вибірки з різних зернин відповідно: 0 (Seed 1), 543787 (Seed 2) та 1912346 (Seed 3) (числа набиралися випадковим набиранням цифр з клавіатури). Продемонструємо застосування на вибірках обсягу: 500, 1000, 10000. Усі висновки будуть зроблені на око, хоча для $N = 1000$ зробимо графік оцінки MISE (по $B = 100$ повторних вибірках) як функції від параметра згладжування $h > 0$.

N = 500	**Seed 1**	**Seed 2**	**Seed 3**
h.silv.s	0.9267131	0.8979114	0.9840562
h.silv.i	0.9267131	0.8979114	0.9840562
h.nonp	0.4211175	0.4072796	0.4461268
h.cv	1.310031	1.088848	0.6199906
h.theor	**0.9397807**		
N = 1000	**Seed 1**	**Seed 2**	**Seed 3**
h.silv.s	0.8497688	0.8291105	0.8668361
h.silv.i	0.8497688	0.8291105	0.8668361
h.nonp	0.3439314	0.3354204	0.3504855
h.cv	1.071575	1.054707	0.6810802
h.theor	**0.8181266**		
N = 10000	**Seed 1**	**Seed 2**	**Seed 3**
h.silv.s	0.5279088	0.5191472	0.5237334
h.silv.i	0.5279088	0.5191472	0.5237334
h.nonp	0.1470581	0.1446232	0.1459301
h.cv	0.5381125	0.5670538	0.5278922
h.theor	**0.516203**		

Табл. 1: Значення параметрів згладжування при різних N та на різних вибірках. h.silv.s – просте правило Сільвермана, h.silv.i – поліпшене правило Сільвермана, h.nonp – непараметрична адаптивна оцінка, h.cv – приблизна точка мінімуму модифікованого функціонала кросс-валідації, h.theor – теоретично оптимальне значення.

Чесно кажучи, автор роботи на око підібрав би приблизно такі самі значення, що відповідають теоретично оптимальним. Зосередимося на графічних результатах.

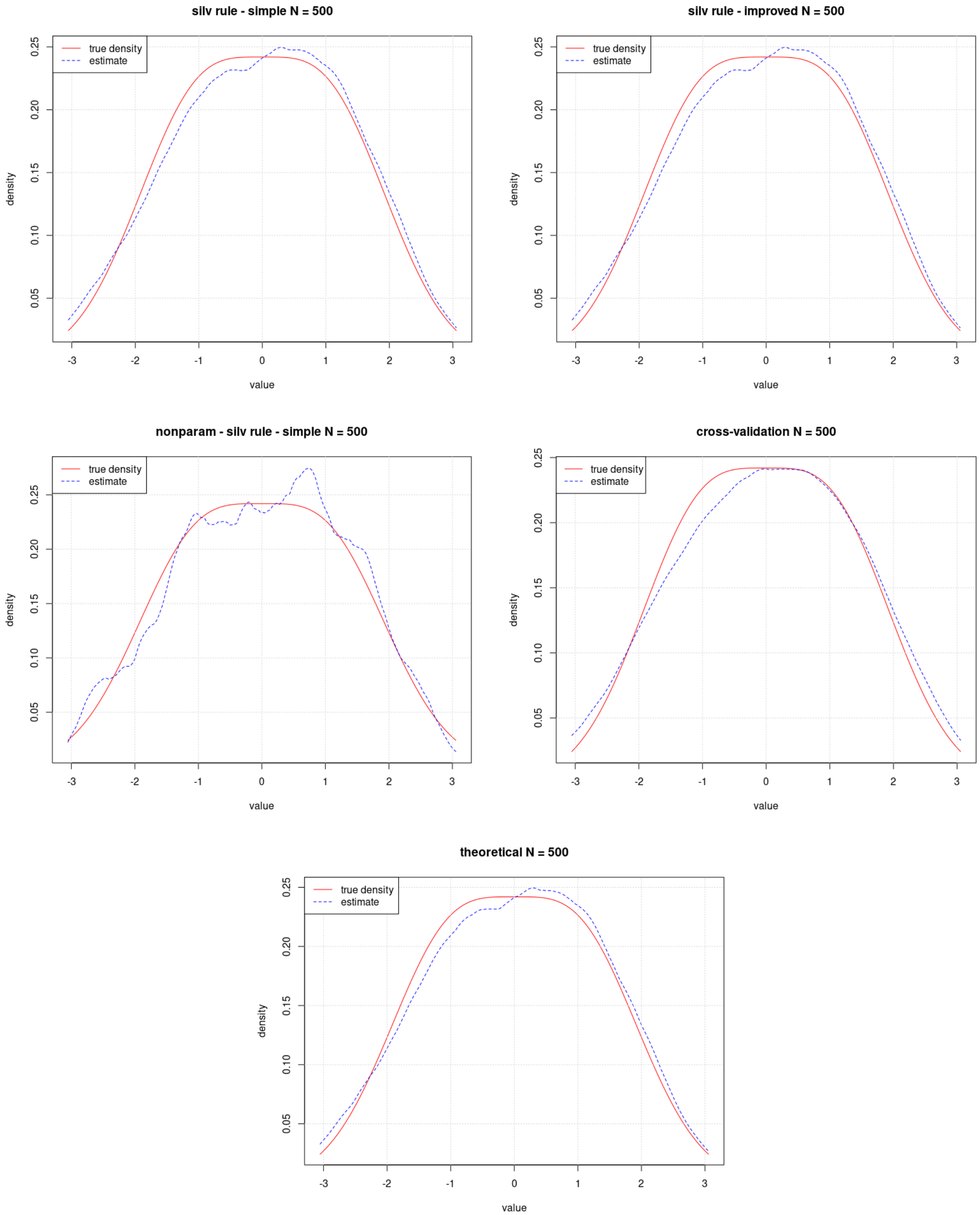

Рис. 1: Графіки ядерних оцінок: Seed 1, N = 500.

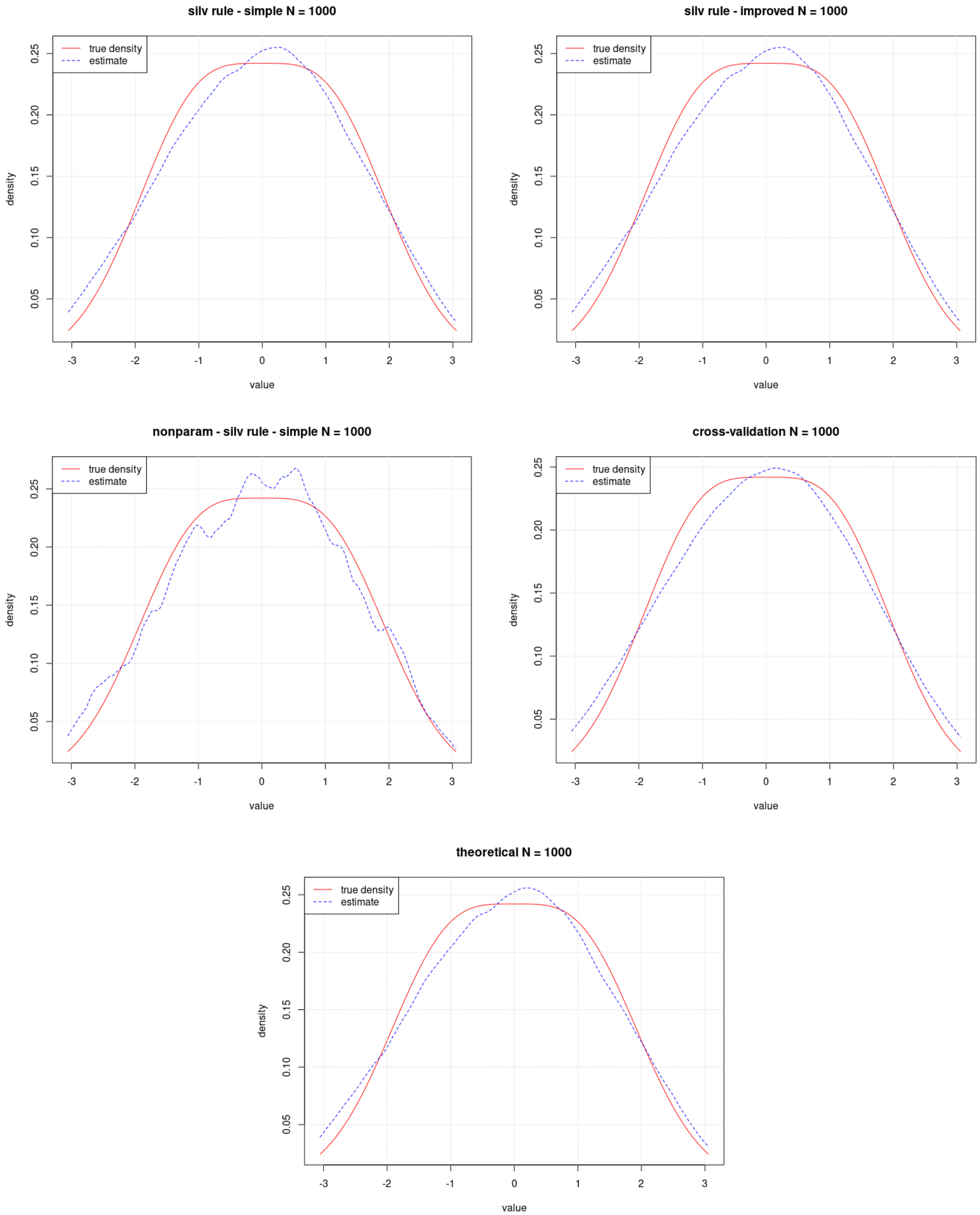

Рис. 2: Графіки ядерних оцінок: Seed 1, N = 1000.

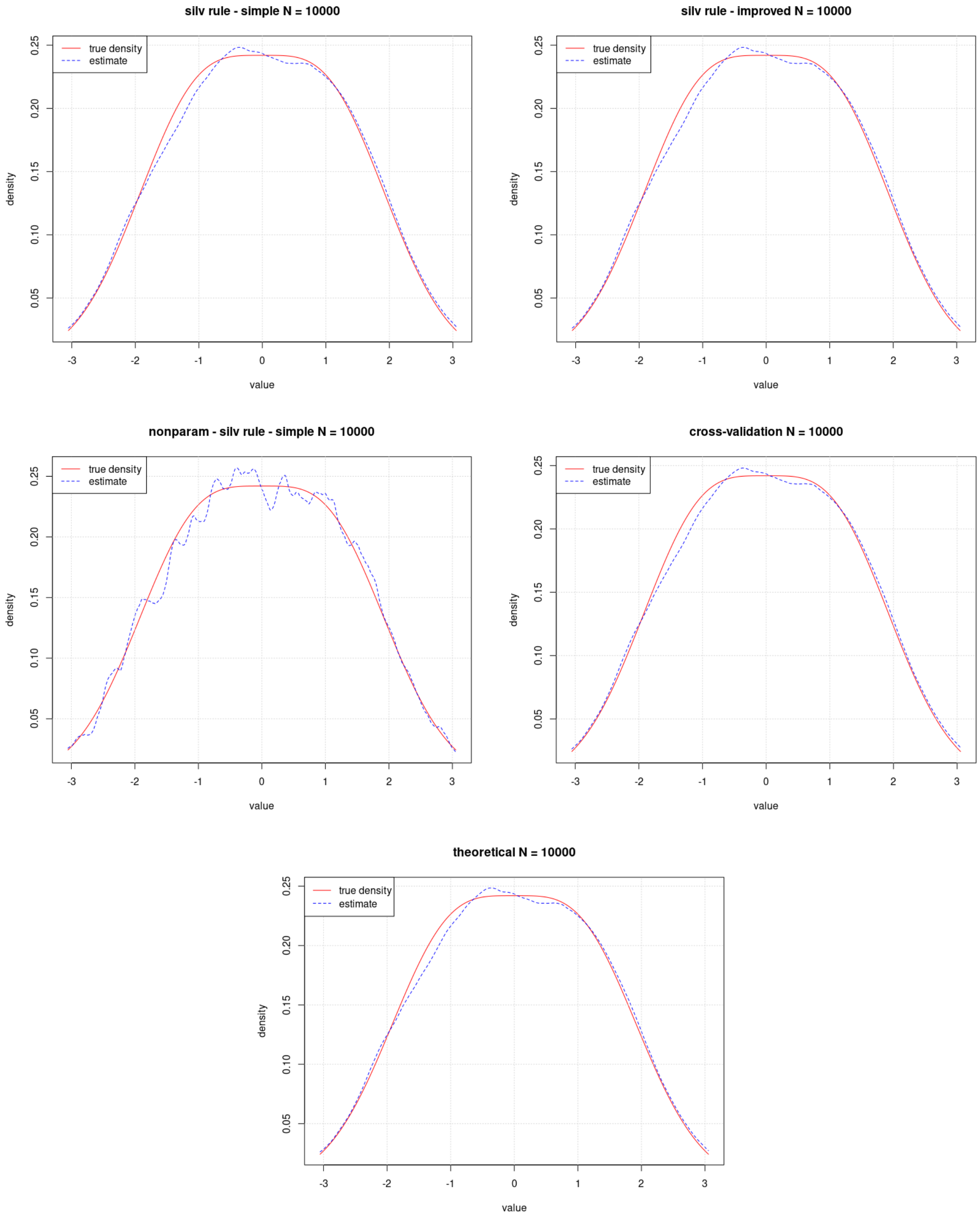

Рис. 3: Графіки ядерних оцінок: Seed 1, N = 10000.

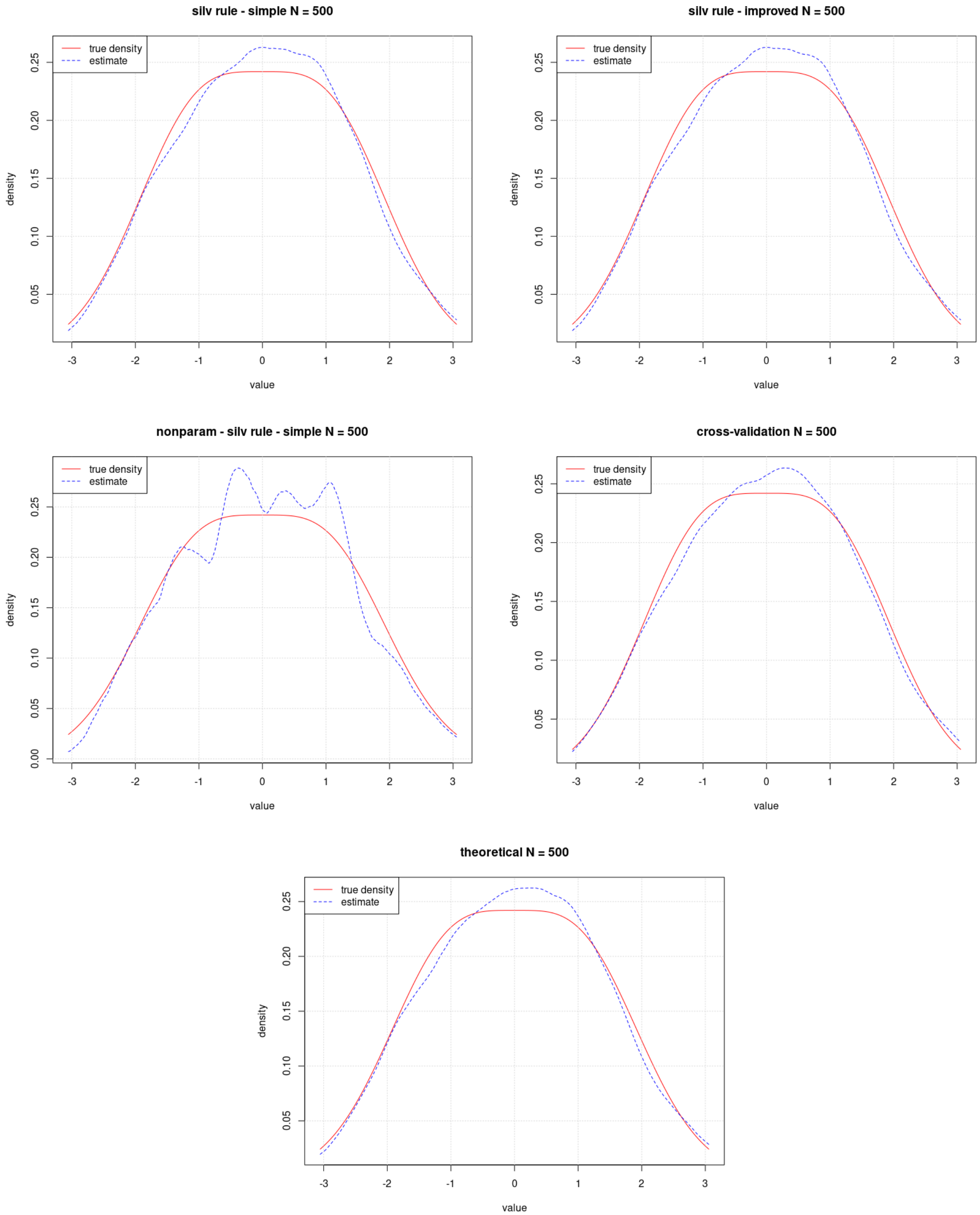

Рис. 4: Графіки ядерних оцінок: Seed 2, N = 500.

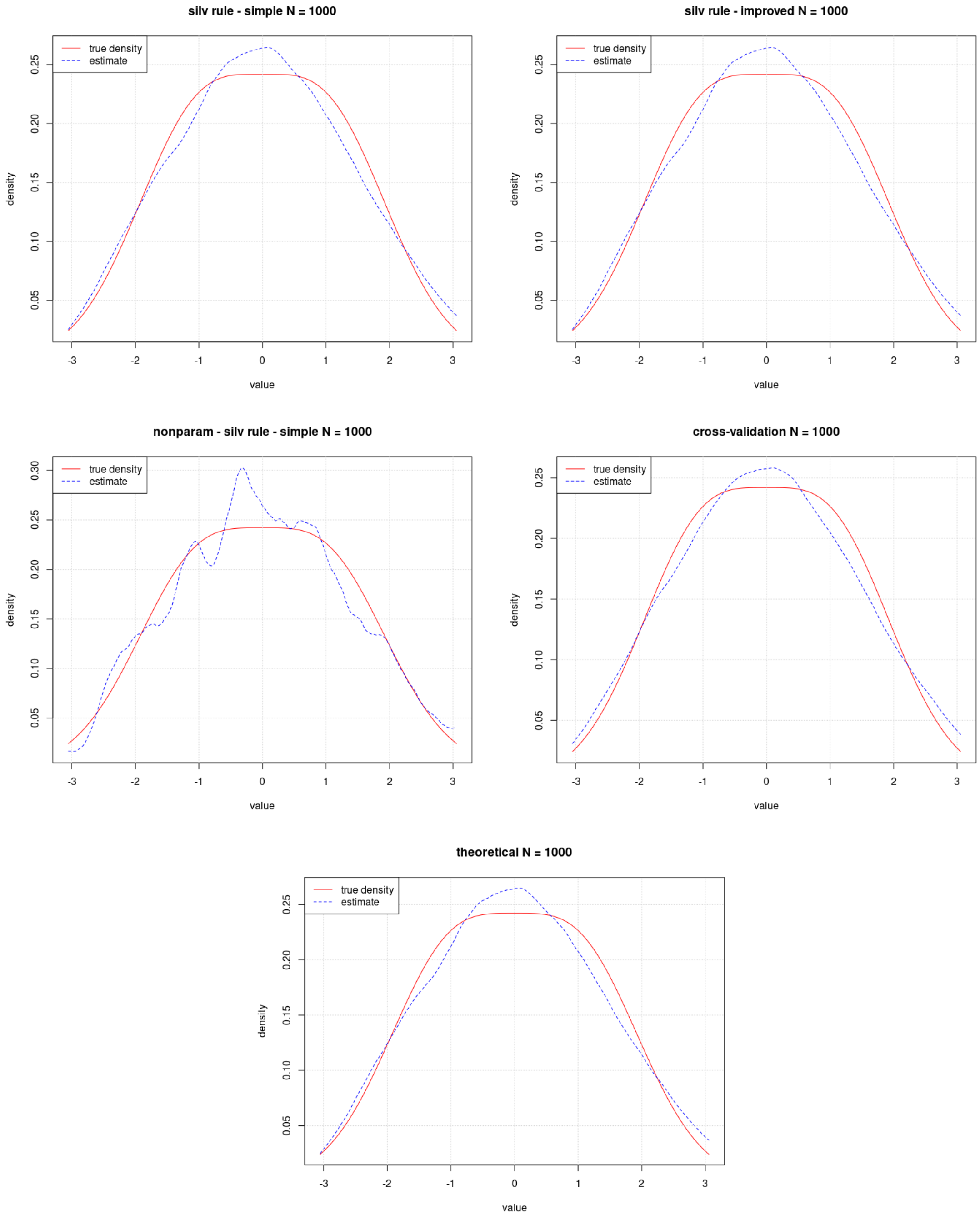

Рис. 5: Графіки ядерних оцінок: Seed 2, N = 1000.

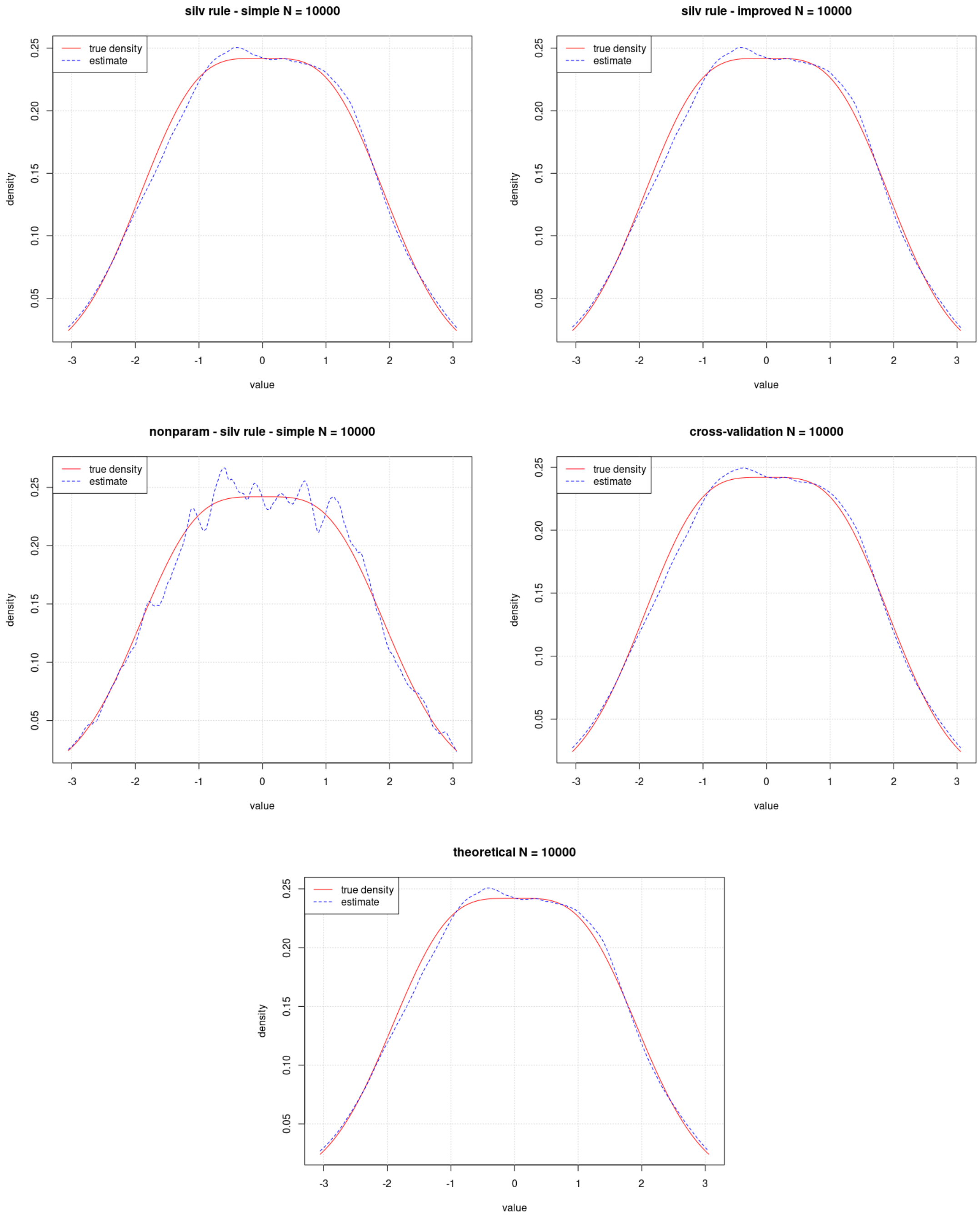

Рис. 6: Графіки ядерних оцінок: Seed 2, N = 10000.

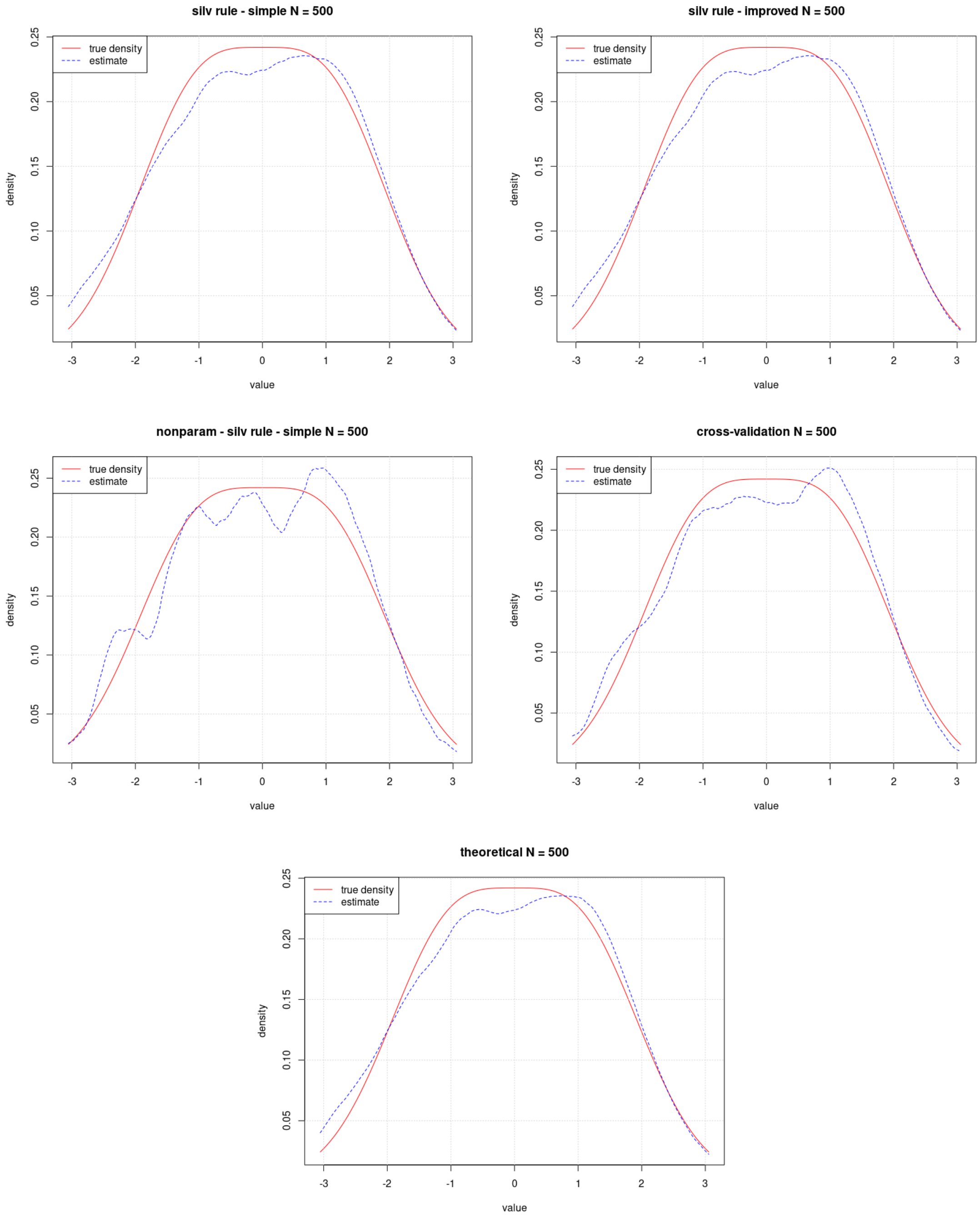

Рис. 7: Графіки ядерних оцінок: Seed 3, N = 500.

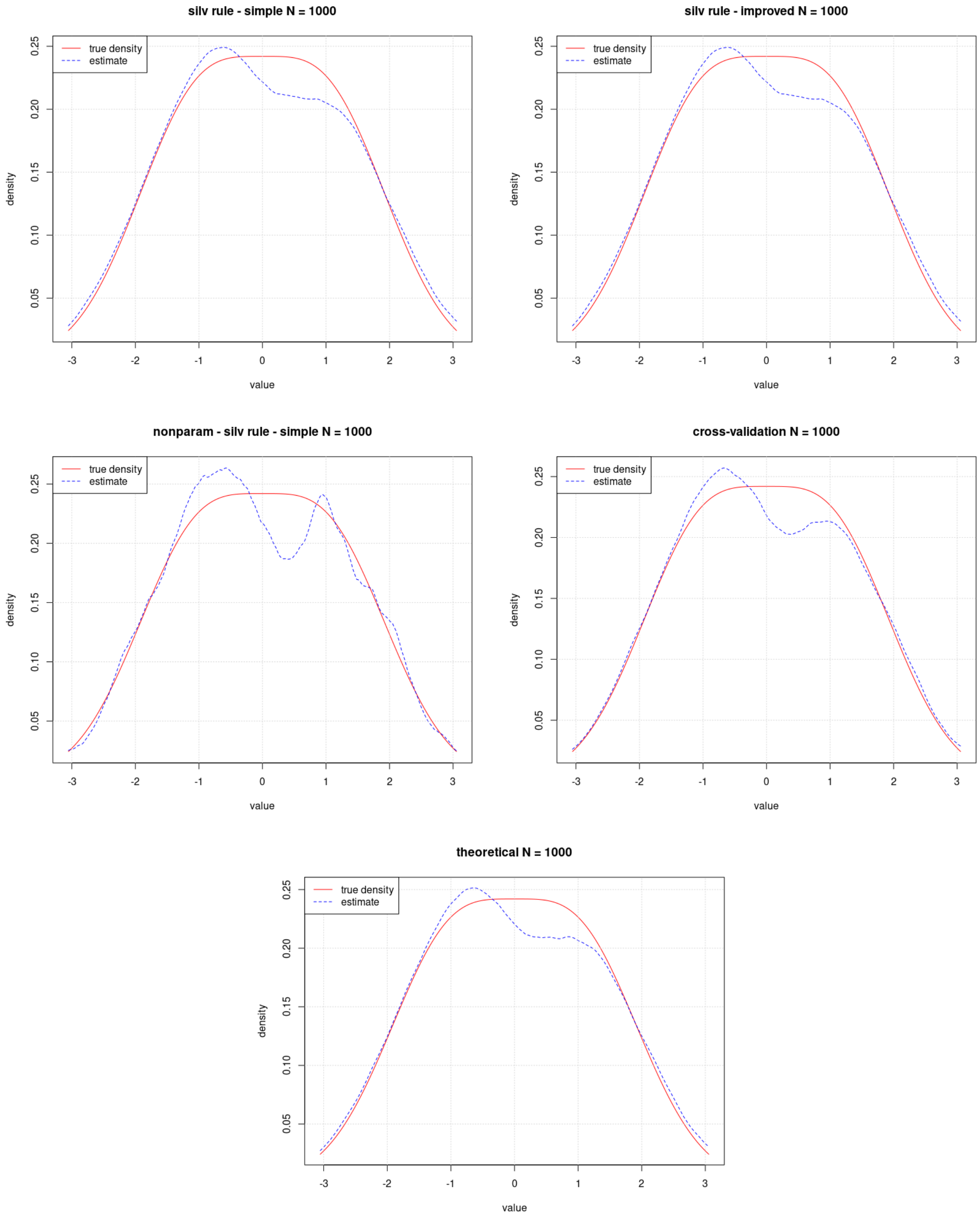

Рис. 8: Графіки ядерних оцінок: Seed 3, N = 1000.

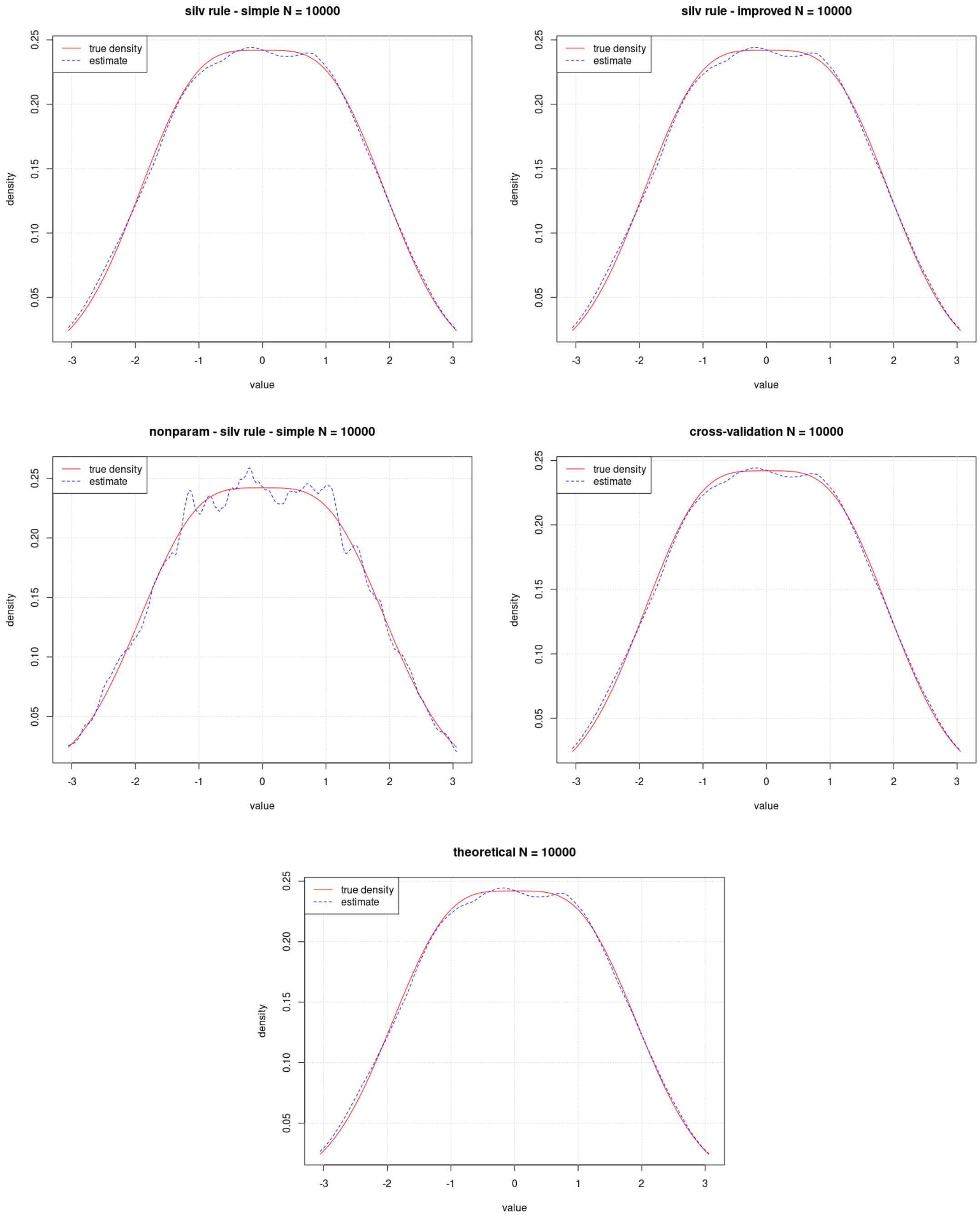

Рис. 9: Графіки ядерних оцінок: Seed 3, N = 10000.

Далі зобразимо графіки емпіричного MISE та aMISE для різних N = 500, 1000, 10000. На графіках відмічені значення різних оцінок параметра згладжування за вибіркою з зернини "Seed 1". Зокрема відмічено теоретично оптимальне значення параметра згладжування.

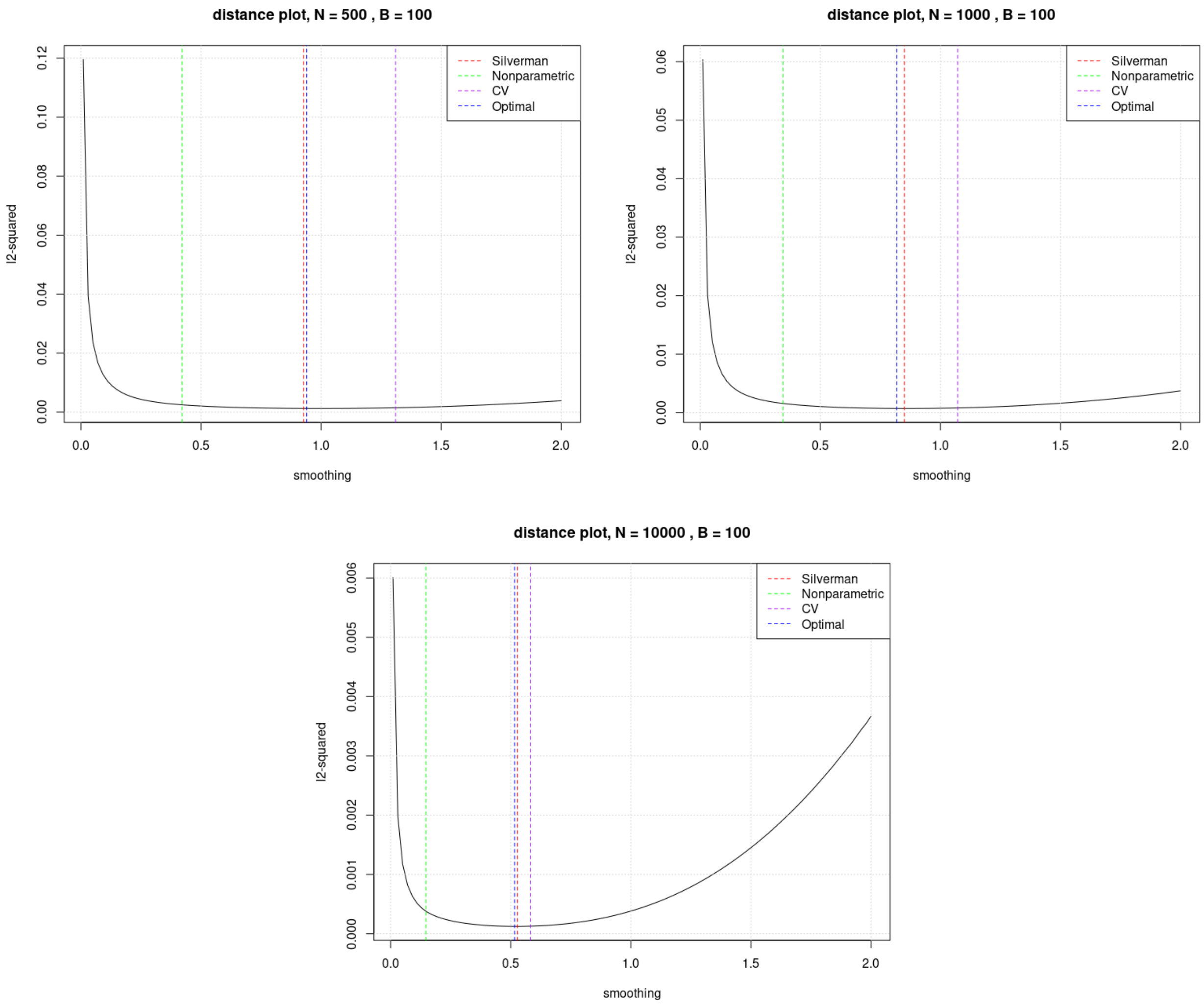

Рис. 10: Графіки емпіричного аналогу MISE, N = 500, 1000, 10000; Seed 1. Вертикальними лініями відмічені значення оцінок параметрів згладжування різними методами та, зокрема, оптимального значення.

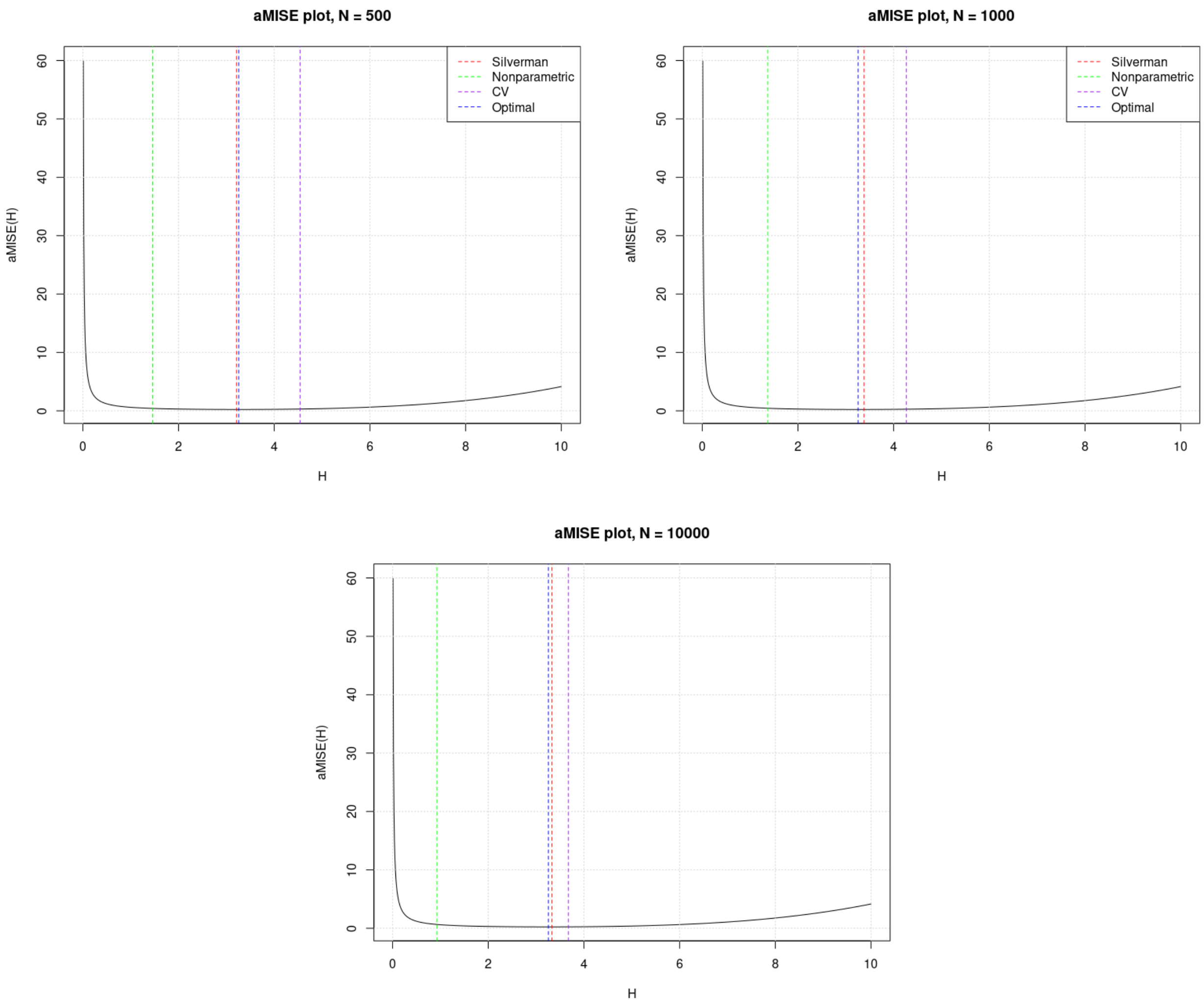

Рис. 11: Графіки емпіричного аналогу aMISE, N = 500, 1000, 10000; Seed 1. Вертикальними лініями відмічені значення оцінок параметрів згладжування різними методами та, зокрема, оптимального значення.

Висновки.

Теорема Стоуна для оцінки параметра згладжування методом кросс-валідації справджується на практиці: при збільшенні обсягу вибірки, відповідна оцінка наближається до оптимального. Теоретична щільність розподілу є унімодальною, що відіграло на доцільність використання правила Сільвермана. Загалом як кросс-валідаційні, так і за правилом Сільвермана оцінки дають непогані результати. Непараметрична (адаптивна) оцінка підвела у даній задачі: оцінене значення виходить занадто малим, що призводить до ефекту недозгладжування (варто звернути увагу на "шершавість" графіків ядерних оцінок з відповідним параметром згладжування). Втім, сильного відхилення від оптимальних мало чим відрізняється в сенсі досягнутого квадратичного ризику. Використовувати непараметричну оцінку здається не найкращим з рішень, оскільки простіша в обчисленні оцінка за правилом Сільвермана дає кращі наближення.