Теорія ймовірностей
та математична статистика
Вип. 96, 2017, с. 131–141

Teoriya Ĭmovirnosteĭ
ta Matematychna Statystyka
No. 96, 2017, pp. 131–141

УДК 519.21

# АСИМПТОТИЧНА НОРМАЛЬНІСТЬ ОЦІНОК КАПЛАНА – МЕЙЄРА ДЛЯ СУМІШЕЙ ЗІ ЗМІННИМИ КОНЦЕНТРАЦІЯМИ

Р. МАЙБОРОДА

Анотація. Розглядається модифікація оцінки Каплана – Мейєра для розподілу компонентів суміші зі змінними концентраціями за цензурованими даними. Доведено асимптотичну нормальність цих оцінок у рівномірній нормі.

*Ключові слова і фрази.* Оцінка Каплана – Мейєра, моделі сумішей зі змінними концентраціями, асимптотична нормальність, цензурування.

## 1. Вступ

Класична оцінка Каплана – Мейєра (КМЕ) широко застосовується в аналізі даних типу тривалості життя як непараметрична оцінка функції розподілу (ф. р.) за цензурованими даними. У [11] запропоновано модифікацію цієї оцінки (mKME) для випадку, коли спостереження отримано із суміші кількох популяцій (компонентів) зі змінними концентраціями (модель MVC). Консистентність mKME та оцінки для швидкості її збіжності одержано у [11].

У цій роботі доводиться асимптотична нормальність mKME в рівномірній нормі на скінченному інтервалі. Указаний результат є узагальненням на випадок mKME класичної теореми про асимптотичну нормальність КМЕ [5] і дозволяє отримати аналог формули Грінвуда для асимптотичної дисперсії mKME. Доведення ґрунтується на асимптотичній теорії навантажених емпіричних функцій розподілу [9, 10], теорії продакт-інтегралів з [5] та класичних результатах зі слабкої збіжності ймовірнісних мір у функціональних просторах [2].

Далі в п. 2.1 нагадується означення КМЕ для однорідних вибірок та результати щодо її асимптотичної нормальності. Оцінювання функцій розподілу в моделі MVC за відсутності цензурування розглянуто у п. 2.2. Означення mKME та основний результат статті по асимптотичній нормальності mKME міститься у п. 3. Доведення наведено у п. 4.

## 2. Попередні відомості

2.1. **Цензурування та оцінка Каплана – Мейєра за однорідною вибіркою.** Почнемо з опису стандартної моделі випадкового цензурування справа і конструкції КМЕ.

Нехай $\xi_j$, $j = 1, \dots, n$, є тривалостями життя деяких об'єктів, причому вони вважаються незалежними, однаково розподіленими невід'ємними випадковими величинами. Ці тривалості спостерігаються, якщо вони є меншими ніж моменти цензурування $C_j$ для відповідних об'єктів (тобто, якщо об'єкт загинув раніше, ніж був відцензурований). Якщо цензурування передувало загибелі, то спостерігається лише

2000 *Mathematics Subject Classification.* 62N05, 62G05.

момент цензурування. Для кожного об'єкта відомо також, чи відбулось цензурування.

Таким чином, маємо такі спостережувані дані: $\mathbf{X} = (\xi_j^*, \delta_j, j = 1\dots, n)$, де $\xi_j^* = \min(\xi_j, C_j)$ — цензуровані моменти загибелі, $\delta_j = \mathbb{1}\{\xi_j \le C_j\}$ — індикатори того, що цензурування не відбулось.

Ф. р. $F$ випадкових величин $\xi_j$ вважається невідомою. КМЕ є оцінкою емпіричного методу найбільшої вірогідності для $F$ за даними $\mathbf{X}$ [13]. Вона визначається таким чином.

Позначимо

$$\widehat{Y}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{\xi_j^* \ge t\right\},$$

$$\widehat{N}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{\xi_j^* \le t, \delta_j = 1\right\}.$$

Нехай усі $\xi_j^*$ у вибірці є різними. Тоді класична оцінка Каплана–Мейєра для $F(x)$ визначається як

$$\widehat{F}(t) = 1 - \prod_{j:\xi_j^*\le t}\left(1 - \frac{\Delta\widehat{N}_n(\xi_j^*)}{\widehat{Y}_n(\xi_j^*)}\right) = 1 - \prod_{j:\xi_j^*\le t}\left(1 - \frac{\delta_j}{n - \sum\limits_{i:\xi_i^*<\xi_j^*} 1}\right), \tag{1}$$

де $\Delta\widehat{N}_n(t) = \widehat{N}_n(t) - \widehat{N}_n(t-)$ це стрибок функції $\widehat{N}_n$ у точці $t$. (Границю функції $F$ зліва будемо позначати $F(t-) = \lim_{s<t, s\to t} F(s)$.)

Припустимо, що моменти цензурування $C_j$ також є незалежними та однаково розподіленими із ф. р. $G$. Якщо $F$ і $G$ є неперервними і $F(t) < 1$, $G(t) < 1$ для деякого $t > 0$, то

$$\sqrt{n}\left(\widehat{F}_n(t) - F(t)\right) \xrightarrow{\text{w}} N\left(0, \sigma^2(t)\right), \tag{2}$$

де

$$\sigma^2(t) = \big(1 - F(t)\big)^2 \int_0^t \frac{F(du)}{\big(1 - G(u)\big)\big(1 - F(u)\big)^2} \tag{3}$$

(див. [5], тут і далі $\xrightarrow{\text{w}}$ позначає слабку збіжність). Рівняння (3) є асимптотичною версією класичної формули Грінвуда для КМЕ [6, рівняння (3.2.31)].

2.2. **Суміші зі змінними концентраціями.** У моделі MVC вважається, що кожен спостережуваний об'єкт $O$ належить одній із $M$ різних підпопуляцій. Справжній номер $\text{ind}(O)$ популяції, якій належить $O$, — невідомий. Спостерігається деяка характеристика $\xi = \xi(O)$, що вважається випадковою величиною із ф. р. $F_m$, залежною від того, якій підпопуляції належить $O$, тобто

$$F_m(t) = \mathsf{P}\{\xi(O) \le t \mid \text{ind}(O) = m\}.$$

Отже ф. р. спостережуваного $\xi(O)$ є сумішшю $F_m$. Функції $F_m$ вважаються невідомими, але відомі концентрації компонентів у суміші, які є різними для різних спостережень. (У [1, 3, 10] розглянуто різні задачі, пов'язані з цією моделлю.)

Отже, якщо спостерігається $n$ незалежних об'єктів $O_j$, статистичні дані набувають вигляду $(\xi_{j;n},\ j = 1, \dots, n)$, де $\xi_{j;n} = \xi(O_j)$ має ф. р.

$$\mathsf{P}\{\xi_{j;n} \le t\} = \sum_{m=1}^{M} p_j^m F_m(t). \tag{4}$$

Тут

$$p_j^m = \mathsf{P}\{\text{ind}(O_j) = m\}$$

— концентрація (змішуюча ймовірність) для $m$-го компонента (субпопуляції) у суміші під час спостереження $j$-го об'єкта.

Набір усіх концентрацій $\{p^m_{j;n}, j = 1, \dots, n;\ m = 1, \dots, M;\ n = 1, 2, \dots\}$ будемо позначати $\mathbf{p}$. $\mathbf{p}^m_{;n} = (p^m_{1,n}, \dots, p^m_{n;n})^T$ — вектор-стовпець концентрацій $m$-го компонента, $\mathbf{p}_{j;n} = (p^1_{j,n}, \dots, p^M_{j,n})^T$ — вектор-стовпець концентрацій на момент $j$-го спостереження, $\mathbf{p}_{;n} = (p^m_{j,n}, j = 1, \dots, n;\ m = 1, \dots, M)$ — матриця концентрацій для вибірки з $n$ елементів, що має $n$ стовпців і $M$ рядків.

Аналогічні позначення використовуються для набору вагових коефіцієнтів

$$\mathbf{a} = \{a^m_{j;n}, j = 1, \dots, n;\ m = 1, \dots, M;\ n = 1, 2, \dots\},$$

які будуть введені пізніше.

Усереднення по всій вибірці (тобто, за індексом $j$) позначається кутовими дужками:

$$\left\langle \mathbf{p}^m \mathbf{a}^k \right\rangle_n = \frac{1}{n} \sum_{j=1}^{n} p^m_{j;n} a^k_{j;n}, \quad \left\langle (\mathbf{a}^k)^2 \right\rangle_n = \frac{1}{n} \sum_{j=1}^{n} \left(a^k_{j;n}\right)^2$$

і т. д. (Додавання, множення, піднесення до степеня всередині кутових дужок виконується покоординатно.) Кутові дужки без нижнього індексу позначають границю $\left\langle \mathbf{p}^m \mathbf{a}^k \right\rangle = \lim_{n\to\infty} \left\langle \mathbf{p}^m \mathbf{a}^k \right\rangle_n$ у припущенні, що вона існує.

Позначимо

$$\mathbf{\Gamma}_n = \frac{1}{n} \mathbf{p}_{;n} \mathbf{p}^T_{;n} = \left( \left\langle \mathbf{p}^m \mathbf{p}^k \right\rangle_n \right)^M_{m,k=1}, \quad \mathbf{\Gamma} = \lim_{n\to\infty} \mathbf{\Gamma}_n = \left( \left\langle \mathbf{p}^m \mathbf{p}^k \right\rangle \right)^M_{m,k=1}.$$

У [8, 9] (див. також [10]), для оцінювання $F_m(t)$ по $\mathbf{X}$ запропоновано використовувати

$$\widehat{F}_{\mathbf{a}}(t) = \frac{1}{n} \sum_{j=1}^{n} a_{j;n} \mathbb{1}\left\{\xi_{j;n} \le t\right\}.$$

У [10] показано, що $\widehat{F}_m(t) = \widehat{F}_{\mathbf{a}^m}(t)$ є мінімаксною оцінкою $F_m$ у класі всіх незміщених оцінок, якщо

$$\mathbf{a}^m_{;n} = \mathbf{\Gamma}^{-1} \mathbf{p}_{;n}. \tag{5}$$

Вагові коефіцієнти $\mathbf{a}^m_{;n}$ називають мінімаксними для $m$-го компонента суміші.

## 3. Основні результати

Розглянемо тепер цензуровані дані, отримані із суміші зі змінними концентраціями.

Нехай спостерігається $n$ об'єктів $O_{j;n}$ кожен з яких належить одному із $M$ компонентів (субпопуляцій), $\xi_{j;n} = \xi(O_{j;n}) > 0$ — змінна, що досліджується (тривалість життя) об'єкта $O_{j;n}$, $C_{j;n} = C(O_{j;n})$ — момент цензурування для $O_{j;n}$. Вважається, що $\xi(O)$ і $C(O)$ є незалежними при фіксованому компоненті, якому належить $O$, і $(\xi_{j;n}, C_{j;n})$, $j = 1, \dots, n$, є незалежними при фіксованому $n$. (Далі розглядається асимптотична теорія при $n \to \infty$ у схемі серій. При цьому не накладаються жодні припущення про зв'язок між спостереженнями при різних $n$.)

Спостерігається цензурована вибірка $\mathbf{X}_n = \left(\xi^*_{j;n}, \delta_{j;n}, j = 1, \dots, n\right)$, у якій $\xi^*_{j;n} = \min\left(\xi_{j;n}, C_{j;n}\right)$ є цензурованою тривалістю життя $j$-го спостережуваного об'єкта і $\delta_{j;n} = \mathbb{1}\left\{\xi_{j;n} < C_{j;n}\right\}$ є індикатором того, що цензурування не відбулось (він дорівнює 1 тоді і тільки тоді, коли $j$-те спостереження не було цензуроване).

Відмітимо, що у цій моделі розподіл моменту цензурування є однаковим для всіх спостережень з одного компонента і може бути різним для різних компонентів. Іншу модель цензурування для даних із суміші зі змінними концентраціями розглянуто у [12].

Нехай $\kappa_{j;n} = \mathrm{ind}(O_{j;n})$ — номер компонента, якому належить $O_{j;n}$. Концентрації компонентів $p^m_{j;n} = \mathsf{P}\{\kappa_{j;n} = m\}$ вважаються відомими.

Позначимо

$$F_m(x) = \mathsf{P}\{\xi(O) \le x \mid \mathrm{ind}(O) = m\}$$

— ф. р. тривалості життя для $m$-го компонента,

$$G_m(x) = \mathsf{P}\{C(O) \le x \mid \mathrm{ind}(O) = m\}$$

— ф. р. моментів цензурування для $m$-го компонента.

Далі для будь-якої ф.р. $F$ відповідну функцію виживання будемо позначати $\bar{F}(t) = 1 - F(t)$.

Для функцій $F(t)$ з обмеженою варіацією на довільній вимірній підмножині $A$ дійсної прямої введемо позначення

$$F(A) = \int_A F(dt).$$

Для $A = ]t_1, t_2]$ це рівносильно $F(A) = F(t_2) - F(t_1)$. Останнє позначення будемо використовувати і тоді, коли $A$ є інтервалом, але $F$ має необмежену варіацію.

Ф. р. $F_m$ та $G_m$, $m = 1, \dots, M$, є невідомими. Визначимо модифікацію оцінки Каплана–Мейєра (mKME) для $F_k$ для деякого фіксованого $1 \le k \le M$. Позначимо

$$\widehat{Y}_{m;n}(t) = \frac{1}{n} \sum_{j=1}^{n} a^m_j \mathbb{1}\{\xi^*_{j;n} \ge t\}$$

— навантажена емпірична функція розподілу цензурованих даних із ваговими коефіцієнтами $a^k_{j;n}$;

$$\widehat{N}_{m;n}(t) = \frac{1}{n} \sum_{j=1}^{n} a^m_j \mathbb{1}\{\xi^*_{j:n} \le t, \delta_{j;n} = 1\}$$

— навантажена емпірична функція для нецензурованих даних. Тепер mKME для $F_k(t)$ визначається як

$$\widehat{F}_{k,n}(t) = 1 - \prod_{j:\xi^*_{j;n} \le t} \left(1 - \frac{\Delta \widehat{N}_{k;n}(\xi^*_{j;n})}{\widehat{Y}_{k;n}(\xi^*_{j;n})}\right) = 1 - \prod_{j:\xi^*_{j;n} \le t} \left(1 - \frac{a^k_j \delta_j}{n - \sum\limits_{i:\xi^*_{i;n} < \xi^*_{j;n}} a^k_i}\right). \quad (6)$$

Консистентність цієї оцінки доведена у [11].

Основним результатом даної статті є теорема про асимптотичну нормальність емпіричного процесу

$$U_{k;n} = \sqrt{n}\left(\widehat{F}_{k;n}(t) - F_k(t)\right) \quad (7)$$

як елемента $D[0,T]$ — простору функцій без стрибків другого роду на $[0,T]$, де $T$ — будь-яке число на $]0, +\infty[$, таке, що $F_m(T) < 1$, $G_m(T) < 1$ для всіх $m = 1, \dots, M$.

Щоб описати граничний гауссів процес нам будуть потрібні додаткові позначення.

Уведемо $\xi_{(m)}, C_{(m)}$ — незалежні випадкові величини з розподілами, відповідно $F_m$ і $G_m$. Ці величини можна трактувати як тривалість життя та момент цензурування для об'єкта, вибраного навмання з $m$-го компонента. Тоді $\xi^*_{(m)} = \min\left(\xi_{(m)}, C_{(m)}\right)$ — цензурована тривалість життя, $\delta^*_{(m)} = \mathbb{1}\left\{\xi_{(m)} < C_{(m)}\right\}$ — індикатор відсутності цензурування для об'єкта, вибраного навмання з $m$-го компонента.

Функцію виживання для цензурованої тривалості життя об'єктів з $m$-го компонента позначимо

$$Y_m(t) = \mathsf{P}\left\{\xi^*_{(m)} \ge t\right\} = \bar{G}_m(t-)\bar{F}_m(t-).$$

Крім того, уведемо позначення

$$N_m(t) = \mathsf{P}\left\{\xi^*_{(m)} \le t, \delta_{(m)} = 1\right\} = \int_{]0,t]} \bar{G}_m(s-)\, F_m(ds).$$

Тоді

$$\Lambda_m(t) = \int_{]0,t]} \frac{N_m(dt)}{Y_m(t)}$$

— це інтегральна інтенсивність смертності для об'єктів з $m$-го компонента. Далі буде використане позначення

$$R_m(A) = N_m(A) - \int_A Y_m(t)\, \frac{N_k(dt)}{Y_k(t)}.$$

Помітимо, що $R_m(A)$ залежить також від $k$ — номера компонента, для якого будується оцінка, але ми не будемо вказувати цю залежність явно для спрощення позначень.

Визначимо тепер функцію $\rho$ що описує коваріації приростів гауссового процесу $Z_k$, який буде використано для опису границі $U_{k;n}$. Для довільних $0 \le u_1 < u_2 \le \le u_3 < u_4$ позначимо $A_1 = ]u_1, u_2]$, $A_2 = ]u_3, u_4]$.

Задамо функцію $\rho$ таким чином:

$$\rho(A_1, A_2) = \sum_{m=1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^m \right\rangle \left[ \int_{A_2} Y_m(t_2) \Lambda_k(dt_2) \int_{A_1} \Lambda_k(dt_1) - N_m(A) \int_{A_1} \Lambda_k(dt) \right] - \\ - \sum_{m_1, m_2 = 1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^{m_1} \mathbf{p}^{m_2} \right\rangle R_{m_1}(A_1) R_{m_2}(A_2) \tag{8}$$

і

$$\rho(A_1, A_1) = \sum_{m=1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^m \right\rangle \left[ N_m(A_1) - 2 \int_{A_1} N_m(]t, u_2]) \Lambda_k(dt) + \right. \\ \left. + \int_{A_1 \times A_1} Y_m\big(\max(t_1, t_2)\big) \Lambda_k(dt_1) \Lambda_k(dt_2) \right] - \\ - \sum_{m_1, m_2 = 1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^{m_1} \mathbf{p}^{m_2} \right\rangle R_{m_1}(A_1) R_{m_2}(A_1). \tag{9}$$

Визначимо $Z_k(t)$ як гауссів випадковий процес із траєкторіями, що належать $D[0,T]$, із нульовим математичним сподіванням та

$$\mathsf{E}\, Z_k(A_1) Z_k(A_2) = \rho(A_1, A_2), \quad \mathsf{E}\big(Z_k(A_1)\big)^2 = \rho(A_1, A_1).$$

Для визначеності будемо вважати, що $Z_k(0) = 0$. Отже, для $u_1 < u_2$,

$$\operatorname{Cov}\big(Z_k(u_1), Z_k(u_2)\big) = C_Z(u_1, u_2) = \mathsf{E}\, Z_k(]0, u_1]) Z_k(]0, u_2]) = \\ = \mathsf{E}\big(Z_k(]0, u_1])\big)^2 + \mathsf{E}\, Z_k(]0, u_1]) Z_k(]u_1, u_2]) = \\ = \rho(]0, u_1], ]0, u_1]) + \rho(]0, u_1], ]u_1, u_2]). \tag{10}$$

Процес $Z_k$ можна також описати як гауссів процес із нульовим математичним сподіванням та коваріаційною функцією, заданою (10). Зрозуміло, що це означення буде змістовним лише тоді, коли (10) визначає якусь справжню коваріаційну функцію. Ми покажемо, що в умовах теореми 3.1 це дійсно так.

**Теорема 3.1.** *Нехай*

1. $\det \mathbf{\Gamma} \ne 0$.
2. $\left\langle (\mathbf{a}^k) \mathbf{p}^{m_1} \mathbf{p}^{m_2} \right\rangle$ *існують для всіх* $m_1, m_2 = 1, \dots, M$.

*3. Для всіх* $m = 1, \dots, M$, $F_m(T) < 1$, $G_m(T) < 1$.
*4. Для всіх* $m = 1, \dots, M$, *функції* $F_m$ *і* $G_m$ *є неперервно диференційовними на* $[0, T]$.

*Тоді емпіричні процеси* $U_{k;n}$, *визначені формулою (7), слабко збігаються при* $n \to \infty$ *у* $D[0, T]$ *із рівномірною нормою до граничного процесу* $U_k$, *визначеного як*

$$U_k(t) = \big(1 - F_k(t)\big) \int_{]0,t]} \frac{Z_k(du)}{Y_k(u)}. \tag{11}$$

*Зауваження* 3.1. У припущеннях теореми $Z_k(t)$ є неперервним майже напевно на $[0, T]$, і функція $1/Y_k(u)$ має обмежену варіацію. Тому інтеграл у (11) можна розглядати як потраєкторний інтеграл Рімана–Стілтьєса. З іншого боку, його можна також визначити як границю у середньому квадратичному інтегральних сум (quadratic mean, QM-інтеграл). Обидві інтерпретації приводять до того ж самого розподілу $U_k(t)$.

QM-інтегральна інтерпретація (11) дозволяє обчислювати дисперсію та коваріаційну функцію $U_k(t)$. Наприклад, у наслідку 3.1 отримано аналог асимптотичної версії формули Грінвуда.

Позначимо

$$\sigma_t^2 = \big(1 - F_k(t)\big)^2 \left[ \sum_{m=1}^{M} \left( \big\langle (\mathbf{a}^k)^2 \mathbf{p}^m \big\rangle \int_0^t \frac{N'_m(u)\, du}{(Y_k(u))^2} \right) + \iint_{S_t} \frac{\partial^2 \rho(u_1, u_2)}{\partial u_1 \partial u_2} \frac{du_1\, du_2}{Y_k(u_1) Y_k(u_2)} \right], \tag{12}$$

де $S_t = \{(u_1, u_2) \in [0, t],\ u_1 \neq u_2\}$, $N'_m(u) = dN_m(u)/du$.

**Наслідок 3.1.** *Нехай виконані умови теореми 3.1. Тоді для будь-якого* $t \in [0, T]$ *розподіл* $U_{k;n}(t)$ *слабко збігається при* $n \to \infty$ *до нормального розподілу з нульовим математичним сподіванням і дисперсією* $\sigma_t^2$.

## 4. Доведення

Доведення теореми ґрунтується на ідеях, використаних у [5] для отримання асимптотичної нормальності КМЕ для випадку однорідних цензурованих даних. Ми почнемо з двох технічних лем.

Нехай $T$ — довільне фіксоване додатне число. Для функцій $f\colon [0, T] \to \mathbb{R}$ позначимо $\|f\|_\infty = \sup_{t \in [0,T]} |f(t)|$ — рівномірна норма, $\|f\|_V$ — варіація $f$ на $[0, T]$.

**Лема 4.1.** *Нехай* $G_n$, $G$, $F_n$ *і* $F$ *— деякі функції з* $[0, T]$ *у* $\mathbb{R}$, *що мають обмежену варіацію і задовольняють наступні умови:*

1. *Існує таке* $K_F < \infty$, *що* $\|F_n\|_V < K_F$ *для всіх* $n = 1, 2, \dots$
2. *Існують неперервні функції* $f, g\colon [0, T] \to \mathbb{R}$, *такі, що* $g_n = \sqrt{n}(G_n - G) \to g$ *і* $f_n = \sqrt{n}(F_n - F) \to f$ *при* $n \to \infty$ *у* $\|\cdot\|_\infty$*-нормі.*

*Тоді* $\|I_n - I\|_\infty \to 0$ *при* $n \to \infty$, *де*

$$I_n(t) = \sqrt{n} \left( \int_0^t G_n(u) F_n(du) - \int_0^t G(u) F(du) \right),$$
$$I(t) = \int_0^t G(u) f(du) + \int_0^t g(u) F(du).$$

*Зауваження* 4.1. Ми не припускаємо тут, що $f$ має обмежену варіацію, але $f$ є неперервною і $\|G\|_V < \infty$. Тому $\int_0^t G(u) f(du)$ існує як границя інтегральних сум Рімана–Стілтьєса і дорівнює $G(u) f(u)\big|_0^t - \int_0^t f(u) G(du)$.

Існування інтегралів $\int_0^T G_n(u)F_n(du)$ як єдиних границь інтегральних сум Ріма-на–Стілтьєса не гарантоване умовами леми, але її твердження є правильним для будь-яких часткових границь таких сум.

*Доведення.* Розглянемо спочатку

$$J_n(t) = \sqrt{n}\int_0^t \big(G_n(u) - G(u)\big)\big(F_n(du) - F(du)\big).$$

Покажемо, що

$$\|J_n\|_\infty \to 0 \quad \text{при } n \to \infty. \tag{13}$$

Для будь-якого $m$

$$\begin{aligned}|J_n(t)| &= \left|\int_0^t g_n(u)\big(F_n(du) - F(du)\big)\right| \le \\ &\le \left|\int_0^t \big(g_n(u) - g_m(u)\big)\big(F_n(du) - F(du)\big)\right| + \left|\int_0^t g_m(u)\big(F_n(du) - F(du)\big)\right| \le \\ &\le \|g_n - g_m\|_\infty \|F_n - F\|_V + \big(2\|g_m\|_\infty + \|g_m\|_V\big)\|F_n - F\|_\infty .\end{aligned}$$

Оцінюючи другий доданок в останній нерівності, ми інтегрували частинами:

$$\int_0^t g_m(u)\big(F_n(du) - F(du)\big) = g_m(u)\big(F_n(u) - F(u)\big)\Big|_0^t - \int_0^t \big(F_n(u) - F(u)\big)g_m(du).$$

Спрямувавши $n$ до нескінченності, отримуємо

$$\limsup_{n\to\infty}\|J_n\|_\infty \le (K_F + \|F\|_V)\limsup_{n\to\infty}\|g_n - g_m\|_\infty ,$$

оскільки $\|g_m\|_\infty < \infty$, $\|g_m\|_V < \infty$, для фіксованого $m$ і, за другим припущенням леми, дістаємо

$$\|F_n - F\|_\infty = \frac{1}{\sqrt{n}}\|f_n\|_\infty \to 0 \quad \text{при } n \to \infty.$$

Враховуючи $\|g_n - g_m\|_\infty \le \|g_n - g\|_\infty + \|g_m - g\|_\infty$, маємо

$$\limsup_{n\to\infty}\|J_n\|_\infty \le (K_F + \|F\|_V)\|g_m - g\|_\infty$$

для всіх натуральних $m$. Спрямувавши $m \to \infty$, отримуємо (13).

Помітимо, що

$$I_n(t) = I_n^1(t) + I_n^2(t), \tag{14}$$

де

$$\begin{aligned}I_n^1(t) &= \sqrt{n}\int_0^t \big(G_n(u) - G(u)\big)F_n(du), \\ I_n^2(t) &= \sqrt{n}\int_0^t G(u)\big(F_n(du) - F(du)\big).\end{aligned}$$

Тому

$$\begin{aligned}I_n^2 = \int_0^t G(u)f_n(du) = G(u)f_n(u)\Big|_0^t - \int_0^t f_n(u)G(du) \to \\ \to G(u)f(u)\Big|_0^t - \int_0^t f(u)G(du) = \int_0^t G(u)f(du)\end{aligned} \tag{15}$$

при $n \to \infty$ рівномірно по $t \in [0, T]$. Із (13) одержуємо також

$$I_n^1(t) = J_n(t) + \int_0^t g_n(u)F(du) \to \int_0^t g(u)F(du). \tag{16}$$

Враховуючи (14)–(16), отримуємо твердження леми. □

Позначимо

$$\zeta_j(A) = \mathbb{1}\left\{\xi^*_{j;n} \in A, \delta_{j;n} = 1\right\} - \int_A \mathbb{1}\left\{\xi^*_{j;n} > t\right\} \Lambda_k(dt),$$

$$Z_{k;n}(A) = \frac{1}{\sqrt{n}} \sum_{j=1} a^k_{j;n} \zeta_j(A).$$

Для $t > 0$ будемо позначати $Z_{k;n}(t) = Z_{k;n}(]0, t])$.

**Лема 4.2.** *У припущенях теореми 3.1 процес $Z_k(t)$ існує, має неперервні траєкторії на $[0,T]$, і процеси $Z_{k;n}(t)$ слабко збігаються до $Z_k(t)$ у $D[0,T]$ із нормою $\|\cdot\|_\infty$ при $n \to \infty$.*

*Доведення.* Розглянемо випадковий вектор $\left(\xi_{(m)}, \delta_{(m)}, \xi^*_{(m)}\right)$, розподіл якого тотожний умовному розподілу $\big(\xi(O), \delta(O), \xi^*(O)\big)$ за умови $\mathrm{ind}(O) = m$.

Помітимо, що

$$\mathsf{E}\, Z_{k;n}(A) = \frac{1}{n} \sum_{j=1}^{n} \sum_{m=1}^{M} a^k_{j;n} \left( \mathsf{P}\left\{\xi^*_{(m)} \in A, \delta_{(m)} = 1\right\} - \int_A \mathsf{P}\left\{\xi^*_{(m)} > t\right\} \Lambda_k(dt) \right) =$$

$$= \sum_{m=1}^{M} \left\langle \mathbf{a}^k \mathbf{p} \right\rangle_n \left( N_m(A) - \int_A Y_m(t) \frac{N_m(dt)}{Y_k(t)} \right) = 0,$$

оскільки $\left\langle \mathbf{a}^k \mathbf{p} \right\rangle_n = \mathbb{1}\{k = m\}$.

Безпосередніми обчисленнями отримується

$$\mathsf{E}\, Z_{k;n}(A_1) Z_{k;n}(A_2) = \rho_n(A_1, A_2), \quad \mathsf{E}(Z_{k;n}(A_1))^2 = \rho_n(A_1, A_1),$$

де

$$\rho_n(A_1, A_2) = \sum_{m=1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^m \right\rangle_n \left[ \int_{A_2} Y_m(t_2) \Lambda_k(dt_2) \int_{A_1} \Lambda_k(dt_1) - N_m(A) \int_{A_1} \Lambda_k(dt) \right] -$$

$$- \sum_{m_1, m_2 = 1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^{m_1} \mathbf{p}^{m_2} \right\rangle_n R_{m_1}(A_1) R_{m_2}(A_2) \tag{17}$$

і

$$\rho_n(A_1, A_1) = \sum_{m=1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^m \right\rangle_n \left[ N_m(A_1) - 2 \int_{A_1} N_m(]t, u_2]) \Lambda_k(dt) + \right.$$

$$\left. + \int_{A_1 \times A_1} Y_m\big(\max(t_1, t_2)\big) \Lambda_k(dt_1) \Lambda_k(dt_2) \right] -$$

$$- \sum_{m_1, m_2 = 1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^{m_1} \mathbf{p}^{m_2} \right\rangle_n R_{m_1}(A_1) R_{m_2}(A_1). \tag{18}$$

Наприклад, щоб отримати (18), розглянемо

$$\mathsf{E}(Z_{k;n}(A_1))^2 = \frac{1}{n} \sum_{j=1}^{n} \big(a^k_{j;n}\big)^2 \mathsf{E}\big(\zeta_j(A_1) - \mathsf{E}\, \zeta_j(A_1)\big)^2 =$$

$$= \sum_{m=1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^m \right\rangle_n \mathsf{E}\left(\zeta_{(m)}(A_1)\right)^2 -$$

$$- \sum_{m_1, m_2 = 1}^{M} \left\langle (\mathbf{a}^k)^2 \mathbf{p}^{m_1} \mathbf{p}^{m_2} \right\rangle_n \mathsf{E}\, \zeta_{(m_1)}(A_1)\, \mathsf{E}\, \zeta_{(m_2)}(A_1),$$

де

$$\zeta_{(m)}(A) = \mathbb{1}\left\{\xi^*_{(m)} \in A, \delta_{(m)} = 1\right\} - \int_A \mathbb{1}\left\{\xi^*_{(m)} > t\right\} \Lambda_k(dt).$$

Вочевидь

$$\mathsf{E}\,\zeta_{(m)}(A_1) = R_m(A_1),$$

$$\begin{aligned}\mathsf{E}(\zeta_{(m)}(A_1))^2 &= \mathsf{E}\,\mathbb{1}\left\{\xi^*_{(m)} \in A_1, \delta_{(m)} = 1\right\} - \\ &\quad - 2\,\mathsf{E}\int_{A_1} \mathbb{1}\left\{\xi^*_{(m)} \in A_1, \delta_{(m)} = 1\right\} \mathbb{1}\left\{\xi^*_{(m)} > t\right\} \Lambda_k(dt) + \\ &\quad + \mathsf{E}\int_{A_1}\int_{A_1} \mathbb{1}\left\{\xi^*_{(m)} > t_1\right\} \mathbb{1}\left\{\xi^*_{(m)} > t_2\right\} \Lambda_k(dt_1)\Lambda_k(dt_2) = \\ &= N_m(A_1) - 2\int_{A_1} N_m(]t, u_2])\Lambda_k(dt) + \\ &\quad + \int_{A_1\times A_1} Y_m\big(\max(t_1,t_2)\big)\Lambda_k(dt_1)\Lambda_k(dt_2).\end{aligned}$$

Звідси випливає (18). Рівність (17) можна отримати аналогічно.

Отже, в умовах теореми 3.1, коваріаційна функція $Z_{k;n}$ збігається до функції $C_Z(u_1,u_2)$, визначеної (10). Тому $C_Z(u_1,u_2)$ є коваріаційною функцією деякого гауссового процесу на $[0,T]$. У припущеннях теореми 3.1, функції $N_m(t)$, $m = 1,\dots,M$ є неперервно диференційовними і функція $Y_k(t)$ відокремлена від нуля на $[0,T]$. Тому за (9)

$$\mathsf{E}\big(Z_k(t_2) - Z_k(t_1)\big)^2 \le K(t_2 - t_1)^2$$

для деякого $K < \infty$ і всіх $t_1, t_2 \in [0,T]$. Використовуючи теорему Колмогорова [4, теорема 7, п. 5 розділу III], отримуємо, що гауссів процес $Z_k(t)$ можна задати неперервним майже напевно.

Асимптотичну нормальність скінченновимірних розподілів $Z_{k;n}(t)$ при $n \to \infty$ можна довести, використовуючи центральну граничну теорему з умовою Ліндеберга, так, як це зроблено у [10] для навантажених емпіричних функцій. За (17)

$$\mathsf{E}\big(Z_{n;k}(t_3) - Z_{n;k}(t_2)\big)\big(Z_{n;k}(t_2) - Z_{n;k}(t_1)\big) \le K(t_3 - t_2)(t_2 - t_1) \le K(t_3 - t_1)^2$$

для деякого $K < \infty$ і всіх $0 \le t_1 < t_2 < t_3 \le T$. Звідси, з урахуванням збіжності скінченновимірних розподілів, випливає збіжність $Z_{k;n}$ до $Z_k$ у $D[0,T]$ у метриці Скорохода [2, теорема 13.5]. Оскільки граничний процес $Z_k$ є майже напевно неперервним, зі слабкої збіжності у метриці Скорохода випливає слабка збіжність у рівномірній метриці. □

*Начерк доведення теореми 3.1.* Позначимо

$$\begin{aligned}\widehat{\Lambda}_{k;n}(t) &= \int_0^t \frac{\widehat{N}_{k;n}(du)}{\widehat{Y}_{k;n}(u)}, \\ Z^Y_{k;n}(t) &= \sqrt{n}\left(\widehat{Y}_{k;n}(t) - Y_k(t)\right), \\ Z^N_{k;n}(t) &= \sqrt{n}\left(\widehat{N}_{k;n}(t) - N_k(t)\right).\end{aligned}$$

Тоді

$$Z_{k;n}(t) = \int_0^t \left(Z^N_{k;n}(du) - Z^Y_{k;n}(u)\Lambda_k(du)\right).$$

Так само, як у лемі 4.2, можна показати, що процес $\big(Z^Y_{k;n}(\cdot), Z^N_{k;n}(\cdot)\big)$ слабко збігається у просторі функцій $[0,T] \to \mathbb{R}^2$ без розривів другого роду у рівномірній нормі до деякого процесу $\big(Z^Y_k(\cdot), Z^N_k(\cdot)\big)$ з майже напевно неперервними траєкторіями.

Отже, застосовуючи лему 4.1 і метод одного ймовірнісного простору, отримуємо, що

$$\sqrt{n}\left(\widehat{\Lambda}_{k;n}(t)-\Lambda_k(t)\right)=\sqrt{n}\left(\int_0^t\frac{1}{\widehat{Y}_{k;n}(u)}\widehat{N}_{k;n}(du)-\int_0^t\frac{1}{Y_k(u)}N_k(du)\right)$$

збігається слабко на $[0,T]$ до

$$\int_0^t\left(\frac{1}{Y_k(u)}Z_k^N(du)-\frac{Z_k^Y(u)N_k(du)}{(Y_k(u))^2}\right)=\int_0^t\frac{\widetilde{Z}_k(du)}{Y_k(u)},$$

де

$$\widetilde{Z}_k(t)=Z_k^N(t)+\int_0^t Z_k^Y(u)\Lambda_k(du).$$

Легко бачити, що $\widetilde{Z}_k(t)$ є границею $Z_{k;n}(t)$, отже, за лемою 4.2, $\widetilde{Z}_k(t)=Z_k(t)$.

Використовуючи рівність (66) із [5], отримуємо

$$\sqrt{n}\left(\widehat{F}_{k;n}(t)-F_k(t)\right)=\big(1-F_k(t)\big)\int_0^t\frac{1-\widehat{F}_{k;n}(u-)}{1-\widehat{F}_{k;n}(u)}\sqrt{n}\left(\widehat{\Lambda}_{k;n}-\Lambda_k\right)(du).$$

Оскільки $\left\|\widehat{F}_{k;n}-F_k\right\|_\infty\le K\sqrt{\log(n)/n}$ [11], то маємо

$$\frac{1-\widehat{F}_{k;n}(u-)}{1-\widehat{F}_{k;n}(u)}\to 1$$

і

$$\sqrt{n}\left(\widehat{F}_{k;n}(t)-F_k(t)\right)\xrightarrow{\text{w}}\big(1-F(t)\big)\int_0^t\frac{Z_k(du)}{Y_k(u)}$$

при $n\to\infty$ у рівномірній нормі. (Для строгого доведення цієї збіжності потрібно застосувати техніку, подібну до використаної у доведенні леми 4.1.) □

*Доведення наслідку 3.1.* Будемо розглядати інтеграл у (11) як границю у середньому квадратичному інтегральних сум вигляду

$$J(\mathcal{T})=\sum_{i=1}^I\frac{Z_k(]t_{i-1},t_i])}{Y_k(t_i)},$$

коли $\operatorname{diam}(\mathcal{T})\to 0$, де $\mathcal{T}=\{0=t_0<t_1<\dots<t_I=T\}$ — деяке невипадкове розбиття $[0,1]$, $\operatorname{diam}(\mathcal{T})=\sup_{i=1,\dots,I}(t_i-t_{i-1})$. Такі QM-інтеграли зазвичай визначають для $Z_k$, що є процесами з незалежними приростами. У нашому випадку це не так. Теорія QM-інтегрування за процесами із залежними приростами описана у [7, розділ 37.3]. Із неї, зокрема, випливає, що

$$\operatorname{Var}\left(\int_0^T\frac{Z_k(dt)}{Y_k(t)}\right)=\lim_{\operatorname{diam}(\mathcal{T})\to 0}\sum_{i\ne j}\frac{\mathsf{E}\,Z_k(]t_{i-1},t_i])Z_k(]t_{j-1},t_j])}{Y_k(t_i)Y_k(t_j)}+$$
$$+\lim_{\operatorname{diam}(\mathcal{T})\to 0}\sum_{i=1}^I\frac{\mathsf{E}\big(Z_k(]t_{i-1},t_i])\big)^2}{(Y_k(t_i))^2}.\tag{19}$$

В умовах теореми 3.1

$$\mathsf{E}\,Z_k(]t_{i-1},t_i])Z_k(]t_{j-1},t_j])=\frac{\partial^2\rho(t_i,t_j)}{\partial t_i\partial t_j}+o\left(\operatorname{diam}(\mathcal{T})^2\right)$$

і

$$\mathsf{E}\big(Z_k(]t_{i-1},t_i])\big)^2=\sum_{m=1}^M\left\langle(\mathbf{a}^k)^2\mathbf{p}^m\right\rangle N_m'(t_i)+o\big(\operatorname{diam}(\mathcal{T})\big).$$

Звідси випливає (12). □

Автор вдячний рецензенту за увагу до цієї публікації і слушні зауваження.

## Список літератури


1. F. Autin and C. Pouet, *Minimax rates over Besov spaces in ill-conditioned mixture-models with varying mixing-weights*, J. Statist. Plann. Inference **146** (2014), 20–30.
2. P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1999.
3. O. V. Doronin, *Adaptive estimation for a semiparametric model of mixture*, Theory Probab. Math. Statist. **91** (2015), 29–41.
4. I. I. Gihman and A. V. Skorohod, *The Theory of Stochastic Processes*, vol. 1, Nauka, Moscow, 1971; English transl., Springer-Verlag, Berlin–Heidelberg–New York, 1974.
5. R. D. Gill and S. Johansen, *A survey of product-integration with a view toward application in survival analysis*, Ann. Statist. **18** (1990), no. 4, 1501–1555.
6. J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, Second edition, Wiley, 2002.
7. M. Loéve, *Probability Theory II*, Forth edition, Springer, New York, 1978.
8. R. E. Maiboroda, *On the estimation of parameters of variable mixtures*, Theory Probab. Math. Statist. **44** (1991), 87–92.
9. R. E. Maiboroda, *Estimates for distributions of components of mixtures with varying concentrations*, Ukrainian Math. J. **48** (1996), no. 4, 618–622.
10. R. Maiboroda and O. Sugakova, *Statistics of mixtures with varying concentrations with application to DNA microarray data analysis*, J. Nonparametr. Stat. **24** (2012), no. 1, 201–215.
11. R. E. Maiboroda, V. G. Khizanov, *A modified Kaplan–Meier estimator for a model of mixtures with varying concentrations*, Theory Probab. Math. Statist. **92** (2016), 109–116.
12. A. Yu. Ryzhov, *Estimates of distributions of components in a mixture from censoring data*, Teor. Imovir. Mat. Statist. **69** (2003), 154–161; English transl. in Theory Probab. Math. Statist. **69** (2004), 167–174.
13. J. Shao, *Mathematical Statistics*, Springer-Verlag, New York, 2003.


Кафедра теорії ймовірностей, статистики та актуарної математики, механіко-математичний факультет, Київський національний університет імені Тараса Шевченка, вул. Володимирська, 64/13, м. Київ, Україна, 01601

*Адреса електронної пошти*: mre@univ.kiev.ua

Стаття надійшла до редколегії 21.12.2016

## ASYMPTOTIC NORMALITY OF KAPLAN – MEIER ESTIMATOR FOR MIXTURES WITH VARYING CONCENTRATIONS

R. MAIBORODA

Abstract. A modification of Kaplan–Meier estimator is considered for mixture components CDFs estimation by censored data in the case when mixing probabilities vary from observation to observation. Asymptotic normality of the estimators in the sup-norm is demonstrated.

## АСИМПТОТИЧЕСКАЯ НОРМАЛЬНОСТЬ ОЦЕНОК КАПЛАНА – МЕЙЕРА ДЛЯ СМЕСЕЙ С ПЕРЕМЕННЫМИ КОНЦЕНТРАЦИЯМИ

Р. МАЙБОРОДА

Аннотация. Рассматривается модификация оценки Каплана – Мейера для распределения компонентов смеси с переменными концентрациями по цензурированным наблюдениям. Доказана асимптотическая нормальность этих оценок в равномерной норме.