

УДК 519.233.2:519.254

О. М. Мацуга, Г. С. Шубіна

Дніпропетровський національний університет імені Олеся Гончара

ОБЧИСЛЮВАЛЬНА СХЕМА ІДЕНТИФІКАЦІЇ РОЗПОДІЛІВ З КЛАСУ НОРМАЛЬНОГО

Запропоновано обчислювальну схему ідентифікації розподілів з класу нормального, яка відрізняється від існуючих можливістю ідентифікації сплайн-розподілу шляхом автоматизованого пошуку вузлів сплайна на ймовірнісному папері.

Ключові слова: ідентифікація, ймовірнісний папір, обчислювальна схема, сплайн-нормальний розподіл, вузол склеювання.

Предложена вычислительная схема идентификации распределений из класса нормального, которая отличается от существующих возможностью идентификации сплайн-распределения путем автоматизированного поиска узлов сплайна на вероятностной бумаге.

Ключевые слова: идентификация, вероятностная бумага, вычислительная схема, сплайн-нормальное распределение, узел склеивания.

A computing scheme for identification and restoration of distribution from normal distribution class is offered. It is notable for identification of a spline-normal distribution by automatic search of spline nodes on a probability paper.

Key words: identification, probability paper, computing scheme, spline-normal distribution, node.

Постановка проблеми. В інформаційному забезпеченні обробки неоднорідних статистичних даних нині актуальна проблема автоматизованої ідентифікації моделі розподілу за вибірковими даними. До моделей, які описують розподіл неоднорідних даних, відносяться суміші та сплайн-розподіли. Проблема їх ідентифікації лише частково вирішена для сумішей розподілів [1 – 4]. Що стосується сплайн-розподілів, єдиний спосіб їх ідентифікації полягає у візуальному аналізі ймовірнісного паперу, в той час як є необхідність в автоматизованій ідентифікації.

У роботі для вирішення проблеми ідентифікації сплайн-розподілів пропонується на ймовірнісному папері в автоматизованому режимі шукати точки, в яких пряма змінює кут нахилу. Такі точки можна

© О. М. Мацуга, Г. С. Шубіна, 2011

тракувати як вузли склеювання, і тим самим ідентифікувати сплайн-розподіл із заданою кількістю вузлів.

Для відпрацювання запропонованої схеми ідентифікації обрано клас нормального розподілу як найбільш поширений у прикладних задачах.

Аналіз останніх досліджень і публікацій з вирішення проблеми ідентифікації моделей розподілів виявив наявність обчислювальних схем моментної, етропійної та квантильної ідентифікації [1; 2; 5]. Однак вони реалізуються для конкретних типів розподілів і здебільшого придатні під час ідентифікації за однорідними даними. Найбільш потужна є ідентифікація на основі ймовірнісного паперу.

Постановка задачі. Нехай результати спостережень задано у вигляді вибірки $\{x_i; i = \overline{1, N}\}$, де N – кількість спостережень; x_i – спостережуване значення в i -му експерименті. На основі вибірки побудовано варіаційний ряд $\{x_i, n_i, p_i; i = \overline{1, r}\}$, де $x_1 < x_2 < \dots < x_r$; r – кількість варіант; x_i – значення i -ї варіанти; n_i – частота i -ї варіанти; $\sum_{i=1}^r n_i = N$;

$p_i = \frac{n_i}{N}$ – відносна частота i -ї варіанти; $\sum_{i=1}^r p_i = 1$. У кожній варіанті роз-

раховано значення емпіричної функції розподілу $F_N(x_i) = \sum_{j=1}^i p_j$, $i = \overline{1, r}$.

Припускається, що на основі емпіричної функції може бути ідентифіковано розподіл з класу нормального, тобто «чистий» нормальний розподіл або сплайн-нормальний розподіл.

Необхідно розробити обчислювальну схему автоматизованої ідентифікації одного із зазначених розподілів.

Основний матеріал. Пропонується обчислювальна схема ідентифікації та відтворення розподілів з класу нормального. Її відмінність полягає в можливості ідентифікації сплайн-нормального розподілу шляхом автоматизованого пошуку вузлів сплайна на ймовірнісному папері.

Обчислювальна схема складається з таких етапів.

1. Здійснюється момента ідентифікація нормального розподілу на основі коефіцієнтів асиметрії та ексцесу [5], яка зводиться до перевірки основних гіпотез

$$H_0 : A = 0, \quad H_0 : E = 0$$

за відповідних альтернатив $H_1 : A \neq 0$, $H_1 : E \neq 0$. Для перевірки основних гіпотез вводяться статистики

$$u_A = \frac{\hat{A}}{\sigma\{\hat{A}\}}, \quad u_E = \frac{\hat{E}}{\sigma\{\hat{E}\}},$$

де \hat{A} та \hat{E} – оцінки коефіцієнтів асиметрії та ексцесу вибірки відповідно; $\sigma\{\hat{A}\}$ та $\sigma\{\hat{E}\}$ – середньоквадратичні відхилення зазначених оцінок відповідно.

У разі одночасного виконання умов

$$|u_A| \leq u_{\alpha/2}, \quad |u_E| \leq u_{\alpha/2},$$

де $u_{\alpha/2}$ – двобічний квантиль стандартного нормального розподілу, ідентифікується нормальний розподіл і здійснюється перехід до п. 2. В іншому разі відбувається перехід до п. 3.

2. Відтворюється нормальний розподіл за вибірковими даними [5] і обчислення завершуються.

3. Визначається кількість вузлів склеювання та самі вузли сплайн-нормального розподілу за однією з обчислювальних схем 1 – 3.

4. Відтворюється сплайн-нормальний розподіл з ідентифікованими вузлами склеювання [3], після чого всі розрахунки завершуються.

Загальна ідея запропонованих обчислювальних схем 1 – 3 пошуку вузлів склеювання сплайн-розподілу однакова. На ймовірнісному папері фіксується варіанта x_i варіаційного ряду і будуються два відрізки, що перетинаються в цій варіанті. Якщо на ймовірнісному папері пряма у варіанті x_i не змінює кут нахилу, то побудовані відрізки також будуть мати однакові кути нахилу, інакше – нахил побудованих відрізків буде різнитися, що свідчитиме про наявність вузла склеювання у варіанті x_i . Пропоновані обчислювальні схеми різняться способом установаження відмінності кутів нахилу відрізків. Запропоновано три такі способи.

Для побудови ймовірнісного паперу нормального розподілу функція розподілу

$$F(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{1}{2}\left(\frac{y-m}{\sigma}\right)^2\right) dy$$

зводиться до лінійного вигляду шляхом перетворення відносно квантилів:

$$x = m + \sigma u,$$

де u – одnobічний квантиль стандартного нормального розподілу.

На ймовірнісному папері відображується лінеаризована емпірична функція розподілу у вигляді масиву $\{x_i, u_i; i = \overline{1, r}\}$, де x_i – варіанта варіаційного ряду; u_i – однобічний квантиль стандартного нормального розподілу порядку $F_N(x_i)$; $F_N(x_i)$ – значення емпіричної функції розподілу у варіанті x_i .

1. Пошук вузлів склеювання сплайн-розподілу на основі вимірювання кутів.

Фіксується варіанта x_t і формуються масиви:

$$\{x_i, u_i; i = \overline{t-l, t}\}, \quad (1)$$

$$\{x_i, u_i; i = \overline{t, t+l}\}, \quad (2)$$

за якими відтворюються відповідні лінійні регресії:

$$u = k_1 x + b_1, \quad (3)$$

$$u = k_2 x + b_2. \quad (4)$$

Параметри k_1 та k_2 цих моделей являють собою тангенси кутів нахилу регресій до вісі абсцис, тому в якості різниці між кутами нахилу регресій можна ввести показник

$$K = |\arctg |k_2| - \arctg |k_1||. \quad (5)$$

Ті варіанти, в яких значення показника K вищі за деякий поріг, доцільно вважати вузлами склеювання сплайн-розподілу.

Відтворення лінійних регресій (3) та (4) здійснюється методом найменших квадратів шляхом мінімізації залишкових дисперсій [6]:

$$S_1^2 = \sum_{i=t-l}^t (u_i - k_1 x_i - b_1)^2, \quad (6)$$

$$S_2^2 = \sum_{i=t}^{t+l} (u_i - k_2 x_i - b_2)^2. \quad (7)$$

Умова мінімуму залишкових дисперсій (6) та (7) приводить до розв'язання систем лінійних алгебричних рівнянь

$$\begin{cases} b_1 + k_1 \bar{x}^{(1)} = \bar{u}^{(1)}, \\ b_1 \bar{x}^{(1)} + k_1 \bar{x}^2{}^{(1)} = \overline{ux}^{(1)}, \end{cases} \quad \begin{cases} b_2 + k_2 \bar{x}^{(2)} = \bar{u}^{(2)}, \\ b_2 \bar{x}^{(2)} + k_2 \bar{x}^2{}^{(2)} = \overline{ux}^{(2)}, \end{cases}$$

з яких одержують оцінки параметрів:

$$\begin{aligned} \hat{k}_1 &= r_{x,u}^{(1)} \frac{S_u^{(1)}}{S_x^{(1)}}, & \hat{k}_2 &= r_{x,u}^{(2)} \frac{S_u^{(2)}}{S_x^{(2)}}, \\ \hat{b}_1 &= \bar{u}^{(1)} - \hat{k}_1 \bar{x}^{(1)}, & \hat{b}_2 &= \bar{u}^{(2)} - \hat{k}_2 \bar{x}^{(2)}, \end{aligned} \quad (8)$$

де

$$\begin{aligned} r_{x,u}^{(1)} &= \frac{\overline{xu}^{(1)} - \bar{x}^{(1)}\bar{u}^{(1)}}{S_x^{(1)}S_u^{(1)}}; & r_{x,u}^{(2)} &= \frac{\overline{xu}^{(2)} - \bar{x}^{(2)}\bar{u}^{(2)}}{S_x^{(2)}S_u^{(2)}}; \\ \bar{x}^{(1)} &= \frac{1}{l+1} \sum_{i=t-l}^t x_i; & \bar{x}^{(2)} &= \frac{1}{l+1} \sum_{i=t}^{t+l} x_i; \\ \bar{u}^{(1)} &= \frac{1}{l+1} \sum_{i=t-l}^t u_i; & \bar{u}^{(2)} &= \frac{1}{l+1} \sum_{i=t}^{t+l} u_i; \\ \overline{xu}^{(1)} &= \frac{1}{l+1} \sum_{i=t-l}^t x_i u_i; & \overline{xu}^{(2)} &= \frac{1}{l+1} \sum_{i=t}^{t+l} x_i u_i; \\ S_x^{(1)} &= \frac{1}{l+1} \sum_{i=t-l}^t (x_i - \bar{x}^{(1)})^2; & S_x^{(2)} &= \frac{1}{l+1} \sum_{i=t}^{t+l} (x_i - \bar{x}^{(2)})^2; \\ S_u^{(1)} &= \frac{1}{l+1} \sum_{i=t-l}^t (u_i - \bar{u}^{(1)})^2; & S_u^{(2)} &= \frac{1}{l+1} \sum_{i=t}^{t+l} (u_i - \bar{u}^{(2)})^2. \end{aligned}$$

Вищенаведені міркування дозволяють сформулювати обчислювальну схему 1 пошуку вузлів склеювання сплайн-розподілу.

1. Задається l – довжина «плеча», яка визначає довжини масивів (1) та (2).

2. На кожному кроці $t = \overline{l, r-l}$ виконується:

2.1. Відтворення лінійних регресії (3) та (4) на основі масивів (1) та (2) відповідно, яке зводиться до обчислення оцінок параметрів \hat{k}_1 та \hat{k}_2 за формулами (8).

2.2. Обчислення чергового значення показника K за формулою (5), і тим самим, формування масиву $\{K_t; t = \overline{l, r-2l}\}$.

3. Проводиться згладжування даних масиву $\{K_t; t = \overline{l, r-2l}\}$ з метою вилучення шумів. Наприклад, може бути здійснене медіанне згладжування.

4. Знаходяться вузли склеювання сплайн-розподілу шляхом визначення таких елементів у масиві $\{K_t; t = \overline{l, r-2l}\}$, що перевищують зада-

ний поріг. В якості порогу може бути використаний певний відсоток від максимального значення показника K , наприклад, значення $0,9K_{\max}$, де K_{\max} – максимальне значення у масиві $\{K_t; t = \overline{1, r-2l}\}$.

2. Пошук вузлів склеювання сплайн-розподілу на основі порівняння двох регресійних прямих.

Відмінність від попереднього випадку полягає у тому, що визначення наявності відмінності кутів нахилу регресій (3) та (4) зводиться до перевірки статистичної гіпотези

$$H_0 : k_1 = k_2$$

за альтернативи $H_1 : k_1 \neq k_2$.

За масивами (1) та (2) знаходяться оцінки параметрів \hat{k}_1 та \hat{k}_2 регресії (3) та (4) відповідно. Тоді перевірка головної гіпотези здійснюється на основі статистики [6]

$$K = \left| \frac{\hat{k}_1 - \hat{k}_2}{S \sqrt{\frac{1}{(l-1)S_x^{(1)2}} + \frac{1}{(l-1)S_x^{(2)2}}} \right|, \quad (9)$$

де

$$S^2 = S_x^{(1)2} + S_x^{(2)2}.$$

Одержуємо обчислювальну схему 2 пошуку вузлів склеювання сплайн-розподілу.

1. Задається l – довжина «плеча», яка визначає довжини масивів (1) та (2).

2. На кожному кроці $t = \overline{l, r-l}$ виконується:

2.1. Обчислення оцінок параметрів \hat{k}_1 та \hat{k}_2 за формулами (8) за масивами (1) та (2) відповідно.

2.2. Обчислення чергового значення показника K за формулою (9), і тим самим, формування масиву $\{K_t; t = \overline{1, r-2l}\}$.

3. Проводиться згладжування (наприклад, медіанне) даних масиву $\{K_t; t = \overline{1, r-2l}\}$ з метою видалення шумів.

4. Знаходяться вузли склеювання сплайн-розподілу шляхом визначення таких елементів у масиві $\{K_t; t = \overline{1, r-2l}\}$, що перевищують заданий поріг. Поріг може бути обраний як в обчислювальній схемі 1.

3. Пошук вузлів склеювання сплайн-розподілу на основі порівняння двох регресійних прямих, оцінених шляхом відтворення лінійної сплайн-регресії.

У двох попередніх випадках відрізки на ймовірнісному папері будуються шляхом середньоквадратичного наближення до точок масиву (1) та (2), тому побудовані відрізки не обов'язково перетинаються у варіанті x_t . У даному випадку відрізки будуються перетинними у варіанті x_t . Для реалізації цієї ідеї фіксується варіанта x_t варіаційного ряду і розглядається масив

$$\{x_t, u_i; i = \overline{t-l, t+l}\}, \quad (10)$$

за яким відтворюється лінійна сплайн-регресія вигляду

$$u = \begin{cases} k_1 x + b_1, & x \leq x_t, \\ k_2 (x - x_t) + k_1 x_t + b_1, & x \geq x_t. \end{cases} \quad (11)$$

Як і раніше, параметри k_1 та k_2 являють собою тангенси кутів нахилу регресій до вісі абсцис, тому подальший пошук вузлів склеювання сплайн-розподілу аналогічний першому випадку.

Обчислення оцінок параметрів моделі (11) здійснюється за допомогою методу найменших квадратів з умови мінімуму залишкової дисперсії [6]

$$S^2 = \frac{1}{2l-3} \left[\sum_{i=t-l}^t (u_i - k_1 x_i - b_1)^2 + \sum_{i=t+1}^{t+l} (u_i - k_2 (x_i - x_t) - k_1 x_t - b_1)^2 \right],$$

яка еквівалентна розв'язуванню системи алгебричних рівнянь

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} \begin{pmatrix} b_1 \\ k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix},$$

де

$$a_{11} = 2l + 1; \quad a_{12} = \sum_{i=t-l}^t x_i + x_t l; \quad a_{13} = \sum_{i=t+1}^{t+l} (x_i - x_t);$$

$$a_{22} = \sum_{i=t-l}^t x_i^2 + x_t^2 l; \quad a_{23} = x_t \sum_{i=t+1}^{t+l} (x_i - x_t); \quad a_{33} = \sum_{i=t+1}^{t+l} (x_i - x_t)^2;$$

$$g_1 = \sum_{i=t-l}^{t+l} u_i ; \quad g_2 = \sum_{i=t-l}^t x_i u_i + x_t \sum_{i=t+1}^{t+l} u_i ; \quad g_3 = \sum_{i=t+1}^{t+l} (x_i - x_t) u_i .$$

Шукані оцінки параметрів знаходяться за формулами:

$$\hat{k}_1 = \frac{\det \begin{pmatrix} a_{11} & g_1 & a_{13} \\ a_{12} & g_2 & a_{23} \\ a_{13} & g_3 & a_{33} \end{pmatrix}}{\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}}, \quad \hat{k}_2 = \frac{\det \begin{pmatrix} a_{11} & a_{12} & g_1 \\ a_{12} & a_{22} & g_2 \\ a_{13} & a_{23} & g_3 \end{pmatrix}}{\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}}. \quad (12)$$

Приходимо до обчислювальної схеми 3 пошуку вузлів склеювання сплайн-розподілу.

1. Задається довжина «плеча» l , яка визначає довжину масиву (10).

2. На кожному кроці $t = \overline{l, r-l}$ виконується:

2.1. Обчислення оцінок параметрів \hat{k}_1 та \hat{k}_2 сплайн-регресії за масивом вигляду (10) за формулами (12).

2.2. Обчислення чергового значення показника K за формулою (5), і тим самим, формування масиву $\{K_t; t = \overline{1, r-2l}\}$.

3. Проводиться згладжування (наприклад, медіанне) даних масиву $\{K_t; t = \overline{1, r-2l}\}$ з метою вилучення шумів.

4. Знаходяться вузли склеювання сплайн-розподілу шляхом визначення таких елементів у масиві $\{K_t; t = \overline{1, r-2l}\}$, що перевищують заданий поріг. Поріг може бути обраний як в обчислювальній схемі 1.

Запропоновані обчислювальні схеми 1–3 пошуку вузлів склеювання сплайн-розподілу та сформульована на їх основі обчислювальна схема ідентифікації та відтворення розподілів з класу нормального були реалізовані у вигляді програмного забезпечення «SplineNormalDistribution». Адекватність обчислювальних схем підтверджено результатами обчислювальних експериментів на даних імітаційного моделювання. За результатами експериментів найбільш адекватні результати з пошуку вузлів сплайн-розподілу були одержані за обчислювальною схемою 2.

Нижче наводяться результати експерименту, під час якого моделювалась вибірка обсягом $n = 500$ зі сплайн-нормального розподілу з од-

ним вузлом з параметрами $x_0 = 105$, $m = 100$, $\sigma_1 = 20$, $\sigma_2 = 90$. На ймовірнісному папері нормального розподілу (рис. 1) чітко видно дві прямі з різними кутами нахилу, що відповідає сплайн-розподілу з одним вузлом.

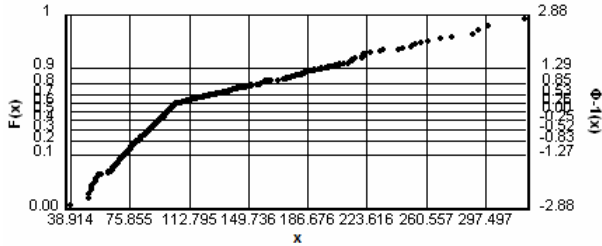


Рис. 1. Ймовірнісний папір нормального розподілу в експерименті 1

Результати роботи обчислювальних схем 1–3 з пошуку вузлів склеювання (рис. 2–4) близькі, характер поведінки показника K дуже схожий, у всіх випадках на графіку цього показника спостерігається один максимум у точці, близькій до заданого під час моделювання вузла склеювання.

За обчислювальною схемою 1 (рис. 2) ідентифіковано один вузол склеювання 103,6. Обчислювальна схема 2 (рис. 3) дозволила ідентифікувати один вузол склеювання 104,5. За обчислювальною схемою 3 (рис. 4) ідентифіковано п'ять вузлів 101,2, 102,4, 102,7, 103,5, 103,7, але всі вони близькі до параметра моделювання.

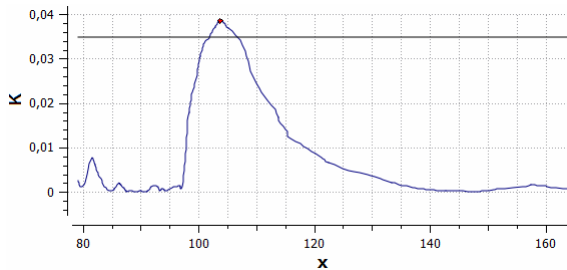


Рис. 2. Графік показника K , одержаний за обчислювальною схемою 1 в експерименті 1

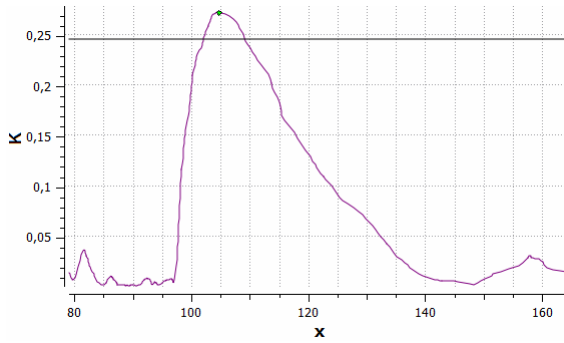


Рис. 3. Графік показника K , одержаний за обчислювальною схемою 2 в експерименті 1

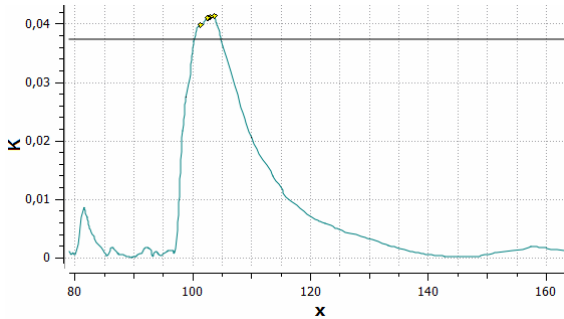


Рис. 4. Графік показника K , одержаний за обчислювальною схемою 3 в експерименті 1

Також подаються результати експерименту, під час якого моделювалась вибірка обсягом $n = 500$ зі сплайн-нормального розподілу з двома вузлами з параметрами $x_0 = 95$, $x_1 = 120$, $m = 100$, $\sigma_1 = 10$, $\sigma_2 = 30$, $\sigma_3 = 10$. На ймовірнісному папері нормального розподілу (рис. 5) чітко виділяються три прями з різними кутами нахилу, що відповідає випадку сплайн-розподілу з двома вузлами.

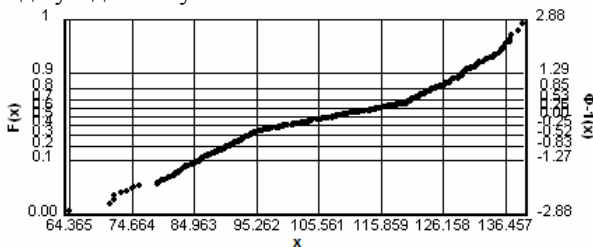


Рис. 5. Ймовірнісний папір нормального розподілу в експерименті 2

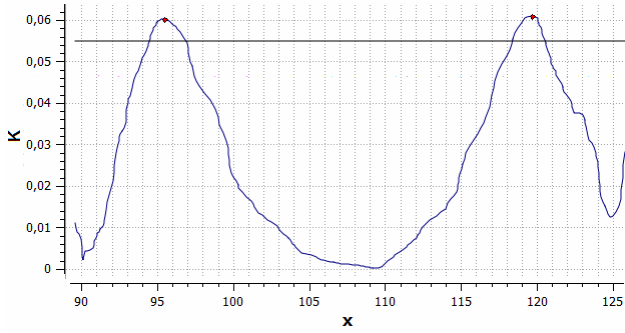


Рис. 6. Графік показника K , одержаний за обчислювальною схемою 1 в експерименті 2

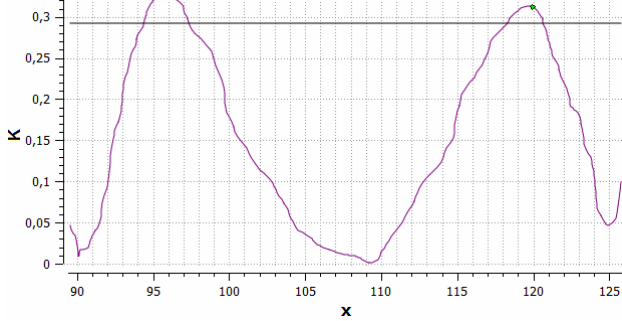


Рис. 7. Графік показника K , одержаний за обчислювальною схемою 2 в експерименті 2

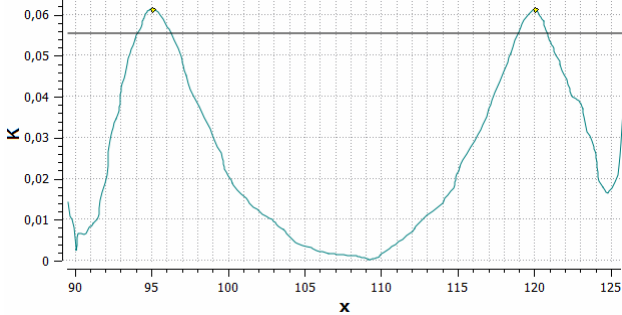


Рис. 8. Графік показника K , одержаний за обчислювальною схемою 3 в експерименті 2

У другому експерименті результати роботи обчислювальних схем 1–3 з пошуку вузлів склеювання (рис. 6–8) також близькі, характер пове-

дінки показника K дуже схожий, у всіх випадках на графіку цього показника спостерігається два максимуми у точках, близьких до заданих під час моделювання вузлів склеювання.

За всіма обчислювальними схемами ідентифіковано два вузли склеювання. Згідно обчислювальної схеми 1 (рис. 6) це 95,4 та 119,6, за обчислювальною схемою 2 (рис. 7) це 95,8 й 119,9, за обчислювальною схемою 3 (рис. 8) ідентифіковано вузли 95,1 та 120,1.

Висновки. Розроблено та реалізовано обчислювальну схему ідентифікації та відтворення розподілів з класу нормального, особливість якої у запропонованому способі ідентифікації сплайн-нормального розподілу шляхом пошуку вузлів сплайна на ймовірнісному папері. Розроблено та реалізовано три обчислювальні схеми пошуку вузлів сплайна на ймовірнісному папері. Їх тестування на даних імітаційного моделювання показало, що всі схеми дають адекватні результати. Результати їх роботи дуже близькі, незначну перевагу має друга схема, оскільки знайдені на її основі оцінки вузлів ближчі до параметрів моделювання. Усі схеми пошуку вузлів склеювання можуть знаходити декілька близько розташованих вузлів. На подолання цього недоліку спрямовані подальші розвідки.

Бібліографічні посилання

1. **Приставка О. П.** Сплайн-розподіли у статистичному аналізі / О. П. Приставка. – Д., 1995. – 152 с.
2. **Приставка А. Ф.** Смеси и сплайн-распределения на неоднородных нормальных пространствах / А. Ф. Приставка, О. В. Райко – Д., 1987. – 233 с. – Деп. в ВИНТИ 11.01.88, №33–В88.
3. **Миленський А. В.** Классификация сигналов в условиях неопределенности / А. В. Миленський. – М., 1975. – 328 с.
4. **Апраушева Н. Н.** Новый подход к обнаружению кластеров / Н. Н. Апраушева. – М., 1993. – 65 с.
5. **Приставка А. Ф.** Идентификация и восстановление распределений на ЭВМ: справ. пособ. / А. Ф. Приставка, О. В. Райко, Н. Л. Малаховская, А. К. Мымриков. – Д., 1991. – 216 с.
6. **Приставка А. Ф.** Вычислительный методы и программная среда корреляционного и регрессионного анализа / А. Ф. Приставка, А.И. Передерий, О.В. Райко, В.М. Остропицкий. – Д., 1996. – 192 с.

Надійшла до редколегії 15.06.11