УДК 519.652:519.254

DOI: 10.18372/2073-4751.74.17879

Зівакін В.Д.,
orcid.org/0000-0002-0420-0558,
Приставка П.О., д.т.н.,
orcid.org/0000-0002-0360-2459,

ДОСЛІДЖЕННЯ ІМІТАЦІЇ ДВОВИМІРНИХ ВИБІРОК З ВИКОРИСТАННЯМ ПОЛІНОМІАЛЬНИХ СПЛАЙНІВ

Національний авіаційний університет

valerii.zivakin@npp.nau.edu.ua,
pylyp.prystavka@npp.nau.edu.ua

Питанням моделювання двовимірних наборів інформації присвячений ряд робіт, наприклад [2-4]. В них згадуються авторегресійні моделі випадкових полів, двопараметричні та інші математичні моделі. Самі роботи присвячені таким методам: на основі математичних моделей масивів з урахуванням особливостей їх формування [2], на основі нелінійної регресії величин за допомогою нормалізуючих перетворень [3] або на основі аналітичної моделі даних із відповідним типом розподілу [4].

Так як в [1] було вказано, що є певний інтерес в подальшому розширені описаних в роботі концепцій на імітацію вже двовимірних даних, однією з цілей роботи була розробка методу імітації вибірок на основі двовимірної сплайн-апроксимації функції щільності. Такий метод мав би перевагу у відсутності необхідності побудови аналітичних моделей та розрахунку параметрів, при необхідності лише якісної апроксимації функції щільності.

Для апроксимації оцінок функції щільності використанні сплайн-оцінки, отримані за допомогою локальних поліноміальних сплайнів, близьких до інтрерполяційних в середньому, на основі В-сплайнів другого порядку. [5]. Відзначимо, що апарат згладжування на основі сплайнів не поступається іншим в оцінці функції щільності і функції розподілу, що обґрунтовано в [5]. Крім того, він має свої переваги за рахунок більшою гнучкості локальної апроксимації з точки зору врахування локальних особливостей функцій (особливо при їхній неоднорідності).

Врахувавши все сказане вище поставлено наступну задачу: нехай маємо двовимірну вибірку $\Omega_{1..N,1..N}$ з деякої генеральної сукупності Ω, необхідно розробити алгоритм та програмний додаток для моделювання нових вибірок $\Omega^*_{1..N,1..N}$, які б належали тій самій генеральній сукупності та перевірити якість його роботи.

Отже, процес імітації починається з генераційної вибірки-основи $\Omega_{1..N,1..N}$, розподіл якої має співпадати з розподілом вибірки-імітації. В якості генераційних вибірок було вирішено моделювати суміші двох двовимірних нормально розподілених величин. Такий підхід забезпечив різну степінь однорідності та рознесеності даних у вибірках-основах.

$$xs_i = w_1 x'_i + w_2 x''_i \qquad (1)$$

В (1) наведено загальну формулу отримання і-го елемента суміші (xs_i) двох розподілів, де w_1, w_2 – вагові коефіцієнти кожного розподілу в суміші, а x'_i, x''_i - і-ті елементи цих розподілів відповідно.

В свою чергу, кожний з розподілів суміші моделюється на основі генерації двох одновимірних нормально розподілених вибірок із заданим коефіцієнтом кореляції:

$$x_1 = M_1 + \delta_1 * z_1,\ x_2 = M_2 + \delta_2 * (z_2 * \sqrt{1 - r^2} + z_1 * r) \qquad (2)$$

де x_1 та x_2 – елементи одновимірних нормально розподілених вибірок, які в сукупності дають елемент двовимірної вибірки, M_1та M_2 – середні значення цих вибірок, δ_1 та δ_2 – їхні середньоквадратичні відхилення, а r – коефіцієнт кореляції міх ними. z_1 та z_2 , в свою чергу – випадкові

елементи, що отримані в результаті моделювання стандартизованої вибірки.

Після отримання вибірки-основи необхідно оцінити її функцію щільності, для чого було використано згаданий вище сплайн-апарат. Слідуючи [5] маємо наступне: нехай за рівномірним розбиттям

$$\Delta_{h_t,h_q}: t_i = ih_t, q_j = jh_q;$$

$$\tilde{\Delta}_{h_t,h_q}: t_i = (i+0{,}5)h_t, q_j = (j+0{,}5)h_q;\ i,j \in Z;\ h_t, h_q > 0 \qquad (3)$$

вісі реалізацій випадкової величини ξ, на підставі вибірки $\Omega_{1..N,1..N} = \{t_i, q_j; i,j = \overline{1,N}\}$ проведено гістограмну оцінку, а отже одержано

$$f_{i,j}, F_{1..N_i,1..N_j}, t \in [t_i; t_{i+1}), q \in [q_j; q_{j+1}), i,j \in Z -$$

масиви оцінок усереднених значень функцій щільності та розподілу. При оцінюванні функції щільності $f(t,q)$ для $\forall t \in [t_{min}; t_{max}]$ та $\forall q \in [q_{min}; q_{max}]$ доцільним є застосування сплайнів на основі B-сплайнів другого та четвертого порядків, серед яких сплайн $S_{2,0}(f,t,q)$ має найпростішу обчислювальну схему

$$S_{2,0}(f,t,q) = \frac{1}{64}\Big(f_{i-1,j-1}(1-2x-2y+4xy+x^2+y^2-2x^2y-2xy^2+x^2y^2) + f_{i-1,j}(6-12x+6x^2-2y^2+4xy^2-2x^2y^2) + f_{i-1,j+1}(1-2x+2y-4xy+x^2+y^2+2x^2y-2xy^2+x^2y^2) + f_{i,j-1}(6-12y-2x^2+6y^2+4x^2y-2x^2y^2) + f_{i,j}(36-12x^2-12y^2+4x^2y^2) + f_{i,j+1}(6+12y-2x^2+6y^2-4x^2y-2x^2y^2) + f_{i+1,j-1}(1+2x-2y-4\times y+x^2+y^2-2x^2y+2xy^2+x^2y^2) + f_{i+1,j}(6+12x+6x^2-2y^2-4\times y^2-2x^2y^2) + f_{1+1,j+1}(1+2x+2y+4xy+x^2+y^2+2x^2y+2\times y^2+x^2y^2)\Big), \qquad (4)$$

де

$$x = \frac{2}{h_t}(t-t_i) - 1;\ i = \left[\frac{t-t_{min}}{h_t}\right] + 1;\ y = \frac{2}{h_q}(q-q_j) - 1;\ j = \left[\frac{q-q_{\min}}{h_q}\right] + 1;$$

$[\,]$ – ціла частина.

Для відновлення функції щільності за допомогою сплайнів є важливою початкова кількість точок, тобто кількість класів на яку розбивається вибірка. В [1] при використанні способу на основі коефіцієнту ексцесу, було отримано найліпший результат:

$$n = \left[\frac{\bar{E}+1.5}{6} * N^{0.4}\right] \qquad (5)$$

де $\bar{E}$ – коефіцієнт ексцесу, n – кількість класів, N – кількість елементів увибірці, [-] – ціла частина.

Після отримання оцінки функції щільності лишається використати її для генерації нової вибірки. В [1] використовувався так званий метод виключень (реджекції), тому було вирішено застосувати його повторно. Нижче приведено опис метода для двовимірного випадку.

Нехай g – тривимірна область, яка з двох боків обмежена деякими інтервалами [a; b], [c; d], а з третього – графіком функції $f_{\eta,\gamma}$, яка є щільністю розподілу випадкової величини , що моделюється.

Помістимо область g всередину іншої області G, що також обмежена інтервалами [a; b], [c; d] і нехай точка (t,q,f) – реалізація випадкового вектора (η,γ,ξ), рівномірно розподіленого в області G. Тоді процедура полягає в тому, що при виконанні нерівності $f_{\eta,\gamma} \geq f$ значення t,q приймаються як реалізація двовимірної випадкової величини.

У підсумку отримуємо наступний алгоритм моделювання на основі заданої вибірки:

1. Провести розбиття вхідної вибірки на класи та визначити їх відносні частоти.

2. На основі отриманих частот відновити функцію щільності розподілу за допомогою поліноміального двовимірного сплайну (2).

3. На основі відновленої функції щільності використати метод виключень для моделювання нової вибірки $\Omega^*_{1..N,1..N}$.

На основі даного алгоритму було проведено експеримент в якому роль вхідних вибіркок-основ виконували вибірки-суміші.

1. Моделюється вибірка суміші двох нормальних розподілів заданої кількості.

2. Для змодельованої вибірки проводиться розбиття на класи (кількість

визначається автоматично за (5)) та рахуються їхні відносні частоти (3).

3. На основі порахованих частот, за допомогою поліноміального сплайну (4), відновлюються значення функції щільності.

4. На основі відновленої функції, за допомогою методу виключень, моделюється нова вибірка із 1000 елементів і для неї проводиться первинний аналіз (кількість класів співпадає із такою у першій вибірці).

5. Пункти з 1 по 4 повторюються по 1000 разів для кожної кількості елементів у першій вибірці (1000, 2000, 5000, 10000). Для імітації різноманітних сумішей математичні очікування розподілів суміші обираються випадково з проміжку від 0 до 100, квадратичні відхилення – від 0 до 20, а кореляції між двома вимірами кожного розподілу, що формують суміш – від 0.2 до 0.7 з випадковим знаком (зведено в табл.). Ваги розподілів у суміші обираються також випадково з проміжку від 0.15 до 0.5 (для першого).

6. Упродовж проведення експерименту вираховуються χ^2 статистики для кожного імітованого розподілу.

Таблиця. Способи вибору параметрів при генерації сумішей

Параметр	Спосіб
Кількість елементів	Стала - 1000, 2000, 5000, 10000
Середні значення одновимірних нормальних розподілів	Випадкові – обираються з інтервалу [0; 100]
Квадратичні відхилення одновимірних нормальних розподілів	Випадкові – обираються з інтервалу [0; 20]
Коефіцієнти кореляції між вимірами у двовимірних нормальних розподілах	Випадкові – обираються з інтервалу [0.2; 0.7]
Ваги для першого розподілу суміші	Випадкові – обираються з інтервалу [0.15; 0.5]

На основі описаної вище процедури було модифіковане програмне забезпечення у середовищі Visual Studio [1]. У ПЗ реалізовані розділи моделювання двовимірних нормально-розподілених вибірок, суміші двох нормальних розподілів та імітування нових двовимірних вибірок на основі попередньо оброблених. Для виконання описаного вище експерименту була виділена окрема кнопка.

Припущення про однорідність генеративних та імітованих вибірок робиться на основі критерію Пірсона, для чого і вираховуються χ^2 статистики у шостому пункті проведення експерименту. Для оцінки на основі двовимірної гістограми необхідно наступне:

1. Обчислення очікуваних значень функції щільності. Для перевірки однорідності вибірок, необхідно обчислити очікувані значень функції щільності для кожної пари варіант, що є центрами класів гістограми. Так як для моделювання вибіроксонов задаються ті ж параметри, що описані в (2), а також ваги сумішей, то очікуване значення функції щільності у відповідних варіантах можна розрахувати наступним чином:

$$f_w(x_t, y_q) = w_1 f_1(x_t, y_q) + w_2 f_2(x_t, y_q),$$

де w_1, w_2 – вагові коефіцієнти кожного розподілу в суміші, f_1, f_2 – значення, що вираховуються з аналітичної формули функції щільності двовимірного нормального розподілу при відомих середніх та квадратичних відхиленнях, а x_t, y_q – відповідні варіанти (центри) кожного класу.

2. Обчислення статистики χ^2. Для цього порівнюються значення функції щільності, що спостерігаються (реальні) та вже обчисленні очікувані. Різниця між реальними та очікуваними значеннями зводиться в квадрат і ділиться на очікувані. Потім підсумовуються такі значення по всіх осередках гістограми. Відомо, що для оцінкою значення функції щільності може слугувати відносна частота класу,

поділена на його площу, тобто $h_t * h_q$ з формули (3):

$$f_r(x_t, y_q) = \frac{p_{t,q}}{h_t * h_q}$$

де $p_{t,q}$ – відносна частота класу (t, q).

Розрахунок статистики:

$$\chi^2 = \sum_{t,q} \frac{\left(f_r(x_t, y_q) - f_w(x_t, y_q)\right)^2}{f_w(x_t, y_q)}$$

3. Оцінка статистичної значимості. Для оцінки статистичної значимості використовується таблиця критичних значень (квантилів) розподілу χ^2. На основі рівня значущості та кількості ступенів свободи, визначається квантиль. Якщо обчислене значення статистики χ^2 перевищує критичне значення, гіпотеза про однорідність вибірок відкидається.

Як зрозуміло з опису експерименту, для кожної початкової кількості елементів у початковій вибірці тисячу разів проводилася імітація двовимірних вибірок кількістю 1000 елементів і для кожного окремого випадку імітації обраховувалась χ^2 статистика. Це дозволило отримати усереднені значення цієї статистики для кожної окремої початкової кількості, що представлено на рисунку.

Так як кількість елементів безпосередньо впливає на розбиття гістограмної сітки, це може викликати збільшення кількості елементів в обрахуванні статистики, тому можна відмітити зростання усередненого значення статистики із зростанням кількості елементів у початковій вибірці.

Та загалом, навіть найбільше значення усередненої статистики дозволяє стверджувати, що імітація була успішною, адже при двох степенях вільності (двовимірні величини) та коефіцієнті значущості 0.05 (стандарт при таких величинах) квантилем розподілу χ^2 є 5.99. Та навіть при збільшені коефіцієнта значущості до 0.9, відповідний квантиль набуває значення 0.211. А отже в будь-якому випадку, приймається гіпотеза про те, що імітовані вибірки є однорідними із вибіркаим, які були задані параметрами.

Був розроблений алгоритм та програмний додаток для імітації двовимір-них вибірок на основі існуючих із вико-ристанням поліноміальних сплайнів. На основі представлених результатів можна стверджувати, що даний алгоритм є ефективним засобом імітації нових вибірок з тієї ж генеральної сукупності, що і вхід-на.

Найголовнішим напрямом подальшої роботи є модифікація алгоритму для його використання в тривимірних та ба-гатовимірних випадках.

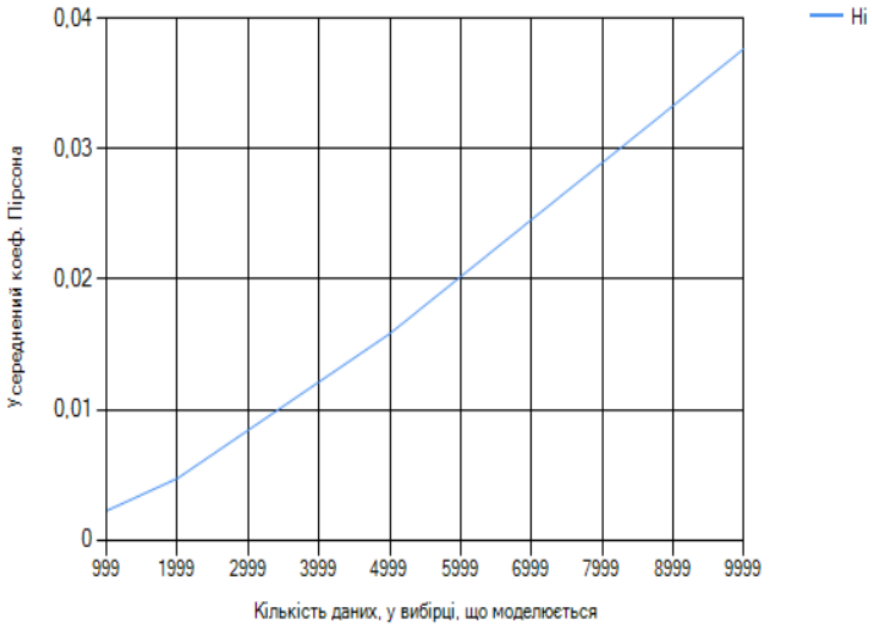

Рис. Залежність усередненого значення χ^2 від початкової кількості даних

Література

1. Зівакін В. Дослідження імітації одновимірних вибірок із використанням поліноміальних сплайнів. *Таврійський науковий вісник. Серія: Технічні науки*. 2021. В. 6. С. 23–30.

2. Подчашинський Ю.О. Розробка методу моделювання масивів двовимірної інформації про механічні величини. *Eastern-European Journal of Enterprise Technologies*. 2010. Т. 1, № 7. С. 14–19.

3. Приходько Н., Приходько С., Кудін О. Комп'ютерне моделювання залежної негаусівської випадкової величини за нелінійною регресійною моделлю на основі нормалізуючого перетворення. *Обробка сигналів і негаусівських процесів*: пр. VII Міжнар. науково-практ. конф., Черкаси, 23-24 трав. 2019 р. / ЧТДУ. Черкаси, Україна, 2019. С. 176–178.

4. Вижва З., Демидов В., Вижва А. Статистичне моделювання випадкових процесів та двовимірних полів в аеромагнітометрії. *Вісник Київського національного університету ім. Тараса Шевченка. Геологія*. 2012. В. 56. С. 52–55.

5. Приставка П.О. Поліноміальні сплайни при обробці даних: монографія. Дніпро : Вид-во Дніпропетр. ун-ту, 2004. 236 с.

Зівакін В.Д., Приставка П.О.

ДОСЛІДЖЕННЯ ІМІТАЦІЇ ДВОВИМІРНИХ ВИБІРОК З ВИКОРИСТАННЯМ ПОЛІНОМІАЛЬНИХ СПЛАЙНІВ

Моделювання використовується для вирішення різноманітних завдань, через певний набір причин: імітація "критичних" режимів, що в умовах реальної експлуатації може бути небезпечним, економія часових і матеріальних ресурсів, можливість дистанційного тренінгу та ін. Зокрема, у випадку проведення досліджень у багатовимірних просторах, актуальним є не моделювання роботи системи, а саме послідовностей даних деякого визначеного вигляду могло б вирішити проблему нестачі таких даних.

В [1] сказано, що при імітаційному моделюванні вибірок перше, з чого необхідно виходити – це модель розподілу, яку необхідно отримати. Модель може бути визначена деяким аналітичним законом розподілу (нормальний, Вейбула, рівномірний, тощо), і в цьому вона залежать від параметрів (параметрична модель). Зазвичай обирають моделі такі, щоб їхні параметри несли деяку змістовну інтерпретацію (a, b – початок та кінець інтервалу в рівномірному розподілі, λ – інтенсивність в експоненціальному, тощо). Іншим класом моделей, що відтворюють функції розподілу є непараметричні (ядерні методи, гістограмні оцінки емпіричної функції розподілу, сплайн – апроксимація [4]). Основною проблемою методів, що ґрунтуються на параметрах, є обмеженість, особливо в двох випадках:

1. При моделюванні багатовимірних даних – в цьому випадку робота завжди призводить до переходу до багатовимірного нормального розподілу.

2. При моделюванні неоднорідних вибірок, які є сумішшю декількох розподілів (не обов'язково з одного класу), усічених або тих, що містять пропуски спостережень.

В цьому контексті використання параметричних моделей об'єктивно є неможливим у чистому вигляді. Отже, наявність інструменту, який добре апроксимує неоднорідні дані є бажаною для вирішення задачі генерації неоднорідних багатовимірних сукупностей.

Ключові слова: *моделювання, імітація, дані, сплайн, щільність розподілу, гістограмна оцінка, апроксимація.*

Zivakin V.D., Prystavka P.O.

RESEARCH OF SIMULATION OF TWO-DIMENSIONAL SAMPLES USING POLYNOMIAL SPLINES

Simulation is used to solve a variety of problems, for a certain set of reasons: imitation of "critical" modes that can be dangerous in real operation, saving time and material resources, the possibility of remote training, etc. In particular, in the case of research in multidimensional spaces, it is important not to model the operation of the system, but rather the data sequences of a certain type, could solve the problem of the lack of such data.

In [1] it is said that when simulating samples, the first thing to start from is the distribution model that needs to be obtained. The model can be defined by some analytical distribution law (normal, Weibull, uniform, etc.), and in this it depends on the parameters (parametric model). Usually, models are chosen such that their parameters carry some meaningful interpretation (a, b - the beginning and end of the interval in a uniform distribution, λ – intensity in an exponential distribution, etc.). Another class of models that reproduce distribution functions are nonparametric ones (kernel methods, histogram estimates of the empirical distribution function, spline approximation [4]). The main problem with parameter-based methods is their limitation, especially in two cases:

When modeling multivariate data - in this case, the work always leads to the transition to a multivariate normal distribution.

When modeling heterogeneous samples that are a mixture of several distributions (not necessarily from the same class), truncated or containing missing observations.

In this context, the use of parametric models is objectively impossible in its pure form. Therefore, having a tool that approximates heterogeneous data well is desirable for solving the problem of generating heterogeneous multivariate populations.

***Keywords**: modeling, simulation, data, spline, distribution density, histogram estimation, approximation.*