

КОМПЮТЪРНО ГЕНЕРИРАНЕ НА СЛУЧАЙНИ ВЕКТОРИ С ГАРАНТИРАНЕ НА СТАТИСТИЧЕСКИТЕ ПАРАМЕТРИ

Димитър Ст. Тянев

УДК 519.23. Коментират се дефектите на компютърно генерирани нормално разпределени статистически извадки и невъзможността генераторите да гарантират техните отнапред зададени статистически параметри. Предлага се метод за присвояване на статистическите извадки на отнапред известните параметри – среден вектор и ковариационна матрица, при пълно удовлетворяване на закона $N^{(k)}(\nu, C)$.

1. Компютърно генерирани извадки - дефекти

Използването на статистическото оценяване на стойностите на дадена величина, известно като метод Монте-Карло [15, 12, 8, 4], става все по-актуално, което се дължи на универсалните възможности на компютърните приложения. Статистическото изследване изисква генерираните експериментални данни да имат отнапред известни статистически параметри. Тези данни обикновено се получават чрез компютърно генериране. Фактичното положение при този подход обаче е такова, че липсват гаранции за точното постигане на необходимите статистически параметри на закона на разпределение на експерименталните данни. Например, числовите поредици, които трябва да се подчиняват на равномерен закон, са фактически с неравномерно разпределение. Нещо повече, обикновено те са псевдослучайни, тъй като са периодично повтарящи се [2, 8, 12]. Когато пък са равномерни [15], те не са случайни. В условията на многомерно пространство се включва изискването за независимост на едномерните генератори [5]. Практическите ситуации се характеризират още с наличие на силна корелираност.

Изследвайки множество методи и програми за генериране на стандартно нормално разпределени извадки от случайни k -мерни вектори [4, 6, 8, 11, 14 и др.], са направени следните по-важни констатации:

1) Извадките се характеризират с изместване и техният среден вектор е различен от нулевия.

Тази констатация има своята обективна природа. Ако в частност се разгледа едномерната случайна величина X , с известен закон на разпределение: $N_x(\mu, \sigma^2)$, то е известно [17] статистическото поведение на оценките на неговите параметри. Известно е например, че извадковата средна стойност \bar{X} , е случайна величина, разпределена по закона $N_{\bar{X}}(\mu, d^2)$. С други думи, средната стойност на отделните извадки е разсеяна около числото μ с дисперсия $d^2 = \sigma^2/n$, където n е означен обемът на отделните извадки. Това разсейване съществува реално, като може да бъде намалено чрез нарастване на обема на извадките.

2) Статистическата структура на извадките е хиперелипсоидална вместо очакваната хиперсферична.

Тази констатация също има своята природа. За споменатата вече едномерна случайна величина X е известно още, че нейната осреднена извадкова дисперсия $\overline{\sigma^2}$ има разпределението $\chi^2(n-1)$. Нейната поправена оценка се получава чрез формулата

$$\overline{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

За генерираната многомерна извадка това означава, че по направление на отделните координатни оси нейните елементи ще бъдат с различна степен на разсейване, а това намира смисъл в собствените стойности на нормалните статистически матрици. С други думи не е гарантирано, че пространствената структура на такава извадка ще бъде хиперсферична.

3) *Собственият им векторен базис е ориентиран случайно и в общо положение*, вместо да съвпада с координатната система на набора от едномерни и независими генератори, както се очаква да бъде;

4) *Извадките имат произволни пълни ковариационни матрици вместо очакваната единична матрица.*

Последните две констатации са следствие от природата на случайната величина. Тъй като тук ние се интересуваме от точното постигане на зададените параметри във всяка генерирана извадка, няма да разглеждаме методите за оценка на доверителните интервали на статистическите оценки на параметрите на закона.

Казаното частично е илюстрирано с експериментални резултати, представени в таблица 1, която съдържа нормата на средните вектори и собствените стойности на ковариационните матрици на конкретни извадки. Извадките са генерирани с помощта на библиотечната подпрограма RNNOA на езика FORTRAN-90 [11], като всяка една от тях съдържа n на брой стандартно и нормално разпределени k -мерни случайни вектори $x_i(k)$, $i=1,2,3,\dots,n$. Ето кратък цитат от библиотечното ръководство за програмиста, относно избраната за експеримента подпрограма:

"Generate pseudorandom numbers from a standard normal distribution using an acceptance/rejection method.

Usage

CALL RNNOA (NR, R)

Arguments

NR — Number of random numbers to generate. (Input)

R — Vector of length NR containing the random standard normal deviates. (Output)

Algorithm

Routine RNNOA generates pseudorandom numbers from a standard normal (Gaussian) distribution using an acceptance/rejection technique due to Kinderman and Ramage (1976). In this method, the normal density is represented as a mixture of densities over which a variety of acceptance/rejection methods due to Marsaglia (1964), Marsaglia and Bray (1964), and Marsaglia et al. (1964) are applied. This method is faster than the inverse CDF technique used in RNNOR to generate standard normal deviates."

Обемите n , за които са получени числените стойности, са подбрани в интервала от 200 до 5000, който е приет за достатъчно представителен. Числените стойности се отнасят за евклидовата норма на средния вектор на всяка извадка $|m|$ и за собствените стойности λ_j , $j = 1,2,3,\dots,k$ на ковариационната матрица $K(k,k)$ за всяка извадка. Резултатите в таблица 1 се отнасят за извадки, генерирани в 7-мерно пространство. Средният вектор характеризира отместването на извадката спрямо началото на координатната система, а собствените стойности – степента на разсейване на елементите на извадката по направление на собствените ѝ вектори.

От представените резултати може да се установи, че собствените стойности са с отклонения спрямо единицата, достигащи до 14% (λ_1 за $n=200$). При големите обеми, от порядъка на $n=5000$, отклоненията достигат 2,5% въпреки, че достоверно намаляват. Подобно е положението и с отместването на извадките. Нормата на средния вектор, която трябва да бъде равна на нула, достига например стойност $|m|=0,21848$ (за $n=200$). Като се има предвид, че статистическото разсейването в пространството в този случай е с радиус не по-голям от $3.\sigma=3$, такова отместване на извадката може да се оцени на повече от 7%. За големите обеми – $n=5000$, отклоненията в отместването спадат под 1%. Извадките с големи обеми обаче не винаги са възможни като вариант, освен това те създават известни изчислителни проблеми [4].

Такива отклонения на посочените параметри от желаните стойности намираме за твърде съществени.

Таблица 1

	$n=200$	$n=500$	$n=1500$	$n=3000$	$n=5000$
$ m $	0.21848	0.15392	0.05995	0.03903	0.02322
λ_1	0.85786	0.90645	1.00208	0.98623	1.02536
λ_2	0.95876	1.03759	0.98528	1.03707	1.00983
λ_3	0.94530	0.96545	1.03142	0.99591	1.00133
λ_4	0.91762	1.07776	0.96478	0.97215	1.01239
λ_5	0.97808	0.94318	0.95656	0.95317	0.98087
λ_6	0.95815	1.06364	1.05786	1.03465	0.99783
λ_7	1.07937	0.95622	0.97793	0.99684	0.99186

Посочените стойности на отклоненията се отнасят за конкретния генератор и за конкретните извадки, което означава, че при други извадки, получени чрез други генератори, могат да се получат други по-големи или по-малки стойности. Този факт на случайно поведение и невъзможност отнапред да бъдат гарантирани статистическите параметри на конкретно генерираните извадки е неприемлив в случаите, когато се търси статистическа оценка на дадена величина, при предварително поставени условия за провеждане на нейното изследване. Такива са мотивите за да се търси възможност за постигане на тези условия върху генерираните експериментални данни.

2. Формулиране на задачата

Изказаните до момента констатации недвусмислено водят до формулиране на задачата за "поправка" на генерираната извадка по такъв начин, че тя винаги да получава исканата статистическа структура. Формалната постановка е следната: дадена е матрицата $X(n,k)$, чиито редове са генерираните (n на брой) k -мерни случайни и нормално разсеяни в пространството вектори, т.е. матрицата X представя генерираната извадка. Фактическите параметри на тази извадка могат да бъдат изчислени и нека те са: $m(k)$ – среден вектор и $K(k,k)$ - ковариационна матрица. Така разпределението, което представя матрицата X , удовлетворява нормалния закон $N^{(k)}(m, K)$. В същото време за извадката се изисква да има параметрите на закона $N^{(k)}(v, C)$. Ясно се разбира, че се извяват две формални задачи:

- задача за преместване на статистическия център от точка m в точка v ;
- задача за преобразуване на статистическата структура, представяна от ковариационната матрица K , в необходимата - представяна от матрицата C .

3. Решение на задачите

Решенията на формулираните задачи не са непосредствено свързани, което е улесняващо обстоятелство. Първата задача се решава лесно, като необходимата трансляция на статистическия център се постига с помощта на разликов вектор r .

Втората задача изисква статистическата извадка X да се преобразува в нова - Y , която да има исканите статистически параметри. Решението се търси във вид на линеен оператор с преобразуваща матрица S . При тази постановка всеки вектор y_i ще се получава така:

$$(1) \quad y_i = S^T \cdot x_i + r, \quad i = \overline{1, n}.$$

Статистическите параметри на извадката Y се оценяват както следва [9]:

✓ за средния вектор:

$$(2) \quad E_i \{ y_i \} = v = E_i \{ S^T \cdot x_i + r \} = r + E_i \{ S^T \cdot x_i \} =$$

$$= \mathbf{r} + \mathbf{S}^T \cdot E_i \{ \mathbf{x}_i \} = \mathbf{r} + \mathbf{S}^T \cdot \mathbf{m}$$

От получения израз може да се определи разликовият вектор \mathbf{r} :

$$(3) \quad \mathbf{r} = \mathbf{v} - \mathbf{S}^T \cdot \mathbf{m}$$

✓ за елементите на ковариационната матрица:

$$(4) \quad E_i \{ (\mathbf{y}_i - \mathbf{v}) \cdot (\mathbf{y}_i - \mathbf{v})^T \} = \mathbf{C} = E_i \{ (\mathbf{S}^T \cdot \mathbf{x}_i + \mathbf{r} - \mathbf{v}) (\mathbf{S}^T \cdot \mathbf{x}_i + \mathbf{r} - \mathbf{v})^T \} =$$

$$= E_i \{ [\mathbf{S}^T \cdot \mathbf{x}_i + \mathbf{r} - (\mathbf{r} + \mathbf{S}^T \cdot \mathbf{m})] \cdot [\mathbf{S}^T \cdot \mathbf{x}_i + \mathbf{r} - (\mathbf{r} + \mathbf{S}^T \cdot \mathbf{m})]^T \} =$$

$$= E_i \{ [\mathbf{S}^T \cdot (\mathbf{x}_i - \mathbf{m})] \cdot [\mathbf{S}^T \cdot (\mathbf{x}_i - \mathbf{m})]^T \} =$$

$$= \mathbf{S}^T \cdot E_i \{ (\mathbf{x}_i - \mathbf{m}) \cdot (\mathbf{x}_i - \mathbf{m})^T \} \cdot \mathbf{S} = \mathbf{S}^T \cdot \mathbf{K} \cdot \mathbf{S} \dots$$

От последния израз се вижда как преобразуващата матрица \mathbf{S} на линейния оператор (1) свързва фактическата ковариационна матрица \mathbf{K} с необходимата - \mathbf{C} . Според [1] необходимото и достатъчно условие за съществуване на матрицата \mathbf{S} е матриците \mathbf{K} и \mathbf{C} да са комутативни. Тук това условие е изпълнено, тъй като ковариационните матрици са симетрични. Построяването на линейния оператор (1) се основава на следната теорема:

Теорема: Ако симетричните матрици \mathbf{K} и \mathbf{C} са положително определени и тяхната ортогонална декомпозиция е означена както следва:

$$(5) \quad \mathbf{K} = \mathbf{V} \cdot \mathbf{A} \cdot \mathbf{V}^T ;$$

$$(6) \quad \mathbf{C} = \mathbf{\Phi} \cdot \mathbf{D} \cdot \mathbf{\Phi}^T ,$$

където \mathbf{A} и \mathbf{D} са диагонални матрици, съдържащи съответните им собствени стойности, а \mathbf{V} и $\mathbf{\Phi}$ са ортонормирани матрици от съответстващите им собствените вектори, то матрицата на линейното преобразуване \mathbf{S} има вида:

$$(7) \quad \mathbf{S} = \mathbf{V} \cdot \mathbf{A}^{-1/2} \cdot \mathbf{D}^{1/2} \cdot \mathbf{\Phi}^T .$$

Доказателство: Заместваме (7) в (4):

$$\left(\mathbf{V} \cdot \mathbf{A}^{-1/2} \cdot \mathbf{D}^{1/2} \cdot \mathbf{\Phi}^T \right)^T \cdot \mathbf{K} \cdot \left(\mathbf{V} \cdot \mathbf{A}^{-1/2} \cdot \mathbf{D}^{1/2} \cdot \mathbf{\Phi}^T \right) = \mathbf{\Phi} \cdot \mathbf{D}^{1/2} \cdot \mathbf{A}^{-1/2} \cdot \left(\mathbf{V}^T \cdot \mathbf{K} \cdot \mathbf{V} \right) \cdot \mathbf{A}^{-1/2} \cdot \mathbf{D}^{1/2} \cdot \mathbf{\Phi}^T$$

Според уравнение (5), записаното в скобите от последния израз е равно на \mathbf{A} , от което следва:

$$(8) \quad \mathbf{\Phi} \cdot \mathbf{D}^{1/2} \cdot \mathbf{A}^{-1/2} \cdot \mathbf{A} \cdot \mathbf{A}^{-1/2} \cdot \mathbf{D}^{1/2} \cdot \mathbf{\Phi}^T = \mathbf{\Phi} \cdot \mathbf{D} \cdot \mathbf{\Phi}^T = \mathbf{C} .$$

Последното показва, че матрицата \mathbf{S} преобразува матрицата \mathbf{K} в \mathbf{C} , както изисква уравнение (4).

4. Случай на изродена ковариационна матрица

В случай, че ковариационната матрица \mathbf{K} е положително полуопределена, при ранг h , ($h < k$), то тя има $(k-h)$ на брой собствени стойности, равни на нула. Тогава диагоналната матрица \mathbf{A} , в декомпозицията (5), има следния блочен вид:

$$(9) \quad \mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_h, \lambda_{h+1} = 0, \dots, \lambda_k = 0) =$$

$$= \begin{bmatrix} \mathbf{A}_h & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Такава матрица има сигнатурно число $q=k-h$ и матрица на инерцията $I(\Lambda) = \text{diag}(I_h, \theta_q)$ [13]. При тези фактически условия многомерната статистическа извадка X е с нулево разсейване в направленията на онези собствени вектори, на които съответствуват нулеви собствени стойности. Този факт прави формирането на матрицата (7) невъзможно, тъй като в нея участва реципрочният квадратен корен $\Lambda^{-1/2}$.

В този случай, за да се гарантира преобразованието (1), ще нулираме сигнатурното число на матрицата Λ , като приемем за нейна матрица на инерцията единичната матрица и като формираме следната блочна матрица:

$$(10) \quad \tilde{\Lambda} = \begin{bmatrix} \Lambda_h & \theta \\ \theta & I_q \end{bmatrix}.$$

Такава матрица има пълен ранг и тогава матрицата $(\tilde{\Lambda})^{-1/2}$ съществува. Така при условията на формулираната теорема преобразованието (1) може да се постигне с помощта на следната преобразуваща матрица:

$$(11) \quad \tilde{S} = V \cdot (\tilde{\Lambda})^{-1/2} \cdot D^{1/2} \cdot \Phi^T.$$

Замествайки (11) в (4), получаваме:

$$(12) \quad \begin{aligned} & \left(V \cdot (\tilde{\Lambda})^{-1/2} \cdot D^{1/2} \cdot \Phi^T \right)^T \cdot K \cdot \left(V \cdot (\tilde{\Lambda})^{-1/2} \cdot D^{1/2} \cdot \Phi^T \right) = \\ & = \Phi \cdot D^{1/2} \cdot \left((\tilde{\Lambda})^{-1/2} \cdot \Lambda \cdot (\tilde{\Lambda})^{-1/2} \right) \cdot D^{1/2} \cdot \Phi^T = \\ & = \Phi \cdot D^{1/2} \cdot I(\Lambda) \cdot D^{1/2} \cdot \Phi^T = \Phi \cdot \begin{bmatrix} D_h & \theta \\ \theta & \theta \end{bmatrix} \cdot \Phi^T. \end{aligned}$$

В последния израз, произведението от трите диагонални матрици има вида (9), който показва, че статистическата извадка Y ще получи желаните параметри, но само в h -мерното подпространство, определено от матрицата на инерцията $I(\Lambda)$. При това положение извадката Y няма да може да получи произволно зададената чрез матрицата C структура. Последната ще бъде постигната частично (само в h -мерното подпространство). В останалата част (q -мерното допълващо подпространство) желаната структура няма да може да бъде постигната - там тя ще бъде игнорирана, което е в пълно съответствие с теоремата на Силвестър за инерцията, според която връзката (4) е възможна, ако двете матрици K и C имат една и съща инерция [13].

5. Заключение

Получаваните при експериментите резултати съответствуват на очакваните. Например, при зададена като желана за статистическата извадка ковариационна матрица C от 3-ти ред:

$$C = \begin{bmatrix} 7 & -3 & 0 \\ -3 & 7 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

без изпълнение на алгоритъма за поправка, при споменатите вече условия и средства, генерираната случайна извадка от 1000 3-мерни вектора, има ковариационна матрица C' , чиито елементи, съществено се отличават от зададените:

$$C' = \begin{bmatrix} 0.6637366 E + 01 & -0.2957354 E + 01 & -0.1840812 E - 01 \\ -0.2957354 E + 01 & 0.6627336 E + 01 & -0.1899240 E + 00 \\ -0.1840812 E - 01 & -0.1899240 E + 00 & 0.7171562 E + 00 \end{bmatrix}$$

В същото време, когато така генерираната извадка се обработи с помощта на алгоритъма за поправка, тя има следната ковариационна матрица:

$$C = \begin{bmatrix} 0.7000002 E + 01 & -0.3000002 E + 01 & 0.7675064 E - 07 \\ -0.3000002 E + 01 & 0.7000001 E + 01 & -0.1636595 E - 06 \\ 0.7675064 E - 07 & -0.1636595 E - 06 & 0.1000000 E + 01 \end{bmatrix}$$

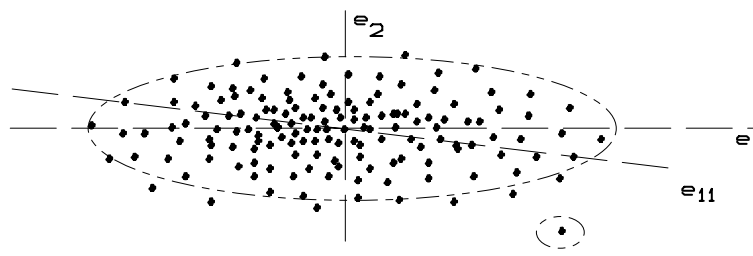
Може да се приеме, че съответствието е постигнато, тъй като то зависи само от точността на компютърните изчисления.

В заключение може да се каже, че представеният метод е в състояние да преобразува една произволна нормална статистическа структура в друга произволно избрана такава, но при условията на формулираната теорема. Това общо изказване важи и за частния случай на положително полуопределена ковариационна матрица C . В този случай матрицата от собствените ѝ стойности D , в декомпозицията (6), ще има вида (9) и за нея ще са в сила изказаните по този повод свойства. Тъй като обаче в (7), както и в (11), участва нейният положителен квадратен корен $D^{1/2}$, този факт не създава затруднения при изпълнение на преобразованието (1). Необходимо е обаче да се разбира, че в този случай, първоначално зададената k -мерна извадка X , след преход в пространството Y , ще загуби разсейването на своите вектори в онези собствени направления, на които съответствуват нулевите собствени стойности на матрицата C . Може да се каже, че в този случай преобразованието (1) променя и размерността на собственото пространство ($k \rightarrow h$) - една способност, която с успех може да се използва при решаването на някои други задачи.

5. Още веднъж за дефектите

Въпреки, че с помощта на предложения метод за преобразуване всяка статистическа структура може да се промени както е необходимо, тя продължава да съхранява някои свои дефекти. Такива дефекти, които не могат да бъдат отстранени с това преобразование, можем да наричаме неотстраними, а борбата с тях да водим с други средства. Типичен пример за такъв дефект в статистическата структура представлява случайното изхвърляне - малко вероятна реализация, неподчиняваща се на правилото на трите сигми (3σ) при генериране на случайните нормално разпределени вектори. В многомерното пространство такава реализация представлява точка, стояща нетипично (недискретно) и явно встрани от основната група точки, каквато е например точката, показана в четвърти квадрант на фигура 1.

В следствие наличието в статистическата извадка на такава случайно изхвърлена точка, главните направления на максималното разсейване ще бъдат определени с отклонение спрямо най-вероятното си положение, което би се получило, ако тази точка не съществуваше - например вместо e_1 ще се получи e_{11} . Наличието в извадките на случайно изхвърлени точки зависи от качествата на генератора с равномерно разпределение, както и от качествата на генератора с нормално разпределение, следователно борбата с такива дефекти следва да води до повишаване на тяхното качество.



Фиг. 1 Влияние на случайно изхвърлена реализация върху главните направления

Литература:

- [1]. Bellman R., *Introduction to matrix analysis*, McGraw-Hill Book company, Inc., 1960.
- [2]. Deley D.W., *Computer Generated Random Numbers*, SYNOPSIS, 30 Apr. 1996, <http://WWW.VIRTUALSCHOOL.EDU/MON/CRIPTO/RANDOMNUMBERMATH>,
- [3]. Jambu M., *Classification automatique pour l'analyse des donnees*, Dunod, Paris, 1978.
- [4]. Forsythe G., Malcolm M., Moler C., *Computer Methods for Mathematical Computations*, Prentice-Hall, Inc., 1977.
- [5]. Fukunaga K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.
- [6]. Haas A., *The multiple prime random number generator*, ACM Transaction on Mathematical Software, vol.13, 4, XII.1987, p. 368-381.
- [7]. Horn R.A., Johnson C.R., *Matrix Analysis*, Cambridge, University Press, 1986.
- [8]. Knuth D.E., *The Art of Computer Programming. Seminumerical Algorithms, vol. 2*, 2nd ed., 1981, Addison Wesley, Reading, MA.
- [9]. Lloyd E., Ledermann W., *Handbook of applicable mathematics, vol. 6: Statistics*, John Wiley&Sons Ltd., 1984.
- [10]. Manasiev L., Konstantinova P., Djerassi E., *Using sequential URAND random generator in distributed random generation*, Proceedings of 11-th International Conference "Systems for automation of engineering and research and DECUS NUG Seminar'97, St. Konstantin resort - Varna, Bulgaria, 20-21 Sept. 1997, p. 86-90.
- [11]. Microsoft FORTRAN Power Station 4.0.
- [12]. Niederreiter H., *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, PA, 1992.
- [13]. Parlett B.N., *The Symmetric Eigenvalue Problem*, Prentice-Hall, 1980.
- [14]. Rabiner L.R., Gold B., *Theory and application of digital signal processing*, Prentice-Hall, New Jersey, 1975.
- [15]. Соболев И.М., *Метод Монте-Карло*, Москва, Издательство "Наука", 1985.
- [16]. Соболев И.М., *Точки, равномерно заполняющие многомерный куб*, Москва, Издательство "Знание", 1985.
- [17]. Afifi A.A., Azen S.P., *Statistical Analysis a Computer Oriented Approach*, Academic Press, New York, 1979.