МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ 2004 год, том 16, номер 2, стр. 118-122

ОБ ИСПОЛЬЗОВАНИИ КВАЗИ-МОНТЕ-КАРЛО В ОЦЕНКАХ BOOTSTRAP

© *И.М.Соболь, Е.Е.Мышецкая*

Институт математического моделирования РАН

Работа выполнена при поддержке РФФИ (грант №02–01–00709)

Оценка bootstrap позволяет в некоторых задачах оценить точность единичного статистического эксперимента. Обычно эта оценка реализуется с помощью метода Монте-Карло.
В настоящей статье построен алгоритм квази-Монте-Карло, скорость сходимости которого в некоторых задачах заметно больше.

ON THE USE OF QUASI-MONTE CARLO IN BOOTSTRAP ESTIMATES

I.M.Sobol', E.E.Myshetskaya

The bootstrap estimate allows to evaluate the accuracy of a single statistical experiment in certain problems, As a rule, this estimate includes a Monte Carlo computation.
In this paper, a quasi-Monte Carlo algorithm is constructed whose convergence rate for certain problems increases considerably.

Введение. В ходе решения многих практических задач, возникают трудности, связанные с недостатком информации об исходных данных.

Обозначим через η случайную величину с неизвестной нам функцией распределения $F_\eta(x)$. Предположим, что $\hat{x}_1,...,\hat{x}_n$ независимые значения η, полученные в эксперименте, и по ним вычисляется значение заданной функции $g(x_1,...,x_n)$. Таким образом, результат эксперимента

$$g(\hat{x}_1,...,\hat{x}_n). \tag{1}$$

Как оценить точность результата (1)?

Оценка bootstrap. Эта оценка была предложена Эфроном (B.Efron) в 1979 году, ей посвящена монография [1]. Основная идея: использовать всевозможные комбинации значений $\hat{x}_1,...,\hat{x}_n$ для приближенного расчета дисперсии $Dg=Dg(\eta_1,\ldots,\eta_n)$.

Во-первых, аппроксимируется математическое ожидание

$$Mg \approx \frac{1}{n^n} \sum_{i_1,...,i_n=1}^{n} g(\hat{x}_{i_1},...,\hat{x}_{i_n}). \tag{2}$$

Точно так же аппроксимируется Mg^2. Тогда $Dg=Mg^2-(Mg)^2$. Заметим сразу, что количество слагаемых в сумме (2) катастрофически растет с увеличением n: уже при $n=8$ значение $n^n=8^8=2^{24}\approx 16\cdot 10^6$. Поэтому обычно сумму (2) вычисляют методом Монте-Карло.

Интерпретация формулы (2). В [2] предложена следующая интерпретация суммы (2). Введем дискретную случайную величину ξ с законом распределения

$$\xi \sim \begin{pmatrix} \hat{x}_1 & \ldots & \hat{x}_n \\ \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}.$$

Тогда среднее (2) равно математическому ожиданию

$$\mathsf{M}g(\xi_1,...,\xi_n)=\frac{1}{n^n}\sum_{i_1,...,i_n=1}^{n} g(\hat{x}_{i_1},...,\hat{x}_{i_n}).$$

Легко видеть, что функция распределения $F_\xi(x)$ совпадает с выборочной функцией распределения $F_n^*(x)$ выборки $\hat{x}_1,...,\hat{x}_n$. Как известно $F_n^*(x)\xrightarrow{P}F_\eta(x)$ при $n\to\infty$ (сходимость имеет место во всех точках непрерывности $F_\eta(x)$). К сожалению, в нашей задаче n фиксировано.

Очевидно, оценка bootstrap смещенная: смещение равно $\mathsf{D}g(\xi_1,...,\xi_n)-\mathsf{D}g(\eta_1,...,\eta_n)$. Модель смещения исследовалась в статье [3]. Здесь мы рассмотрим статистическую погрешность, связанную с приближенным вычислением суммы (2), и возможность уменьшения ее при замене метода Монте-Карло (МК) методом квази-Монте-Карло (кМК).

Алгоритм расчета. Рассмотрим вектор $\xi=(\xi_1,...,\xi_n)$, и пусть $\xi_{(k)}$, k=1,2,..., последовательность независимых реализаций ξ, а $\xi_{(k)}=(\xi_{k1},...,\xi_{kn})$.

Простейший метод МК для оценки. $\mathsf{M}g(\xi)$ – формула средних: при $N\to\infty$

$$\frac{1}{N}\sum_{k=1}^{N} g(\xi_{(n)})\xrightarrow{P}\mathsf{M}g(\xi). \tag{3}$$

Моделировать $\xi_{(k)}$ можно методом обратных функций: выбираем n стандартных случайных чисел $\gamma_{k1},\ldots,\gamma_{kn}$ и полагаем

$$\xi_{kj}=\hat{x}_{1+[n\gamma_{kj}]},\qquad 1\le j\le n;$$

здесь $[z]$ – целая часть z.

Соответствующий алгоритм кМК столь же прост: для расчета вектора $\xi_{(k)}$ выбираем точку $Q_k=(q_{k1},\ldots,q_{kn})$ n-мерной $\mathrm{ЛП}_\tau$-последовательности и полагаем

$$\xi_{kj}=\hat{x}_{1+[nq_{kj}]},\qquad 1\le j\le n.$$

Скорость сходимости (3) для метода МК равна $1/\sqrt{N}$, а для метода кМК скорость сходимости средних в (3) может оказаться $1/N$, или $1/N^\alpha$ с каким-то эффективным показателем $\alpha\le 1$.

Численный эксперимент. Предположим, что случайная величина η равномерно распределена в интервале $0<x<1$, а вычисляемая функция

$$g(x_1,...,x_n)=\prod_{i=1}^{n}\frac{a_i e^{a_i x_i}}{e^{a_i}-1}. \tag{4}$$

Рассматривались 5 вариантов задачи:

Вариант 1:

$$\begin{cases} n=4, & \text{все } a_i=3,\\ \hat{x}_i=0.12,0.38,0.62,0.81. \end{cases}$$

Вариант 2:

$$\begin{cases} n=8, & \text{все } a_i=3,\\ \hat{x}_i=0.12,0.38,0.62,0.81,0.20,0.27,0.53,0.85. \end{cases}$$

Вариант 3:

$$\begin{cases} n=8, & \text{все } a_i=3/i,\\ \hat{x}_i=0.12,0.38,0.62,0.81,0.20,0.27,0.53,0.85. \end{cases}$$

Вариант 4:

$$\begin{cases} n = 8, & \text{все } a_i = 3/(2i), \\ \hat{x}_i = 0.12, 0.38, 0.62, 0.81, 0.20, 0.27, 0.53, 0.85. \end{cases}$$

Вариант 5:

$$\begin{cases} n = 8, & \text{все } a_i = 3/i^2, \\ \hat{x}_i = 0.12, 0.38, 0.62, 0.81, 0.20, 0.27, 0.53, 0.85. \end{cases}$$

Точные значения. Для всех вариантов $I \equiv \mathrm{M}g(\eta) = 1$. Если ввести постоянные

$$c_i = \frac{a_i}{2}\frac{e^{a_i}+1}{e^{a_i}-1},$$

то $I' \equiv \mathrm{M}g^2(\eta) = \prod_{i=1}^{n} c_i$, а дисперсия $\mathrm{D} \equiv \mathrm{D}g(\eta) = I' - I^2$. Для наших вариантов дисперсии D равны соответственно

6.54 – 55.8 – 1.39 – 0.307 – 0.76.

Результаты расчетов. Вместо значений I и I' вычислялись приближения I_N и I'_N по формулам (3), так что $\mathrm{D}_N = I'_N - I^2$. На рис.1 построены значения D_N как функции от $\log_2 N$. Значения, сосчитанные методом kMk, нанесены кружками, а значения, сосчитанные методом МК – крестиками. Очевидно, значения, отмеченные кружками, устанавливаются заметно быстрее. Для вариантов с небольшими дисперсиями (по сравнению с I=1) хорошая точность реализуется уже при N=256.

О грубости оценки bootstrap. Оценка bootstrap предназначена для характеристики точности эксперимента в условиях, когда для классических статистических оценок информации нет. Поэтому с грубостью этой оценки приходится мириться. Посмотрим с этой точки зрения на наш эксперимент.

В наших примерах смещения оценок bootstrap весьма значительны: достаточно сравнить величины

$\mathrm{D}g(\xi)$

1.6 – 4.0 – 0.64 – 0.20 – 0.45

с «идеальными» значениями $\mathrm{D} = \mathrm{D}g(\eta)$, приведенными выше.

Однако, если сопоставить соответствующие средние квадратические погрешности, по которым обычно судят о точности эксперимента, то результат можно считать разумным. В табл.1

$$\sigma = \sqrt{\mathrm{D}g(\eta)}, \qquad \sigma_B = \sqrt{\mathrm{D}g(\xi)}.$$

Таблица 1.

σ и σ_B.

Вариант	1	2	3	4	5
σ	2.6	7.5	1.2	0.55	0.87
σ_B	1.3	2.0	0.80	0.44	0.66

То, что все $\sigma_B < \sigma$, следствие того, что $\hat{x}_i$ были одни и те же.

Очевидно, σ_B – разумная характеристика величины σ.

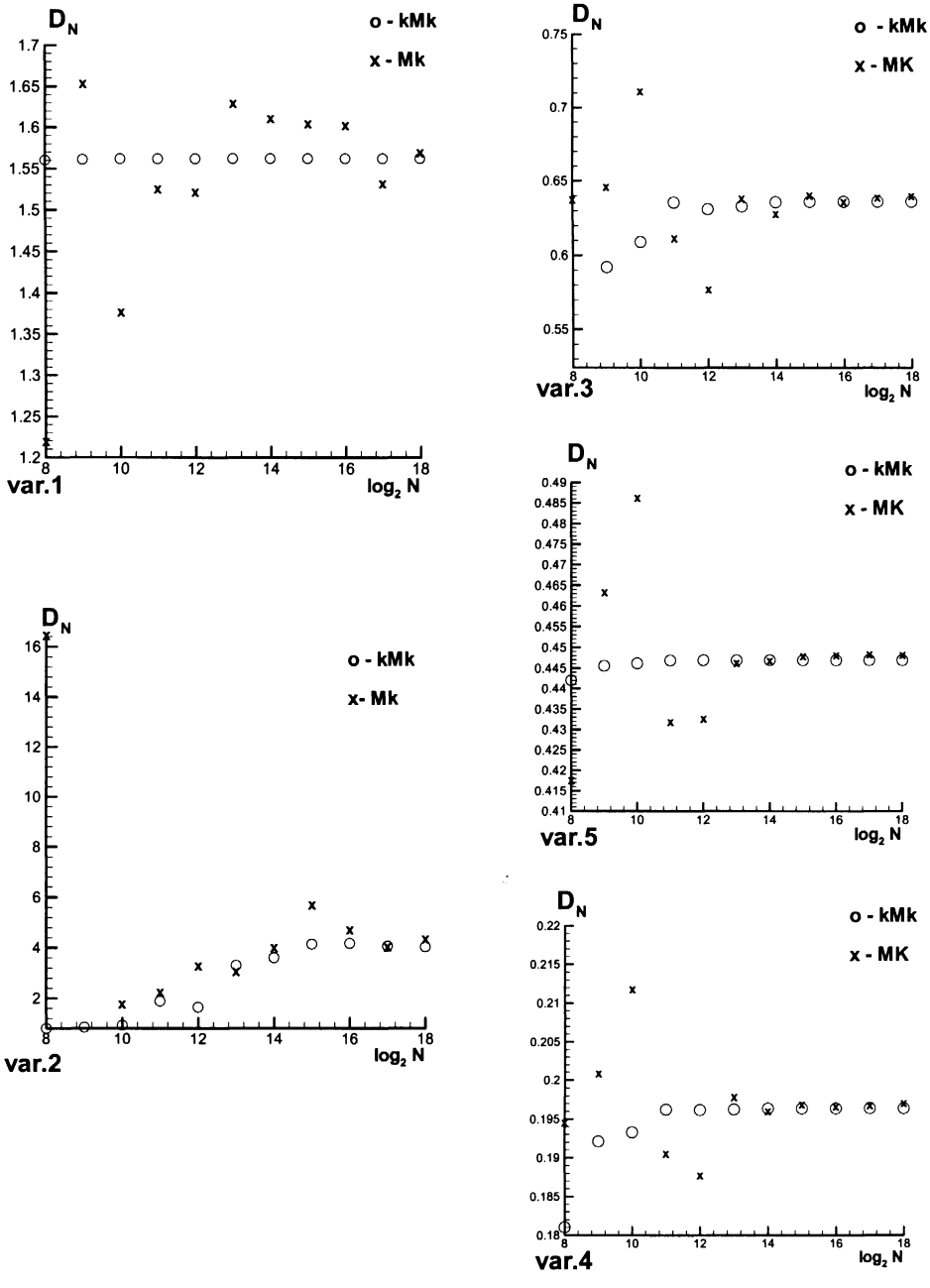

Рис.1. Сосчитанные значения D_N

Замечание: о связи с анализом чувствительности. В [4] введены глобальные показатели чувствительности, позволяющие исследовать структуру нелинейных функций $g(x_i,...,x_n)$. Обозначим через S_i показатель чувствительности первого порядка, соответствующий переменной x_i, и пусть

$$\sum = S_1 + ... + S_n.$$

В [4] доказано, что функция $g(x_i,...,x_n)$ может быть представлена в форме

$$g = g_1(x_1)+...+g_n(x_n) \qquad (5)$$

тогда и только тогда, когда $\sum$=1.

Нетрудно доказать, что средние по точкам $ЛП_\tau$-последовательностей от функций (5) сходятся со скоростью $1/N$ при минимальных требованиях к гладкости $g_i(x_i)$. Поэтому естественно ожидать порядка сходимости, близкого к $1/N$ тогда, когда $\sum$ близко к 1.

Для функции вида (4) все показатели чувствительности легко вычисляются, в частности, показатели первого порядка $s_i = (c_i - 1)/D$.

Для Вар.1: все S_i=0.10, $\sum$=0.40.

Для Вар.2: все S_i=0.012, $\sum$=0.094.

Для Вар.3: S_1=0.47, S_2=0.13, S_3=0.06, S_4=0.03, ... , $\sum$=0.75.

Для Вар.4: S_1=0.60, S_2=0.15, S_3=0.07, S_4=0.04, ... , $\sum$=0.93.

Для Вар.5: S_1=0.86, S_2=0.06, S_3=0.01, S_4=0.004, ... , $\sum$=0.95.

Неудивительно, что для всех вариантов, кроме 2-го, приближения кМк сходятся, как $1/N$. И только для второго варианта сходимость похожа на $1/N^\alpha$ с $\alpha \approx 0.9$.

СПИСОК ЛИТЕРАТУРЫ

1. *B.Efron, R.Tibsharani.* An Introduction to the Bootstrap. Chapman & Hall, – New York, London, 1993.
2. *G.E.B.Archer, A.Saltelli, I.M.Sobol.* Sensitivity measures, ANOVA-like techniques and the use of bootstrap. J.Statist. Comput. Simul., 1997, v.58, p.99-120.
3. *И.М.Соболь, Е.Е.Мышецкая.* Погрешность многомерной квадратурной формулы как модель смещения в методе bootstrap. Кубатурные формулы и их применение. 7-й междунар. семинар, – Красноярск: 2003, с.154-161.
4. *И.М.Соболь.* Об оценке чувствительности для нелинейных математических моделей. Матем. моделирование, 1990, т.2, № 1, с.112-118.

Поступила в редакцию 09.10.2003.