ЭКОНОМИКА И МАТЕМАТИЧЕСКИЕ МЕТОДЫ, 2013, том 49, № 1, с. 119–128

СТАТИСТИЧЕСКИЕ МЕТОДЫ И ТЕОРИЯ ВЕРОЯТНОСТЕЙ

МИНИМАЛЬНЫЙ ОБЪЕМЛЮЩИЙ ПАРАЛЛЕЛЕПИПЕД В ПАРАМЕТРИЧЕСКОМ ОЦЕНИВАНИИ МНОГОМЕРНОГО РАВНОМЕРНОГО РАСПРЕДЕЛЕНИЯ

 Н.И. Киселев

(Москва)

Рассматривается линейная модель генерирования многомерной случайной величины с равномерным распределением в параллелепипеде. Принцип максимального правдоподобия в задачах параметрического оценивания многомерного равномерного распределения формулируется как принцип минимального объема. В общем случае доказаны свойства параллелепипеда минимального объема, включающего все наблюдения выборки. На основе этих свойств обобщается алгоритм комбинаторного типа для нахождения оптимального параллелепипеда. Приводятся результаты численного эксперимента в задаче оценивания центра и ковариационной матрицы двумерной случайной величины, равномерно распределенной в параллелограмме. В эксперименте эффективность оценок минимального объема выше оценок классического метода моментов.

Ключевые слова: объемлющий параллелепипед минимального объема, оценки максимального правдоподобия, равномерное распределение.

ВВЕДЕНИЕ

Нормальное распределение является практически единственной хорошо изученной моделью описания многомерных случайных величин, однако с точки зрения практики описания совокупностей экономических показателей и для сравниельного анализа свойств нормального распределения желательно рассмотрение иных многомерных распределений. В этом плане многомерное равномерное распределение представляет, на наш взгляд, содержательную альтернативу нормальному распределению, которая найдет применение в эконометрических приложениях (Айвазян, 2010) и позволит более глубоко понять свойства и границы применимости нормального распределения.

Принцип максимального правдоподобия, который является основополагающим в задачах параметрического оценивания математической статистики, для многомерного равномерного распределения трансформируется в принцип минимального объема. Если предположить, что нам известен вид формы тела, в котором равномерно распределена случайная величина, то его оценкой максимального правдоподобия по некоторой выборке будет тело данной формы минимального объема, содержащее все наблюдения. В частности, известная в одномерном случае оценка максимального правдоподобия отрезка, на котором случайная величина имеет равномерное распределение, как разности максимальной и минимальной порядковых статистик, очевидно, допускает экстремальную формулировку минимального объема – это оценка, которая следует из условия минимума длины отрезка, содержащего все наблюдения выборки.

В линейной модели образования многомерной случайной величины, где каждая компонента является линейной формой от независимых одномерных равномерно распределенных случайных величин, геометрическим образом формы ее распределения служит параллелепипед. Собственно оценивание этого распределения по выборке состоит в нахождении параллелепипеда минимального объема, содержащего все ее наблюдения. Построение такого параллелепипеда позволяет делать выводы о структуре многомерной случайной величины (Киселев, 2010), вычислять статистики, представляющие интерес для той или иной прикладной задачи.

Задача построения параллелепипеда минимального объема, содержащего все наблюдения, известна, и комбинаторные алгоритмы ее решения для двух- и трехмерного случаев приведены в работах (Вайнштейн, 1990; Vivien, Wicker, 2004). Оба алгоритма основаны на свойстве принадлежности каждой паре параллельных плоскостей оптимального параллелепипеда p + 1-й вершины (p – размерность наблюдений) выпуклой оболочки выборки наблюдений (для общего случая это свойство доказано ниже в разд. 2).

Нахождение параллелепипеда минимального объема формулируется как задача оптимизации с линейными ограничениями и нелинейной целевой функцией. Выписываются необходимые условия оптимальности (условия Куна–Таккера), с помощью которых устанавливаются свойства оптимального решения, позволяющие распространить на общий случай известные для двух- и трехмерной размерностей алгоритмы комбинаторного типа построения искомого параллелепипеда.

Для представления об эффективности оценок минимального объема на конечных выборках проведен численный эксперимент в задаче оценки центра и ковариационной матрицы двумерной равномерно распределенной случайной величины. Приведенные ниже результаты эксперимента показывают их высокую эффективность относительно оценок классического метода моментов.

1. МОДЕЛЬ ОБРАЗОВАНИЯ МНОГОМЕРНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Рассмотрим p-мерную случайную величину x, которая генерируется по правилу

$$x = C\xi, \tag{1}$$

где C – невырожденная матрица размером $p \times p$; ξ – p-мерный вектор, состоящий из независимых одинаково распределенных случайных величин по равномерному закону на отрезке [0, 1].

Геометрическое представление распределения вектора ξ – это p-мерный куб с равномерной плотностью. Так как преобразование (1) аффинное, то геометрическим образом распределения вектора случайной величины x является некоторый параллелепипед с равномерной плотностью. Столбцы матрицы C являются векторами ребер параллелепипеда. Объем $V(C)$ параллелепипеда равен $V(C) = |\det(C)|$. Значение $V(C)$ – это масштаб, по которому пересчитывается объем образа относительно прообраза; $p(x)$ – плотность распределения случайной величины в параллелепипеде, $p(x) = 1/V(C)$.

Заметим, что частные распределения величины x в силу центральной предельной теоремы (Крамер, 1947) при увеличении размерности p (предполагается, что ненулевые элементы матрицы одного порядка и число их в каждой строке стремится к бесконечности при $p \to \infty$) будут сходиться к нормальному распределению. Таким образом, (1) дает пример многомерной величины с равномерным распределением, но каждая из ее компонент при достаточно больших p будет иметь распределение, близкое к нормальному закону.

Пусть $x^{(1)}, ..., x^{(n)}$ выборка независимых наблюдений случайной величины x, где n – объем выборки. Тогда плотность совместного распределения выборки в силу независимости наблюдений:

$$p(x^{(1)}, ..., x^{(n)}) = V(C)^{-n}. \tag{2}$$

Из (2) очевидно, что оценкой максимального правдоподобия матрицы C по выборке $x^{(1)}, ..., x^{(n)}$ будет *матрица* $\hat{C}$ *параллелепипеда минимального объема, который включает все наблюдения выборки*.

2. ФОРМАЛИЗАЦИЯ ОПРЕДЕЛЕНИЯ МИНИМАЛЬНОГО ПАРАЛЛЕЛЕПИПЕДА В ВИДЕ ЗАДАЧИ ОПТИМИЗАЦИИ И ЕГО СВОЙСТВА

Объемлющий параллелепипед, который включает все наблюдения множества $x^{(1)}, ..., x^{(n)}$, описывается системой из $2p$ неравенств:

$$b_j^- \le c_{j1}x_1^{(i)} + ... + c_{jp}x_p^{(i)} \le b_j^+,$$

$$c_{ij} = 1,\ i = 1, ..., n;\quad j = 1, ..., p,$$

где b_j^+ и b_j^- – свободные члены, относящиеся соответственно к верхней и нижней плоскости пары плоскостей j; условие $c_{ij} = 1$, $j = 1,..., p$, фиксирует масштаб направляющих векторов для каждой пары. Тогда нахождение объемлющего параллелепипеда минимального объема может быть представлено в виде задачи оптимизации

$$V = (b_1^+ - b_1^-)...(b_p^+ - b_p^-)/|\det(C)| \to \min \tag{3}$$

при ограничениях, что все наблюдения содержатся в параллелепипеде:

$$Cx^{(i)} \le b^+; \tag{4}$$

$$-Cx^{(i)} \le -b^-; \tag{5}$$

$$c_{jj} = 1, \quad j = 1, ..., p; \quad i = 1, ..., n. \tag{6}$$

Число неизвестных параметров, определяющих пару плоскостей, равно $p + 1$. Тогда общее число неизвестных параметров составляет $p(p + 1)$ при числе ограничений $2pn$.

Выпишем функцию Лагранжа для задачи (3)–(6):

$$L(b, C, \mu, \lambda) = V + \sum_{i=1}^{n}\sum_{j=1}^{p}\left(\lambda_{ij}^+\left(\sum_{k=1}^{p} c_{jk}x_k^{(i)} - b_j^+\right) + \lambda_{ij}^-\left(-\sum_{k=1}^{p} c_{jk}x_k^{(i)} + b_j^-\right)\right).$$

Здесь λ_{ij}^+ и λ_{ij}^- – множители Лагранжа для ограничений, индуцируемых наблюдением i, и означающие его нахождение между верхней и нижней плоскостями пары j.

Приведем необходимые условия (условия Каруша–Куна–Таккера) оптимального решения для данной функции Лагранжа.

1. Условия стационарности:

1.1) $$\frac{\partial L}{\partial b_j^-} = V/(b_j^+ - b_j^-) - \sum_{i=1}^{n}\lambda_{ij}^- = 0, \quad i = 1, ..., n; \quad j = 1, ..., p;$$

1.2) $$\frac{\partial L}{\partial b_j^+} = V/(b_j^+ - b_j^-) - \sum_{i=1}^{n}\lambda_{ij}^+ = 0, \quad i = 1, ..., n; \quad j = 1, ..., p;$$

1.3) $$\frac{\partial L}{\partial c_{jk}} = -(b_1^+ - b_1^-) - ... - (b_p^+ - b_p^-)C_{jk}/\det^2(C) + \sum_{i=1}^{n} x_k^{(i)}(\lambda_{ij}^+ - \lambda_{ij}^-) = 0, \quad j, k = 1, ..., p,$$

где C_{jk} – алгебраическое дополнение элемента c_{jk} в матрице C.

2. Условия дополняющей нежесткости:

$$\lambda_{ij}^+\left(\sum_{k=1}^{p} c_{jk}x_k^{(i)} - b_j^+\right) = 0,$$

$$\lambda_{ij}^-\left(-\sum_{k=1}^{p} c_{jk}x_k^{(i)} + b_j^-\right) = 0,$$

$$i = 1, ..., n; \quad j = 1, ..., p.$$

3. Условия неотрицательности: $\lambda_{ij}^+ \ge 0, \lambda_{ij}^- \ge 0, i = 1, ..., n; j = 1, ..., p.$

Из приведенных необходимых условий оптимального решения следует ряд свойств. Так, из условий 1.1 и 1.2 вытекает, что для любой пары плоскостей сумма значений множителей Лагранжа, наблюдений, которые принадлежат верхней плоскости, равна сумме множителей нижней плоскости, т.е. $\sum_{i=1}^{n}\lambda_{ij}^+ = \sum_{i=1}^{n}\lambda_{ij}^-$ для всех j.

Условия 1.1 и 1.2 можно переписать в виде $\sum_{i=1}^{n}\lambda_{ij}^+ = \sum_{i=1}^{n}\lambda_{ij}^- = V_p/(b_j^+ - b_j^-)$ для всех j, т.е. имеет место утверждение, что для каждой пары плоскостей указанные суммы множителей Лагранжа обратно пропорциональны разности между свободными членами этой пары. Отметим, что приведенные необходимые условия оптимальности решения и полученные следствия указывают на полную симметрию свойств пар плоскостей.

Сформулированная задача оптимизации имеет нелинейную целевую функцию (3) и линейные ограничения (4)–(6). В силу существенной нелинейности целевой функции нам известны решения задачи лишь для двух- и трехмерного случаев. Оба алгоритма основаны на идеях вычислительной геометрии (Препарата, Шеймос, 1989) и состоят из двух этапов. На первом этапе строится многогранник A как выпуклая оболочка множества точек данной выборки $x^{(1)}, ..., x^{(n)}$, на втором – осуществляется перебор среди всех параллелепипедов определенного класса для нахождения минимального.

В двухмерном случае (Вайнштейн, 1990) определение класса возможных параллелограммов основано на свойстве минимального параллелограмма, что в каждой его паре линий одна содержит сторону выпуклого многогранника. Заметим, что этот факт можно представить несколько иначе. По теореме Вейля–Минковского любой выпуклый многогранник можно определить как множество решений некоторой системы линейных неравенств. Тогда указанное свойство перефразируется следующим образом: в каждой паре неравенств, определяющих параллелограмм, одно неравенство совпадает с тем, что имеет место в системе неравенств выпуклого многогранника. Алгоритм, использующий указанное свойство, реализован автором для проведения эксперимента в разд. 4.

В трехмерном случае (Vivien, Wicker, 2004) имеет место обобщение этого свойства – в каждой паре плоскостей минимального параллелепипеда либо одна плоскость содержит грань выпуклого многогранника, либо обе содержат по его ребру, которые не параллельны.

Указанные выше свойства параллелепипедов для размерности $p = 2$ и $p = 3$ можно обобщить на случай любой размерности. Предварительно отметим, что параллелепипед – это выпуклое множество и содержит все точки выборки, следовательно, он содержит выпуклый многогранник A, построенный как выпуклая оболочка всех точек выборки. Тогда внутренняя точка многогранника A будет являться таковой же для параллелепипеда. Следовательно, плоскостям параллелепипеда могут принадлежать только граничные точки многогранника A в виде его граней любой размерности: вершины, ребра и т.д.

Утверждение. *Пусть в конечном множестве* $x^{(1)}, ..., x^{(n)}$ *из* R^p, *где* $n \geq p + 1$, *никакие* $p + 1$ *точки не принадлежат одной плоскости, и пусть объемлющий параллелепипед минимального объема единственный. Тогда каждая пара плоскостей этого параллелепипеда содержит* $p + 1$ *вершину многогранника* A, *построенного как выпуклая оболочка этого множества.*

Предположим, что некоторая пара s плоскостей содержит n_s вершин выпуклого многогранника и $n_s < p + 1$. Тогда направляющий вектор c_s и свободные члены b_s^+, b_s^- пары s удовлетворяют системе линейных уравнений:

$$\sum_{k=1}^{p} c_{sk} x_k^{(i)} - b_s^+ = 0, \quad i = i_1^+, ..., i_{n_s^+}^+; \tag{7}$$

$$\sum_{k=1}^{p} c_{sk} x_k^{(i)} - b_s^- = 0, \quad i = i_1^-, ..., i_{n_s^-}^-; \tag{8}$$

$$c_{ss} = 1, \tag{9}$$

$n_s = n_s^+ + n_s^-$ и пересечение множеств индексов $\{i_1^+, ..., i_{n_s^+}^+\}$ и $\{i_1^-, ..., i_{n_s^-}^-\}$ является пустым. Условие (9) означает выбор масштаба.

В силу $n_s < p + 1$ система (7)–(8) недоопределена относительно неизвестных параметров (их число равно $p + 1$), определяющих пару s плоскостей. Очевидно, что вследствие оптимальности параллелепипеда число вершин, принадлежащих любой паре его плоскостей, не менее двух (как минимум по одной на каждой плоскости). Поэтому всегда можно выразить из (7)–(8) два параметра b_s^+, b_s^- через компоненты направляющего вектора c_s этой пары плоскостей. Пусть для упрощения выкладок b_s^+ выражается из первого уравнения и b_s^- из второго. При подстановке этих параметров в остальные уравнения получим систему вида

$$\sum_{k \subset p_s} c_{sk}(x_k^{(i)} - x_k^{(1)}) = -x_s^{(i)} + x_s^{(1)}; \tag{10}$$

$$\sum_{k \subset p_s} c_{sk}(x_k^{(i)} - x_k^{(2)}) = -x_s^{(i)} + x_s^{(2)}, \tag{11}$$

где множество индексов $p_s = \{1, ..., s-1, s+1, ..., p\}$; число уравнений (10) равно $n_s^+ - 1$, а для (11) – равно $n_s^- - 1$; индекс i в уравнениях (10) и (11) принимает различные значения. Общее число неизвестных коэффициентов c_s равно $p - 1$, тогда как число $n_s - 2$ уравнений в (10)–(11) меньше $p - 1$.

В силу того, что никакие $p + 1$ точки не принадлежат одной плоскости, множество векторов $x^{(i)} - x^{(1)}$ и $x^{(i)} - x^{(2)}$ из (10) и (11) соответственно, общее число которых $n_s - 2$, являются линейно независимыми. В этом случае можно выразить $n_s - 2$ компонент вектора c_s через остальные, обозначим их множество c_s^{rest}. *Изменение компонент из* c_s^{rest} *геометрически приводит к вращению пары s параллельных плоскостей – такому, что изначальные* n_s *вершины, принадлежащие им, остаются на этих плоскостях.* В силу конечного числа точек исходного множества $x^{(1)}, ..., x^{(n)}$ допустимы малые вращения этой пары s, при которых все точки этого множества будут оставаться на границах или между ее плоскостями.

Пусть вектор Δc_s^{rest} – некоторый вектор вариаций компонент из c_s^{rest}, тогда значение нового направляющего вектора $c_s(\Delta)$ при этих вариациях равно $c_s(\Delta) = c_s + B\Delta c_s^{rest}$, где B – матрица размером pr_s, зависящая от координат точек, принадлежащих этой паре плоскостей, и r_s – размерность вектора c_s^{rest}. Соответственно приращение вектора c_s равно

$$c_s(\Delta) - c_s = B\Delta c_c^{rest}, \tag{12}$$

т.е. приращение является линейной функцией от вектора вариаций c_s^{rest}.

Представим объем параллелепипеда (3) как отношение двух линейных форм от вектора c_s. В числителе стоит

$$(b_1^+ - b_1^-)...(b_s^+ - b_s^-)...(b_p^+ - b_p^-) = bc_s^T(x^{(1)} - x^{(2)}), \tag{13}$$

где b – константа, равная произведению разностей в левой части выражений, стоящих в скобках, исключая $(b_s^+ - b_s^-)$, а в знаменателе – модуль детерминанта $|\det(C)|$ матрицы C, и он разложим по строке s, $\det(C) = \sum_{i=1}^{p} c_{si} C_{si} = c_s^{\mathrm{T}} C_s^d$, где верхний индекс Т в обозначении любого вектора означает его транспонирование и одновременно указывает на скалярное произведение, C_{si} – алгебраическое дополнение элемента c_{si} и соответственно C_s^d – вектор этих дополнений. Без потери общности будем полагать, что $\det(C)$ положительная величина. Тогда объем параллелепипеда (3) представляется следующей дробно-линейной функцией от c_s:

$$V = \frac{bc_s^{\mathrm{T}}(x^{(1)} - x^{(2)})}{c_s^T C_s^d}. \tag{14}$$

Соответственно изменение ΔV объема при вращении пары s плоскостей путем вариации Δc_s^{rest} с учетом (12)–(14) равно

$$\Delta V = \frac{(B\Delta c_s^{rest})^{\mathrm{T}}(c_s^{\mathrm{T}} C_s^d(x^{(1)} - x^{(2)}) - c_s^{\mathrm{T}}(x^{(1)} - x^{(2)})C_s^d)}{c_s^{\mathrm{T}} C_s^d(c_s^{\mathrm{T}} C_s^d + \Delta c_s^{\mathrm{T}} C_s^d)}. \tag{15}$$

По условию утверждения минимальный параллелепипед единственный, следовательно, при любом $\Delta c_s^{rest} \neq 0$ значение $\Delta V > 0$. Однако при достаточно малых вариациях $\|\Delta c_s^{rest}\| \leq \varepsilon$ знаменатель в (15) знакопостоянная величина и допустимы вариации Δc_s^{rest} в любом направлении (допустимы в том смысле, что не нарушают условия: все точки лежат внутри или на границах параллелепипеда). Тогда если для некоторой допустимой вариации Δc_s^{rest} изменение объема ΔV положительная величина, то в силу линейной зависимости ΔV от вариаций (15) при изменении их знака на противоположный величина ΔV будет отрицательной. Так как по условию утверждения исходный параллелепипед минимального объема, то полученное противоречие доказывает утверждение.

Замечание 1. Условие единственности минимального параллелепипеда для случая $p = 2$ имеет простую формулировку. Если для каждой пары параллельных сторон ни один из векторов, соединяющих две точки на противоположных сторонах, не параллелен другой паре сторон, то параллелепипед минимального объема единственный.

Замечание 2. Представление объема параллелепипеда (14) как дробно-линейной функции от направляющего вектора c_s (при известном расположении других пар плоскостей) создает предпосылки построения алгоритма поиска минимального параллелепипеда путем многократного решения задачи дробно-линейного программирования, для которой разработаны методы решения (Гольштейн, 1999).

3. КОМБИНАТОРНЫЙ АЛГОРИТМ ПОСТРОЕНИЯ МИНИМАЛЬНОГО ПАРАЛЛЕЛЕПИПЕДА

Полученное утверждение обобщает результаты в работах (Вайнштейн, 1990; Vivien, Wicker, 2004), где указанное свойство оптимального параллелепипеда доказано для двух- и трехмерного случаев. Алгоритм, который можно построить исходя из полученного результата, идейно также совпадает с теми, что приведены в указанных работах. Как уже отмечалось выше, эти алгоритмы состоят из двух этапов – на первом строится выпуклый многогранник A по множеству точек $x^{(1)}, ..., x^{(n)}$, а на втором устраивается перебор среди параллелепипедов, *которые на каждой паре их параллельных плоскостей содержат p + 1 вершину выпуклого многогранника и остальные точки лежат между ними.* С ростом размерности p объем вычислений на каждом этапе значительно увеличивается. В обзоре (Препарата, Шеймос, 1989) представлен набор алгоритмов для построения выпуклого многогранника в двух- и трехмерном случаях. Здесь предложим алгоритм, который можно использовать в задачах большей размерности, хотя комбинаторная сложность задачи, естественно, имеет место.

Собственно, для нахождения параллелепипеда минимального объема нам требуется знание всех вершин многогранника и перечень всех пар параллельных опорных плоскостей, обладающих указанными выше свойствами. Определение вершин можно свести к многократному решению задачи линейного программирования (ЛП). Как известно, симплекс-метод, который реализован в большинстве компьютерных программ, имеет экспоненциальную сложность, но на практике показывает высокую вычислительную эффективность, и потому его многократное применение возможно.

Точка является вершиной, если она не представима выпуклой комбинацией других точек многогранника. Проверку этого можно сделать с помощью решения следующей задачи ЛП.

Пусть $x_1^*, ..., x_k^*$ множество вершин выпуклого многогранника. Точка $\tilde{x}$ не будет вершиной, если существует k-мерный вектор α:

$$\sum_{i=1}^{k} \alpha_i = 1, \alpha_i \geq 0, i = 1, ..., k,$$

такой, что имеет место $\tilde{x} = \alpha_1 x_1^* + \alpha_2 x_2^* + ... + \alpha_k x_k^*$.

Сформулируем задачу нахождения коэффициентов α как задачу линейного программирования:

$$e^{\mathrm{T}}(b^+ + b^-) \to \min, \tag{16}$$

$$X_k \alpha + b^+ - b^- = \tilde{x}, \tag{17}$$

$$e^{\mathrm{T}} \alpha = 1, \ \alpha \geq 0, \tag{18}$$

$$b^+ \geq 0, b^- \geq 0.$$

Здесь X_k – матрица размерности $p \times k$, столбцы матрицы состоят из точек $x_1^*, ..., x_k^*$; $\tilde{x}$ – точка, статус (вершина или не вершина) которой проверяется; e^{T} – единичный вектор-строка; $b^+ \geq 0$, $b^- \geq 0$ – вектора неизвестных положительных невязок.

Итак, если в оптимальном решении задачи (16)–(18) значение целевой функции не равно нулю, то точка $\tilde{x}$ является вершиной, а в противном случае – нет.

Изложенная процедура проверки статуса точки позволяет реализовать алгоритм определения последовательности вершин выпуклого многогранника. На каждом шаге проверяется статус новой точки: если точка не является вершиной, то выбираем следующую, иначе включаем точку в список вершин и множество точек из этого списка проверяем на наличие в нем не вершинных точек, которые могут появиться при включении в него новой точки. Если таковые имеются, исключаем их из списка и переходим к определению статуса следующей точки. На начальном шаге список вершин задается произвольными $p + 1$ точкой выборки, не лежащими в одной плоскости.

После формирования списка вершин необходимо определить все пары параллельных опорных плоскостей, удовлетворяющих выше приведенному условию, что $p + 1$ вершина многогранника принадлежит этим плоскостям. Для этого последовательно определяем пары опорных плоскостей при условии, что на одной плоскости лежит p вершин и на другой одна, затем определяем все пары из условия, что на одной плоскости лежит $p - 1$ вершина и на другой – две и т.д. Процесс завершается при:

– нечетном p на парах плоскостей с одинаковым числом $(p + 1)/2$ вершин на каждой из них;

– четном p на парах плоскостей с распределением вершин – на одной плоскости $p/2 + 1$ и на другой $p/2$ вершин.

После определения полного списка искомых пар опорных плоскостей (обозначим их число N) параллелепипед минимального объема находится простым перебором из всех возможных, число которых C_N^p.

4. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ ПО СРАВНИТЕЛЬНОЙ ЭФФЕКТИВНОСТИ ОЦЕНОК МИНИМАЛЬНОГО ОБЪЕМА ПО СРЕДНЕКВАДРАТИЧНОМУ КРИТЕРИЮ

Эксперимент состоит в том, что для случая $p = 2$ по модели (1) с известной матрицей C генерировались независимые наблюдения случайной величины по 20 выборок с объемами $n = 50$ и $n = 300$. На каждой выборке находилась оценка $\hat{C}$ матрицы, т.е. определялся параллелограмм минимального объема, векторы ребер которого являются столбцами матрицы $\hat{C}$.

Эффективность оценок минимального объема изучается в задачах оценивания центра двухмерной случайной величины из (1) и ее ковариационной матрицы. В качестве меры эффективности используется среднеквадратичный критерий. Оценкой минимального объема для центра случайной величины является точка пересечения диагоналей минимального параллелограмма, а в качестве альтернатив рассматривались классические оценки: выборочное среднее и середина выборочного размаха по каждой координате. Отметим, что все диагонали параллелепипеда пересекаются в одной точке, которая обладает свойством минимаксности. Любая другая точка имеет, хотя бы для одной пары, расстояние до одной из ее плоскостей большее, чем точка пересечения диагоналей.

На рис. 1 и 2 представлены двадцать реализаций оценок центра случайной величины: пересечение диагоналей и выборочного среднего по выборкам объема $n = 50$ и $n = 300$. Точность оценки середины размаха незначительно отличалась от выборочного среднего, и последняя не приводится, чтобы не “утяжелять” рисунки. Заметим, что невысокая эффективность оценки середины размаха объясняется тем, что каждая компонента двухмерной случайной величины представляет сумму двух равномерно распределенных величин и, следовательно, ее распределение уже заметно отличается от равномерного. Тогда как середина выборочного размаха (Дэйвид, 1979) имеет оптимальные свойства для равномерной случайной величины.

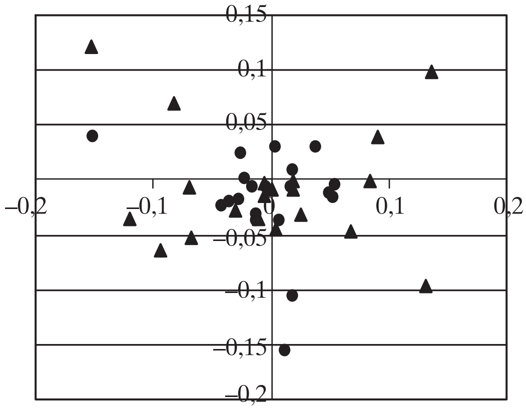

Рис. 1. Объем выборки $n = 50$, где ● – оценки минимального объема, ▲ – выборочные средние

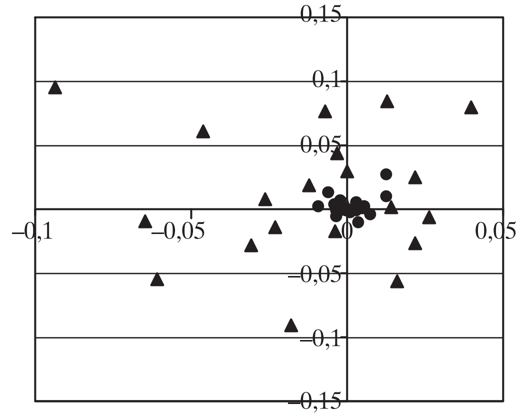

Рис. 2. Объем выборки $n = 300$, где ● – оценки минимального объема, ▲ – выборочные средние

Матрица C размером 2 × 2, которая использовалась для генерирования случайной величины x из (1), равна построчно (1,1; 1,2). При такой матрице коэффициент корреляции компонент случайной величины x равен 0,95. Оценки каждой компоненты также имеют высокую положительную корреляцию и на графике размещаются в узкой прямолинейной полосе. Для получения наглядного графика полученные оценки преобразовались путем умножения на обратную матрицу C^{-1}. На рис. 1–2 помещены уже преобразованные оценки. Среднеквадратичные отклонения, естественно, вычисляются по оценкам без преобразования. Компоненты ξ в (1) представляют две независимые случайные величины, равномерно распределенные на отрезке [–1, 1]. В этом случае истинный центр случайной величины x равен (0, 0).

Как видно на рис. 1–2, при объеме выборки 50 преимущество оценок минимального объема заметно, но не столь очевидно, как при объеме 300, где эти оценки визуально образуют сплошное “пятно” в области истинного значения центра (0, 0). Скорость сходимости оценок минимального объема для центра случайной величины при ее равномерном распределении существенно выше, чем оценок классического метода моментов, что согласуется с известными теоретическими результатами (Дэйвид, 1979).

Средние евклидовы расстояния по 20 оценкам минимального объема, выборочного среднего и середины размахов от истинного центра (0, 0) при объеме выборки $n = 300$ равны соответственно 0,012, 0,091 и 0,099. Относительная эффективность оценки минимального объема по критерию среднего расстояния весьма высока и превосходит среднее выборочное в 0,091/0,012 = 7,5 раза.

Как уже отмечалось, при объеме выборки $n = 50$ оценки минимального объема имеют не столь большое преимущество. Средние расстояния оценок от центра (0, 0) в том же порядке равны 0,086, 0,114 и 0,197, т.е. оценка минимального объема вновь на первом месте (отношение 0,114/0,086 = 1,3), а “худшая” здесь оценка – середина выборочного размаха.

Вторая часть эксперимента состояла в том, что на тех же выборках находились две оценки ковариационной матрицы G: классическим методом через вторые моменты и через оценку минимального параллелограмма. Исходя из модели (1) генерации случайной величины истинная ковариационная матрица равна

$$G = C\overline{\xi\xi^{\mathrm{T}}}C^{\mathrm{T}} = CC^{\mathrm{T}}/3. \qquad (19)$$

Коэффициент 1/3 в (19) появляется в силу того, что в эксперименте ξ распределена равномерно на квадрате с длиной стороны 2. Оценка минимального объема $\hat{C}$ ковариационной матрицы определяется матрицей $\hat{R}$ векторов ребер построенного параллелограмма,

$$\hat{G} = \hat{R}\hat{R}^{\mathrm{T}}/12. \qquad (20)$$

Отметим, что оценка (20) ковариационной матрицы *представляет альтернативу классической оценке*, которая состоит из выборочных вторых моментов. Как мы увидим, в этом эксперименте оценка $\hat{G}$ для двухмерной случайной величины с равномерным распределением в параллелепипеде имеет более высокую эффективность сравнительно с классической.

В качестве критерия точности оценок использовалась матричная евклидова норма разности истинной ковариационной матрицы и ее оценки

$$d = \| G - \hat{G} \| = \sqrt{\sum_{i,j=1}^{2} (g_{ij} - \hat{g}_{ij})^2}.$$

Результаты этого эксперимента весьма близки к тем, что изложены выше для оценки центра случайной величины. При объеме выборки $n = 50$ оценки минимального объема превосходят классические: средние по 20 выборкам расстояния $\bar{d}$ оценок от истинной матрицы равны 0,23 и 0,38 соответственно для оценки минимального объема и классической. При объеме выборки $n = 300$ имеем соответственно 0,04 и 0,14, т.е. здесь оценка минимального объема в 3,5 раза эффективнее классической.

ЗАКЛЮЧЕНИЕ

1. Рассмотрена линейная модель образования многомерной случайной величины, которая приводит к равномерному закону ее распределения в параллелепипеде. Для общего случая доказаны свойства объемлющего параллелепипеда минимального объема. На основе этих свойств обобщается алгоритм комбинаторного типа его нахождения.

2. Для многомерного равномерного распределения принцип максимального правдоподобия в задачах параметрического оценивания формулируется как принцип минимального объема.

3. В приведенном эксперименте на задачах оценивания центра и ковариационной матрицы двухмерной случайной величины, равномерно распределенной в параллелограмме, оценки минимального объема при больших выборках значительно превосходят классические оценки: выборочные средние и ковариационную матрицу, полученную через выборочные моменты. Отмечается высокая скорость сходимости оценок минимального объема при росте объема выборки наблюдений.

СПИСОК ЛИТЕРАТУРЫ

Айвазян С.А. (2010). Методы эконометрики. М.: ИНФРА-М.

Вайнштейн А.Д. (1990). Построение минимального объемлющего параллелограмма // *Дискретная математика.* Т. 2. Вып. 4.

Гольштейн Е.Г. (1999). Двойственный декомпозиционный метод решения общей задачи дробно-линейного программирования // *Экономика и мат. методы*. Т. 35. Вып. 1.

Дэйвид Г. (1979). Порядковые статистики. М.: Наука.

Киселев Н.И. (2010). Альтернативные методы оценки главных компонент // *Прикладная эконометрика.* Т. 8. Вып. 3.

Крамер Г. (1947). Математические методы статистики. М.: Мир.

Препарата Ф., Шеймос М. (1989). Вычислительная геометрия: введение. М.: Мир.

Vivien F., Wicker N. (2004). Minimal Enclosing Parallelepiped in 3D. Computational Geometry // *Theory and applications.* Vol. 29.

Поступила в редакцию
28.03.2011 г.

Minimal Enclosing Parallelepiped in Parametric Estimation of Multidimensional Uniform Distribution

N.I. Kiselev

The linear model of generating a multidimensional random value with uniform distribution in a parallelepiped is discussed. Maximum likelihood principle in problem of parametric estimation is formulated as a principle of the minimal volume. In general case, the distinctive property of a minimal volume parallelepiped, enclosing all set of observations, is proved. On the basis of this property the algorithm of combinatorial class to find this parallelepiped is proposed. The results of numerical experiment on problem of estimation center and covariance matrix are represented for two-dimensional uniform distributed random value. In this experiment, the efficiency of minimum volume estimator is shown to be higher than that of a classical method of the moments.

Keywords: enclosing minimal volume parallelepiped, multidimensional uniform distribution, maximum likelihood estimator.