Один класс многомерных статистических критериев, свободных от распределения *

Клебанов Л.Б.

Посвящается Абраму Ароновичу Зингеру к его 75-ти летию.

Аннотация

Предложен довольно широкий класс свободных от распределения многомерных статистических критериев, основанных на использовании отрицательно определенных ядер. Конструкция близка к введенным и использованным нами ранее вероятностным метрикам (см. [1, 2, 4, 5]) и пригодна также для случая распределений в гильбертовом и некоторых других пространствах. Указаны критерии для проверки однородности, принадлежности аддитивному типу, нормальности с неизвестными параметрами, и родственных гипотез.

AMS Subject Classification: 62G10 Nonparametric inference. Hypothesis testing.

Ключевые слова: свободные от распределения многомерные критерии, вероятностные метрики, отрицательно определенные ядра.

1 Предварительные сведения

Напомним сначала определение отрицательно определенного ядра.

Определение 1.1 *Пусть $\mathfrak{X}$ – некоторое множество, а $\mathcal{L}(x,y)$ – комплекснозначная функция, определенная на $\mathfrak{X}^2$. $\mathcal{L}(x,y)$ называется отрицательно определенным ядром, если неравенство*

$$\sum_{s=1}^{n}\sum_{t=1}^{n}\mathcal{L}(x_s,x_t)h_s\overline{h_t} \leq 0 \tag{1.1}$$

выполнено для любых целых $n \geq 1$, любых точек $x_1,\dots,x_n \in \mathfrak{X}$ и любых комплексных чисел $h_1,\cdots,h_n$, удовлетворяющих условию $\sum_{j=1}^{n} h_j = 0$.

*Работа была поддержана грантами РФФИ 02-01-00262 и 02-01-00483

Определение 1.2 *Ядро $\mathcal{L}(x,y)$ называется строго отрицательно определенным, если равенство в (1.1) достигается тогда и только тогда, когда $h_j = 0$ for all $j = 1,\dots,n$.*

Если $\mathcal{L}(x,y)$ принимает вещественные значения и для любых $x,y \in \mathfrak{X}$

$$\mathcal{L}(x,y) = \mathcal{L}(y,x), \tag{1.2}$$

то в (1.1) достаточно выбырать только вещественные h_j.

Далее мы предполагаем, что $\mathfrak{X}$ - метрическое пространство с метрикой $\mathbf{d}$. Обозначим $\mathfrak{A}$ σ-алгебру всех борелевских подмножеств $\mathfrak{X}$. Говоря далее об отрицательно определенных ядрах мы всегда будем предполагать, что они непрерывны и принимают вещественные значения.

Напомним теперь некоторые результаты о расстояниях, порожденных отрицательно определенными ядрами.

Теорема 1.1 *Пусть $\mathcal{L}(x,y)$ - вещественная непрерывная функция, заданная на $\mathfrak{X}^2$ и удовлетворяющая $\mathcal{L}(x,y) = \mathcal{L}(y,x)$. Неравенство*

$$2\mathbb{E}\mathcal{L}(X,Y) - \mathbb{E}\mathcal{L}(X,X') - \mathbb{E}\mathcal{L}(Y,Y') \geq 0 \tag{1.3}$$

имеет место для всех независимых в совокупности случайных величин X, X', Y, Y' со значениями в $(\mathfrak{X},\mathfrak{A})$ и удовлетворяющими

$$\begin{aligned} X &\stackrel{d}{=} X', & E\mathcal{L}(X,X') &< \infty, \\ Y &\stackrel{d}{=} Y', & E\mathcal{L}(Y,Y') &< \infty \end{aligned} \tag{1.4}$$

тогда и только тогда,когда $\mathcal{L}$ предстваляет собой отрицательно определенное ядро (знак $\stackrel{d}{=}$ используется нами для обозначения равенства распределений).

Доказательство было получено в [1].

Отметим, что определение отрицательно определенного ядра, удовлетворяющего (1.4) может быть переписано в эквивалентном виде как

$$\int_{\mathfrak{X}}\int_{\mathfrak{X}} \mathcal{L}(x,y)h(x)h(y)dQ(x)dQ(y) \leq 0 \tag{1.5}$$

для любой вероятостной меры Q на $(\mathfrak{X},\mathfrak{A})$ и любой интегрируемой функции $h(x)$, для которой $\int_{\mathfrak{X}} h(x)dQ(x) = 0$.

Определение 1.3 *Допустим, что нам дана вероятностная мера Q на $(\mathfrak{X},\mathfrak{A})$ и предположм, что $h(x)$- функция, интегрируемая по мере Q и такая, что $\int_{\mathfrak{X}} h(x)dQ(x) = 0$. Скажем, что $\mathcal{L}$ - сильно орицательно определенное ядро, если $\mathcal{L}$ является отрицательно определенным ядром и равенство*

$$\int_{\mathfrak{X}}\int_{\mathfrak{X}} \mathcal{L}(x,y)h(x)h(y)dQ(x)dQ(y) = 0$$

влечет $h(x) = 0$ Q-почт всюду, не зависимо от выбора меры Q.

Положим

$$\mathcal{N}(X,Y) = 2\mathbb{E}\mathcal{L}(X,Y) - \mathbb{E}\mathcal{L}(X,X') - \mathbb{E}\mathcal{L}(Y,Y').$$

Для сильно отрицательно определенного ядра $\mathcal{L}$ $\mathcal{N}(X,Y)$ представляет собой квадрат расстояния в пространстве функций распределения случайных величин, удовлетворяющих условиям (1.4). Таким образом, если ядро $\mathcal{L}$ является сильно отрицательно определенным, то $\mathcal{N}(X,Y) = 0$ эквивалентно условию $X \stackrel{d}{=} Y$.

Приведем некоторые примеры сильно отрицательно определенных ядер.

Пример 1.1 *Пусть $\mathfrak{X}$ - сепарабельное гильбертово пространство. Допустим, что $f(t)$ - вещественная характеристическая функция безгранично делимой меры на $\mathfrak{X}$. Тогда $\mathcal{L}(t) = -\log f(t)$ - отрицательно определенная функция на $\mathfrak{X}$ т.е., $\mathcal{L}(x-y), x, y \in X$ - отрицательно определенное ядро. Если носитель распределения, соответствующего $f(t)$ совпадает со всем пространством $\mathfrak{X}$, то $\mathcal{L}(x-y), x, y \in X$ - сильно отрицательно определенное ядро.*

Пример 1.2 *Мы знаем, что*

$$\mathcal{L}(t) = \frac{1}{2}(Bt,t) - \int_{\mathfrak{X}} \Big(e^{i(t,x)} - 1 - \frac{i(t,x)}{1+\|x\|^2}\Big)\frac{1+\|x\|^2}{\|x\|^2} d\theta(x),$$

где B - ядерный оператор и θ - конечная мера на $\mathfrak{X}$ такая, что $\theta(\{0\}) = 0$. Ясно, что если $\mathrm{supp}\theta = X$, то $\mathcal{L}$ - сильно отрицательно определенная функция на $\mathfrak{X}$, т.е. $\mathcal{L}(x-y)$ - сильно отрицательно определенное ядро.

В частности, $\|z\|^r$ для $z \in \mathbb{R}^d$ является *отрицательно определенной функцией для $r \in [0,2]$* и *сильно отрицательно определенной для $r \in (0,2)$*.

2 Многомерный критерий однородности

Пусть X, Y - два независимых случайных вектора в $\mathbb{R}^d$. Определим одномерные случайные величины U и V, положив

$$\begin{aligned} U &= \mathcal{L}(X-Y) - \mathcal{L}(X-X'), \\ V &= \mathcal{L}(Y'-Y'') - \mathcal{L}(X''-Y''). \end{aligned} \tag{2.1}$$

Здесь $X \stackrel{d}{=} X' \stackrel{d}{=} X''$, $Y \stackrel{d}{=} Y' \stackrel{d}{=} Y''$, причем векторы X, X', X'', Y, Y', Y'' независимы в совокупности.

Ясно, что условие $\mathcal{N}(X,Y) = 0$ эквивалентно равенству $\mathbb{E}U = \mathbb{E}V$. Однако

$$\mathcal{N}(X,Y) = 0 \iff X \stackrel{d}{=} Y \implies U \stackrel{d}{=} V.$$

Поэтому, при условии (1.4),

$$X \stackrel{d}{=} Y \iff U \stackrel{d}{=} V, \tag{2.2}$$

т.е. гипотеза одинаковой распределенности случайных векторов X и Y эквивалентна гипотезе одинаковой распределенности одномерных случайных величин U и V. Для проверки этой последней гипотезы можно использовать любые известные одномерные свободные от распределения критерии, например, критерий Колмогорова-Смирнова. Разумеется, в случае непрерывных распределений векторов X и Y распределения величин U и V также будут непрерывными, и критерий одинаковой распределенности X и Y, полученный путем применения критерия Колмогорова-Смирнова к U и V, окажется свободным от рапределения.

Пусть $X_1, \dots, X_n, Y_1, \dots, Y_n$ – независимые повторные выборки из генеральных совокупностей X и Y, соответственно. Для применения одномерного критерия к величинам U и V нам нужно построить выборки из этих совокупностей, основываясь на данных $X_1, \dots, X_n$, $Y_1, \dots, Y_n$. Это можно сделать по крайней мере двумя путями.

Один состоит в том чтобы разбить каждую из выборок на три равные части, и по каждую из частей рассматривать как выборку из X, X', X'', и, соответственно, из Y, Y', Y''. Разумеется, этот способ приводит к потере информации, но является теоретически безупречным. Разумеется, он приводит к построению свободного от распределения критерия однородности.

Второй (довольно естественный) подход состоит в том, что мы можем сгенерировать выборки из совокупностей X', X'' (а также Y', Y''), проводя независимые выборы среди наблюдений $X_1, \dots, X_n$ (соответсвенно, $Y_1, \dots, Y_n$). Теоретически этот путь уже не является безукоризненным, так как гипотеза о совпадении генеральных распределений X и Y заменяется гипотезой о совпадении соответствующих эмпирических распределений. Поэтому полученный критерий, очевидно, окажется асимптотически (при $n \to \infty$) свободным от распределения, но, вообще говоря, не будет таковым при фиксированном n. Мы проводили компьютерую симуляцию, которая показала, что эффект отсутствия свободы от распределения в этом случае является практически незаметным при выборках, объема $n \geq 30$.

3 Проверка принадлежности распределений одному аддитивному типу

Пусть $z_1, \dots, z_k$ ($k \geq 3$) - независимые одинаково распределенные случайные векторы в $\mathbb{R}^m$ с функцией распределения $F(x)$. Рассмотрим вектор разностей $Z = (z_2 - z_1, \dots, z_k - z_1)$. Ясно, что распределение этого вектора не измениться, если ко всем z_j ($j = 1, \dots, k$) прибавить один и тот же постоянный вектор $\theta \in \mathbb{R}^m$. Иными словами, распределение Z одно и тоже для аддитивного типа F, т.е. для всех функций распределения вида

$F(x-\theta)$. Задача о восстановлении аддитивного типа распределения по распределению инварианта Z была рассмотрена в [6], где было показано, что восстановление единственно при условии отсутсвия нулей у характеристической функции, соответствующей $F(x)$. Условие отсутсвия нулей может быть заменено требованием аналитичности этой характеристической функции.

Основываясь на указанном выше результате можно построить критерий проверки гипотезы о том что две выборки сделаны из одного и того же аддитивного типа распределения.

Пусть $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ – независимые повторные выборки из генеральных совокупностей X и Y, соответственно. Мы хотим проверить гипотезу о том, что $X \stackrel{d}{=} Y + \theta$ при некотором постоянном неизвестном векторе θ против альтеративы $X \stackrel{d}{\neq} Y + \theta$ при всех θ. Для построения критерия мы можем поступить следующим образом:

1. Путем независимого выбора из значений $X_1, \ldots, X_n$ генерируем $X_1', \ldots, X_n'$ и $X_1'', \ldots, X_n''$.

2. Путем независимого выбора из значений $Y_1, \ldots, Y_n$ генерируем $Y_1', \ldots, Y_n'$ и $Y_1'', \ldots, Y_n''$.

3. Формируем векторные выборки

$$Z_X = ((X_1' - X_1, X_1'' - X_1), \ldots, (X_n' - X_n, X_n'' - X_n))$$

и

$$Z_Y = ((Y_1' - Y_1, Y_1'' - Y_1), \ldots, (Y_n' - Y_n, Y_n'' - Y_n)).$$

4. Используя технику предыдущего раздела проверяем гипотезу о том, что выборки Z_X и Z_Y извлечены из одной совокупности. В силу результатов Коваленко [6] эта гипотеза эквивалентна (в предположении отсутсвия нулей у характеристической функции наблюдений) гипотезе о принадлежности наблюдений одному аддитивному типу.

Ясно, что приведенная методика хороша только асимптотически, так как имеет дефект, связанный с выбором из наблюденных значений. Избежать его снова можно деля исходные выборки на соответсвующее число частей.

4 Провека нормальности наблюдений

Определенный интерес с практической точки зрения представляет собой проверка гипотезы о нормальности векторных наблюдений, причем с неизвестными вектором средних и матрицей ковариаций. Проверка этой гипотезы может быть основана на следующей характеризации нормального закона (см., например, [3]).

Утверждение 4.1 *Пусть Z, Z', Z'', Z''' - четыре независимых одинаково распределенных случайных вектора в $\mathbb{R}^d$. Вектор Z имеет гауссовское распределение тогда и только тогда, когда*

$$Z \stackrel{d}{=} \frac{2}{3}Z' + \frac{2}{3}Z'' - \frac{1}{3}Z'''. \tag{4.1}$$

Пусть $Z_1, \dots, Z_n$ - повторная выборка из генеральной совокупности Z. Критерий проверки гауссовости Z может быть основан на ледующей процедуре:

1. Путем независимого выбора из значений $Z_1, \dots, Z_n$ генерируем $Z'_1, \dots, Z'_n$, $Z''_1, \dots, Z''_n$ и $Z'''_1, \dots, Z'''_n$.

3. Формируем выборки

$$X = (Z_1, \dots, Z_n))$$

и

$$Y = ((\frac{2}{3}Z'_1 + \frac{2}{3}Z''_1 - \frac{1}{3}Z'''_1), \dots, (\frac{2}{3}Z'_n + \frac{2}{3}Z''_n - \frac{1}{3}Z'''_n)).$$

4. Используя результаты раздела 2, проверяем гипотезу однородности для X и Y. В силу Утверждения 4.1 эта гипотеза эквивалентна гипотезе о нормальности Z с произвольными (неизвестными) параметрами.

В заключение заметим, что предлагаемая методика может быть использована для построения критериев проверки независимости, симметрии (см. [2]), сферической симметрии (см. [4]), и некоторых других.

Список литературы

[1] Зингер А.А., Какосян А.В., Клебанов Л.Б. Характеризация распределений с помощью средних значений некоторых статистик в связи с некоторыми вероятностными метриками, Проблемы устойчивости стохастических моделей, ВНИИ системных исследований, Москва, 1989, с. 47-55.

[2] Зингер А.А., Клебанов Л.Б. Характеризация симметрии распределений свойствами моментов, Проблемы устойчивости стохастических моделей, ВНИИ системных исследований, Москва, 1991, с. 70-72.

[3] Kakosyan A.V., Klebanov L.B., Melamed J.A. Characterization of Distributions by the Method of Intensively Monotone Operators, Lecture Notes in Mathematics, 1088, Springer-Verlag, Berlin, 1984.

[4] Klebanov L.B., Kozubovskii T., Rachev S.T., Volkovich V. Characterization of Distributions Symmetric With Respect to a Group of Transformations and Testing of Corresponding Statistical Hypothesis. Statistics & Probability Letters, v. 53, No 3, 2001, 241-247.

[5] Klebanov L.B. A class of Probability Metrics and its Statistical Applications, Statistics in Industry and Technology: Statistical Data Analysis, Yadolah Dodge, Ed. Birkhauser, Basel, Boston, Berlin, 2002, 241-252.

[6] Kovalenko I.N. О восстановлении аддитивного типа распределения по последовательным независимым экспериментам, Труды Всесоюзного совещания по теории вероятностей и математической статистике, Ереван, 1960.