Записки научных
семинаров ПОМИ
Том 341, 2007 г.

В. Н. Судаков, А. В. Судаков

КАПЛИНГ-МЕТРИКИ В ПРОСТРАНСТВЕ ВЫБОРОЧНЫХ МЕР

Пусть X и Y – случайные величины с конечными средними, и $\mathcal{L}(X)$, $\mathcal{L}(Y)$ – их распределения. Расстояние Канторовича $\varkappa(\mathbf{P},\mathbf{Q})$ между двумя вероятностными распределениями $\mathbf{P}$ и $\mathbf{Q}$ на сепарабельном метрическом пространстве $\mathbf{K}$ с метрикой r определяется как

$$\varkappa(\mathbf{P},\mathbf{Q}):=\inf\left\{\iint\limits_{\mathbf{X}\times\mathbf{Y}} r(x,y)m(dxdy) : m\circ\pi_X^{-1}=\mathbf{P}, m\circ\pi_Y^{-1}=\mathbf{Q}\right\},$$

где $\mathbf{X}$ и $\mathbf{Y}$ – две копии пространства $\mathbf{K}$, а x и y – соответственно, их элементы или же элементы пространства $\mathbf{K}$. Каждое вероятностное распределение m на множестве $\mathbf{X}\times\mathbf{Y}$, для которого его маргинальные распределения совпадают с распределениями $\mathbf{P}$ и $\mathbf{Q}$, рассматривается как план перевозки меры $\mathbf{P}$ в меру $\mathbf{Q}$, при котором работа по перевозке единичной массы из точки $x\in\mathbf{K}$ в точку $y\in\mathbf{K}$ равна расстоянию $r(x,y)$.

Как заметил Р. Л. Добрушин (см., например, [1]), расстояние Канторовича $\varkappa(\mathcal{L}(X),\mathcal{L}(Y))$ между вероятностными мерами $\mathcal{L}(X)$ и $\mathcal{L}(Y)$ на числовой прямой $\mathbb{R}$, представимое в этом случае в виде

$$\varkappa(\mathcal{L}(X),\mathcal{L}(Y))=\int|F_X(x)-F_Y(x)|dx$$

или же

$$\varkappa(\mathcal{L}(X),\mathcal{L}(Y))=\int\limits_0^1|F_X^{-1}(t)-F_Y^{-1}(t)|dt$$

(здесь, как обычно, F_Z обозначает функцию распределения случайной величины Z и $\mathcal{L}(Z)$ обозначает закон распределения вероятностей значений случайной величины Z) обладает следующим

Работа выполнена при поддержке гранта НШ 4222.2006.1, РФФИ-ННИО 05-01-00911 и Программой фундаментальных исследований РАН “Современные проблемы теоретической математики”.

(каплинг-)свойством [1]:

$$\varkappa(\mathcal{L}(X),\mathcal{L}(Y)) = \inf\{\mathbf{E}|\xi-\eta| : \mathcal{L}(\xi)=\mathcal{L}(X), \mathcal{L}(\eta)=\mathcal{L}(Y)\},$$

причем ξ_i и η_i – случайные величины, определенные на одном вероятностном пространстве. Короче говоря, для вероятностных распределений $\mathbf{P}$ и $\mathbf{Q}$ на числовой прямой $\mathbb{R}$

$$\varkappa(\mathbf{P},\mathbf{Q}) = \inf\{\|\xi-\eta\|_{L^1} : \mathcal{L}(\xi)=\mathbf{P}, \mathcal{L}(\eta)=\mathbf{Q}\}.$$

Цель настоящей работы – исследовать при определённых ограничениях подобное определение близости двух вероятностных распределений на прямой, возникающее при замене L^1-нормы на другие подходящие нормы. В настоящей работе, возникшей в связи с изучением разброса зависимых гауссовских выборок [2], мы сильно сужаем класс рассматриваемых вероятностных мер, ограничиваясь для наших целей только выборочными распределениями – вероятностными мерами на числовой прямой, являющимися образами равномерных распределённых на конечных множествах фиксированной мощности (т.е. выборками фиксированного объёма, чаще всего без повторяющихся значений).

Поскольку нас будет интересовать всегда выборка в целом, без учёта порядка то конкретная нумерация выборочных значений не играет для нас роли, а выборочным пространством для нас является не координатное пространство $\mathbb{R}^n$, а факторпространство $\mathbb{R}^n/\zeta_T$ этого пространства по соотношению эквивалентности ζ_T, отождествляющем каждую пару элементов координатного пространства, координаты которых получаются друг из друга с помощью некоторой перестановки τ. Подмножество пространства $\mathbb{R}^n$, состоящее из таких элементов, все координаты каждого из которых попарно различны, обозначим $\mathbb{R}_n$. Это – открытое подмножество пространства $\mathbb{R}^n$, состоящее из $n!$ дизъюнктных конусов. Каждый такой конус ограничен $n-1$ гиперплоскостью. Само факторпространство по соотношению эквивалентности ζ_T, рассматриваемому на подмножестве $\mathbb{R}_n$, можно при желании отождествить с группой $T=T_n$ всех перестановок τ. Очевидно, множество T содержит $n!$ перестановок. Будем тривиальную перестановку обозначать $\tau_0 \in T$. Пусть $K_0 \subset \mathbb{R}^n$ обозначает класс эквивалентности, координаты элементов которого монотонно возрастают. Таким образом, соотношение эквивалентности ζ_T позволяет с каждым классом эквивалентности множества $\mathbb{R}_n$

(т.е. с некоторым коническим открытым подмножеством K пространства $\mathbb{R}^n$) канонически связать некоторую перестановку n элементов $\tau_K \in T$, ту, применение которой ранжирует координаты конуса K_0 так же, как ранжированы координаты элементов рассматриваемого конуса K. Тем самым определяется символ $\tau_K : \tau_K K_0 = K$. Группа перестановок T действует на множестве $\mathbb{R}_n/\zeta_T$. Для каждого класса эквивалентности K действие перестановки $(\tau_K)^{-1}$ на этот класс K сводится к расстановке координат каждого элемента $x \in K$ в возрастающем порядке (переход от выборки к вариационному ряду).

Кроме описанной структуры на координатном пространстве $\mathbb{R}^n$, выделившей дизъюнктные открытые конусы K равноранжированных элементов пространства $\mathbb{R}_n$ и дополнительное к их объединению подмножество G, состоящее из граничных точек таких конусов, на этом координатном пространстве рассматривается структура нормированного пространства, т.е. там определена некоторая норма $x \mapsto \|x\|$. (Роль этой нормы для метрики Канторовича, как говорилось выше, играет $L^1(\mathbf{P})$-норма, где $\mathbf{P}$ можно интерпретировать как равномерное вероятностное распределение на единичных ортах пространства, сопряжённого с $\mathbb{R}^n$.) В силу сказанного, естественно предполагать, что эта норма $\|x\|$ не зависит от изменения порядка координат: для любой перестановки $\tau \in T$ и любого элемента $x \in \mathbb{R}^n$ всегда $\|x\| = \|\tau x\|$. Однако, не предполагается часто требуемая в близких ситуациях инвариантность этой нормы относительно замены координат точки их модулями.

Для двух выборочных распределений независимо от наличия или отсутствия совпадающих выборочных значений легко в явной форме выразить расстояние между этими распределениями в метрике Канторовича. Здесь проще всего описать оптимальный план перевозки и по нему подсчитать требуемое расстояние. Критерий Канторовича оптимальности плана перевозки состоит в следующем (см., например, [3]). План перевозки m из $\mathbf{X}$ в $\mathbf{Y}$ является оптимальным (минимизирующим работу по перевозке одного выборочного распределения в другое) тогда и только тогда, когда на метрическом пространстве (K, r) существует такая функция $u : K \to \mathbb{R}$ (потенциальная функция), что

1) функция u удовлетворяет условию Липшица (с константой 1) и

2) m-почти наверное $u(x) - u(y) = r(x, y)$.

Для пары выборочных распределений объёма n покажем, что оптимальным планом перевозки выборочного распределения P_X, отвечающего выборке $\mathbf{x} = \{x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n\}$, в выборочное распределение P_Y, отвечающее выборке $\mathbf{y} = \{y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n\}$, является план перевозки, предусматривающий перенос для каждого k, $k = 1, \ldots n$ массы n^{-1} из точки x_k в точку y_k.(Для краткости мы предполагаем, что обе выборки уже преобразованы в вариационные ряды.) Тем самым мера m для сформулированной задачи является суммой n точечных нагрузок (каждая массы n^{-1}), сосредоточенных в точках координатной плоскости с координатами (x_k, y_k), $k = 1, \ldots, n$. Очевидно, эта мера описывает некоторый план перевозки меры P_X в меру P_y. Потенциальной функцией, подтверждающей экстремальный (минимальный по затратам работы) характер описанного плана служит функция $u : \mathbb{R} \to \mathbb{R}$, определяемая следующим образом. Пусть I_k – последоватедьность всех примыкающих друг к другу отрезков с концами в точках объединённой выборки (их не более $2n$). Функцию $u(t)$ зададим как непрерывную линейную на каждом отрезке I_k функцию, производная которой u' в произвольной точке $t \in \mathbb{R}$ определяется для каждого t следующим правилом:

$u'(t) = 1$, если суммарная масса нагрузок на точках выборки $\mathbf{x}$, лежащих слева от точки t, больше суммарной массы нагрузок на точках выборки $\mathbf{y}$, лежащих в той же области;

$u'(t) = -1$, если суммарная масса нагрузок на точках выборки $\mathbf{x}$, лежащих слева от точки t, меньше суммарной массы нагрузок на точках выборки $\mathbf{y}$, лежащих в той же области;

$u'(t) = 0$, если суммарная масса нагрузок на точках выборки $\mathbf{x}$, лежащих слева от точки t, равна суммарной массе нагрузок на точках выборки $\mathbf{y}$, лежащих в той же области.

Восстанавливаемая по этому описанию с точностью до аддитивной константы функция u, очевидно, липшицева с константой 1. Пусть $x_{(k)}$, $y_{(k)}$ – члены вариационных рядов. Остаётся убедиться, что вероятностное распределение m, сосредоточенное в точках $(x_{(k)}, y_{(k)} \in \mathbb{R}^2$ (ограничимся для простоты изложения случаем выборок без повторяющихся значений) с требуемыми маргинальными распределениями и таково, что относительно него выполнено условие 2), а это почти очевидно. Итак, для метрики Канторовича подсчёт расстояния между двумя выборочными мерами осуществляется без затруднений: от выборок переходим к вариационным рядам и находим сумму модулей разностей:

$$\varkappa(\mathbf{P}_X, \mathbf{P}_Y) = \Sigma_k |x_{(k)} - y_{(k)}|.$$

Если вместе с каждой точкой x пространства $\mathbb{R}^n$ (выборкой) рассматривать и все точки вида τx, то расстояние Канторовича между выборочной мерой, соответствующей выборке x, и выборочной мерой, соответствующей другой выборке $y \in \mathbb{R}^n$, равно минимальному L^1-расстоянию между точкой вида $\tau_i x$ и точкой вида $\tau_j y$. L^1-норма элемента $x \in \mathbb{R}^n$, понимаемого как функция на множестве ортов пространства, сопряженного с пространством $\mathbb{R}^n$ выборок, рассматривается относительно равномерного (с нагрузками по n^{-1}) вероятностного распределения на этом множестве. Тем самым минимальная работа по транспортировке выборочных распределений одно в другое оказывается равной L^1-норме вектора $s_1 - s_2$, где s_1 и s_2 – определяемые по выборкам точки конуса K_0 вариационных рядов. Это – классические результаты. В дальнейшем для простоты мы будем обычно предполагать, что все выборочные значения каждой из двух рассматриваемых выборок попарно различны, т.е. в пространстве $\mathbb{R}^n$ выборочные точки принадлежат каким-то из открытых конусов τK_0. Метрика Канторовича в пространстве выборочных мер произвольного вида вводится вполне аналогично.

Нас, однако, интересуют другие метрики на множестве выборочных распределений, порождаемые различными другими нормами на пространстве выборок $\mathbb{R}^n$.

Рассмотрим теперь другую структуру на нашем пространстве выборок $\mathbb{R}^n$, внешне похожую на только что описанную структуру, порождавшую конусы K и их границы G. Определим для каждого элемента x_k конечного подмножества $N = \{x_1, \ldots, x_m\}$ пока произвольного метрического пространства E с метрикой d подмножество $C_{x_k,N,d} \subset E$, называемое ячейкой близости этого элемента x_k относительно множества N в метрике d и состоящее изо всех таких точек y, для которых элемент x является ближайшим к y из всех элементов множества N.

Определим теперь для нашего случая множество $N \subset \mathbb{R}^n$ следующим образом. Для каждого элемента $x \in K_0$ рассмотрим его образ Tx при воздействии группы перестановок координат $T\colon x \mapsto \{\tau x : \tau \in T\}$. Это множество состоит из $n!$ элементов и играет теперь роль множества N. Каждому элементу $z \in Tx$ сопоставим его ячейку близости $C_{z,Tx,\|\cdot\|} \subset \mathbb{R}^n$ относительно множества Tx, состоящее из всех таких элементов $y \in \mathbb{R}^n$, для которых z – ближайшая к y точка изо всех точек вида τx:

$$C_{z,Tx} = \bigl\{ y \in \mathbb{R}^n : \forall \tau, \tau \in T, \tau \neq \tau_0 \ \|y - z\| \leqslant \|y - \tau z\| \bigr\}.$$

Обозначая $\mathbf{P}_z$ выборочное распределение для выборки $z \in \mathbb{R}^n$, для метрики Канторовича $\varkappa$ в силу теоремы Добрушина [1] и приведённого выше явного выражения для (псевдо)метрики Канторовича на множестве $\mathbb{R}^n$ выборок получаем:

$$\varkappa(\mathbf{P}_x, \mathbf{P}_y) = \|\tau_x^{-1} x - \tau_y^{-1} y\|_{L^1}.$$

(Напомним, что $L^1 = L^1(\mathbf{P}_n)$, где $\mathbf{P}_n$ – вероятностное равномерное распределение на n точках.) Возникает вопрос: насколько существенно эта последняя формула для метрики в пространстве выборочных мер связана с конкретным видом нормы в пространстве выборок $\mathbb{R}^n$, определяющей такую метрику? Для каких норм удобное для измерения расстояния между выборками $x \in \mathbb{R}_n$ и $y \in \mathbb{R}_n$ расстояние Хаусдорфа между двумя подмножествами Tx и Ty измеряется расстоянием между элементами этих подмножеств, попадающими в один и тот же конус одинаково ранжированных элементов, например, в K_0? Легко привести примеры норм в пространстве выборок (например, евклидовых, но не инвариантных относительно операторов перестановок координат τ), для которых это не так. Мы, однако, покажем, что это всегда так для норм, инвариантных относительно перестановок координат. Тем самым кроме более или менее напрашивающегося обобщения метрики Канторовича, возникающего в результате замены L^1-нормы на (координатном) пространстве выборок фиксированного объёма на другую (T-инвариантную) норму, мы получим простую явную формулу для вычисления по двум выборкам равного объёма некоторого нового определённого (не обязательно по метрике Канторовича) расстояния.

Более точно, нас интересует вопрос: совпадают ли разбиение ζ_T на подмножества равноранжированных элементов координатного пространства, определённое выше, и разбиение $\zeta_{C,Tx,\|\cdot\|}$ на ячейки близости для элементов τx, $x \in \mathbb{R}_n$. Если эти разбиения пространства $\mathbb{R}_n$ совпадают, то ближайшим элементом множества Tx к элементу $y_0 \in K_0$ служит элемент $x_0 \in K_0$, и расстояние

Хаусдорфа между двумя подмножествами вида Tx в конечномерном банаховом пространстве $(\mathbb{R}^n, \|\cdot\|)$ равняется расстоянию между двумя точками этих конечных множеств, попавшими в конус K_0 (или в любой другой конус равноранжированных элементов пространства $\mathbb{R}_n$).

Наша ближайшая цель – показать, что такие два разбиения пространства $(\mathbb{R}_n, \|\cdot\|)$ совпадают для произвольной инвариантной относительно перестановки любой пары координат нормы $\|\cdot\|$ и любого элемента $x \in \mathbb{R}_n$. Иначе говоря, требуется показать, что для любого элемента $y \in \mathbb{R}_n$, $y \in K_0$ тот (единственный) элемент y_0 вида τy, который находится в конусе K_0 (и сколь угодно удалён от x_0), удовлетворяет неравенству

$$\|y_0 - x_0\| \leqslant \|y - x_0\|,$$

т.е. содержится в ячейке близости для элемента x_0.

Каждый конус K вида τK_0 представляет собой выпуклый (бесконечный) многогранник в пространстве $\mathbb{R}^n$, ограниченный $n-1$ гиперплоскостью, каждая из которых задаётся для конуса K_0 одним из уравнений

$$x_i = x_{i+1}, \quad i = 1, \ldots, n-1,$$

а сам конус K_0 образован пересечением $n-1$ полупространства вида $x_k < x_{k+1}$.

Предложение. *Пусть на координатном пространстве $\mathbb{R}^n$ рассматривается норма $\|\cdot\|$, инвариантная относительно перестановок координат. Тогда для произвольных точек x_0, y_0 из конуса K_0 и произвольного $\tau \in T$ выполняется неравенство*

$$\|y_0 - x_0\| \leqslant \|\tau y_0 - x_0\|.$$

Доказательство. Рассмотрим сперва тот случай, когла τ – оператор перестановки двух соседних координат, например, y_k и y_{k+1}. Не умаляя общности, предполагаем, что точки x_0 и y_0 находятся в общем положении (в противном случае потребуется простой предельный переход). Конус τK_0 состоит из точек, являющихся зеркальным отражением точек конуса K_0 относительно гиперплоскости $H_{k,k+1}$. (“Зеркальное отражение” здесь понимается как нетождественная изометрия, оставляющая неподвижными точки гиперплоскости $H_{k,k+1}$. Понятие “ортогональность” в наших предположениях лишено смысла.) Пусть

$$z(\lambda) = \lambda x_0 + (1-\lambda)\tau y_0, z(\lambda_0) \in H_{k,k+1}$$

В силу инвариантности нормы относительно перестановок соседних координат

$$\|y_0 - z(\lambda_0)\| = \|\tau y_0 - z(\lambda_0)\|.$$

Тем самым, по неравенству треугольника (не всегда строгому) получаем

$$\|y_0 - x_0\| \leqslant \|\tau y_0 - x_0\|.$$

Для произвольного $\tau \in T$, $\tau \neq \tau_0$, последовательно рассматриваем гиперплоскости H_{kl}, пересекаемые отрезком τy_0, x_0, и повторяем это рассуждение.

Заметим, что пересекаемые гиперплоскости соответствуют последовательности операторов перестановки пары координат, последовательное применение которых в обратной последовательности соответствует применению упомянутого оператора перестановки τ.

Литература

1. А. Н. Ширяев, *Задачи по теории вероятностей.* Издательство МЦНМО, М. (2006).
2. В. Н. Судаков, А. В. Судаков. *Зависимые гауссовские выборки: оценки разброса, свободные от объёма выборки.* — Зап. научн. семин. ПОМИ **320**, (2004), 166–173.
3. Л. В. Канторович, Г. П. Акилов, *Функциональный анализ.* Наука, М. (1977).

Sudakov V. N., Sudakov A. N. Coupling-metrics for sample measures.

A class of metrics on the space of sample distributions is investigated. The distance between sample measures is considered as a Banach norm of the difference of ordered samples.

С.-Петербургское отделение
Математического института
им. В. А. Стеклова РАН

E-mail: sudakov@pdmi.ras.ru

Поступило 23 апреля 2007 г.