УДК 004.62

ОПРЕДЕЛЕНИЕ ГЛУБИНЫ ДАННЫХ НА МНОГОМЕРНЫХ ВЫБОРКАХ

Т.И.Ланге[1], П.Ф.Можаровский[2]
[1]*Университет прикладных наук, Мерзебург,Германия*
[2] *Национальный технический университет Украины «КПИ», Киев, Украина*

Розглянуто поняття глибини даних та глибинно-впорядкованих регіонів, наведені їх основні властивості та класифікація. Детально викладено 5 нотацій функції глибини та результати їх дослідження щодо відповідності вимогам до глибини та регіонів. Викладено одну із останніх форм – глибину зоноїда, для неї вказано спосіб обчислення глибини та алгоритм побудови регіонів. Математичний апарат надає потужний інструментарій для непараметричного аналізу даних, сфера застосування якого охоплює кластерний аналіз, оцінки положення та розсіяння, а також оцінку ризиків.
Ключові слова: глибина даних, глибинно-впорядковані регіони, зоноїд, багатовимірна вибірка.

Concepts of depth and depth trimmed regions are considered, their main properties and classification are presented. 5 notions of data depth and results of investigating their correspondence with requirements, imposed on depth and regions, are discussed. One of the latest forms of depth – zonoid depth and algorithms for depth computation and constructing regions are outlined. These mathematical tools for nonparametric data analysis can be applied in location and dispersion estimation, cluster analysis and risk estimation.
Keywords: data depth, depth trimmed regions, zonoid, multivariate sample.

Рассмотрены понятия глубины данных и глубинно-упорядоченных регионов, приведены их основные свойства и классификация. Подробно изложены 5 нотаций функции глубины и результаты их исследования на предмет соответствия требованиям к глубине и регионам. Изложена одна из последних форм – глубина зоноида, для нее указаны способ вычисления глубины и алгоритм построения регионов. Математический аппарат представляет собой мощный инструмент непараметрического анализа данных, сфера применения которого охватывает кластерный анализ, оценки положения и разброса, а также оценку рисков.
Ключевые слова: глубина данных, глубинно-упорядоченные регионы, зоноид, многомерная выборка.

1 Введение

Функция глубины предоставляет удобства работы с одномерными данными при рассмотрении эмпирических выборок в евклидовом пространстве любой размерности. К преимуществам представления данных на одной оси можно отнести единственность признака, простоту ранжирования, независимость от выбора системы координат и пр.

В самом общем случае глубиной точки $x \in R^d$ называется любая функция $D(x;P)$ для распределения P, заданного в евклидовом d-мерном пространстве R^d, выполняющая упорядочивание по убыванию от центра распределения, основанная на P.

Функцией глубины (в статистическом смысле) является мера степени близости точки к центру выборки в соответствии с распределением вероятностей или эмпирическими данными.

В статье приводятся определения глубины и глубинно-упорядоченных регионов, их основные свойства. Рассмотрены 5 нотаций функции глубины, результаты их исследования на предмет соответствия требованиям, предъявляемым к глубине и регионам, и существующая на сегодняшний день их классификация по способу определения.

В разделе 2 приводятся общие определения глубины и глубинно-упорядоченных регионов, их свойства. В 3-ем разделе рассмотрена исторически первая глубина Тюки, принадлежащая типу D. Раздел 4 содержит классификацию глубин, а представители оставшихся 3 типов рассмотрены в разделах 5 (симплексная глубина), 6 (глубина Ойя) и 7 (глубина Махаланобиса). В 8-м разделе обсуждается наиболее совершенная на сегодняшний день форма глубины – глубина зоноида. Здесь изменен порядок изложения материала (сначала рассмотрены глубинно-упорядоченные регионы и лишь затем глубина) в связи с тем, что глубина определяется с помощью регионов. В 9-м разделе подведены основные итоги и сделаны выводы.

2 Определение и свойства регионов и глубины

На текущий момент существует несколько способов определения глубины, не нацеленных на удовлетворение каких либо практических требований или критериев, а, следовательно, не существует и способа определения, какая из глубин лучше. В [1] проводится рассмотрение свойств функции глубины по 3 основным вопросам:

1) каким желаемым свойствам должна удовлетворять статистическая функция глубины?

2) какие конструктивные подходы позволяют получить привлекательные в смысле удовлетворения свойств функции глубины?

3) существуют ли функции глубины, удовлетворяющие всем желаемым свойствам?

В этой же работе на основании 4-х определяющих свойств функций глубины, впервые сформулированных в [2], вводится общее определение «статистической функции глубины».

Рассмотрим функцию глубины: отображение $D(\cdot;P_X):R^d \mapsto R^1$, или $D(\cdot;P_X):R^d \mapsto [0,1]$, так как более удобно работать именно на этом отрезке. Чтобы глубина была эффективным подходом при упорядочивании по убыванию от центра, она должна удовлетворять следующим требованиям:

1) Аффинная инвариантность. Глубина точки $x \in R^d$ не должна зависеть от выбранной системы координат или шкалы измерения, претерпевать изменение вследствие применения аффинного оператора к выборке и рассматриваемой точке: $D(A \cdot x + b; P_{A \cdot X + b}) = D(x; P_X)$ для любого $b \in R^d$ и любой несингулярной матрицы A размерности $d \times d$.

2) Максимальность в центре. Для выборки с единственной «центральной» точкой (например, точкой симметрии по каким-то

соображениям) функция глубины должна достигать максимального значения в этой точке: $D(\theta; P_X) = \sup_{x \in R^d} D(x; P_X)$ в случае, если θ является центром P_X.

3) Монотонность по отношению к наиболее глубокой точке. Глубина точки должна монотонно убывать при движении вдоль прямой, проходящей через наиболее глубокую точку (точка, в которой функция глубины достигает максимального значения; для симметричных распределений – центр) в направлении от наиболее глубокой точки: если $\theta = \arg\max_x D(x; P_X)$, то $D(x; P_X) \le D(\theta + \alpha \cdot (x - \theta); P_X)$ для любого $\alpha \in [0,1]$.

4) Стремление к нулю в бесконечности. Глубина точки стремится к нулю, если ее мера $\|x\|$ стремится к бесконечности: $\lim_{\|x\| \to \infty} D(x; P_X) = 0$.

Функция $D(\cdot; P_X): R^d \mapsto R^1$, удовлетворяющая всем 4-м требованиям, называется «статистической функцией глубины».

Глубинно-упорядоченным регионом (центральным регионом) называется множество точек, у которых глубина не меньше наперед заданного значения – глубины региона:

$$D_\alpha(P_X) = \left\{ x \in R^d : D(x; P_X) \ge \alpha \right\}. \quad (1)$$

Приведем некоторые из свойств, которым могут удовлетворять глубинно-упорядоченные регионы (часть из них была рассмотрена в [3], изложенные ниже подытожены в [4]).

1) Аффинная эквивариантность: $D_\alpha(P_{A \cdot X + b}) = A \cdot D_\alpha(P_X) + b$ для любых случайного вектора $X \in R^d$, несингулярной матрицы A размерности $d \times d$ и $b \in R^d$.

2) Вложенность: если $\alpha \ge \beta$, то $D_\alpha(P_X) \subseteq D_\beta(P_X)$.

3) Монотонность: если $Y \le X$ (покомпонентно), то $D_\alpha(P_X) + R_+^d \subseteq D_\alpha(P_Y) + R_+^d$.

4) Компактность: $D_\alpha(P_X)$ компактный.

5) Выпуклость (или по меньшей мере связность): $D_\alpha(P_X)$ выпуклый или, по меньшей мере, связный.

6) Субаддитивность: $D_\alpha(P_{X+Y}) \subseteq D_\alpha(P_X) + D_\alpha(P_Y)$, где под суммой глубинно-упорядоченных регионов понимается сумма Минковского (покомпонентно).

Рассмотрим здесь же используемые в статье нотации симметрии распределения. Величина $x \in R^d$ называется центрально симметричной (C-симметричной) относительно точки θ, если $x - \theta \overset{d}{=} \theta - x$, где $\overset{d}{=}$ означает равенство по распределению. $x \in R^d$ называется симметричной в угловом отношении (A-симметричной) относительно точки θ, если величина $\frac{(x-\theta)}{|x-\theta|}$

центрально симметрична относительно начала координат. $x \in R^d$ называется симметричной в полупространственном отношении (H-симметричной) относительно точки θ, если $P(x \in H) \geq \frac{1}{2}$ для любого замкнутого полупространства H, содержащего θ.

3 Полупространственная глубина

Исторически первой считается полупространственная глубина, введенная Тюки в 1975 [5]. Посвятив жизнь изучению выбросов, своей главной задачей американский математик ставил отыскание удобного и интуитивно понятного способа представления данных, позволяющего судить о свойствах неочевидных, прежде не рассматриваемых. По сути, глубина Тюки является обобщением порядка «чем ближе к краю, тем меньше глубина», заданного на одномерной выборке, на евклидово многомерное пространство путем отсечения точек гиперплоскостями. Чтобы сформировать отчетливое представление о глубине вообще и полупространственной глубине в частности попытаемся проследовать философии автора.

Начнем с одномерного случая, задав порядок на выборке, состоящей из n точек: $x_1, x_2, ..., x_n$. Для этого определим симметричную функцию $x_{i|n}$, равную значению i-й точки в упорядоченной по возрастанию последовательности (т. е. $x_{1|n}$ – наименьшее значение в выборке, $x_{2|n}$ – следующее по возрастанию и так до $x_{n|n}$). Таким образом, получим порядковую статистику $x_{1|n} \leq x_{2|n} \leq ... \leq x_{n|n}$. Глубину точки $x_{i|n}$ определим как наименьшее из i и $(n+1-i)$, что соответствует номеру точки в упорядоченной последовательности при счете от края; тогда глубина медианы равна $\frac{1}{2}(1+n)$ (рис. 1).

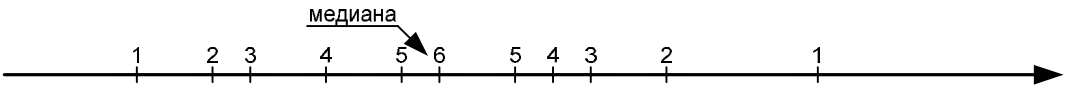

Рис. 1 – Глубина Тюки для 11 одномерных точек

Следуя индукции, распространим понятие глубины на плоскость. Будем отсекать точки от выборки направленной прямой (i, j), слева от которой или на ней находится не меньше i точек, а строго слева от нее лежат не больше j точек. Множество прямых $(i, i-2)$ формирует замкнутый многоугольник (так называемый $\left(i-\frac{1}{2}\right)$-полигон), по сути являющийся оболочкой $(i-1)$-региона, так как прямая, проведенная через любую точку, находящуюся внутри, отсекает от выборки не меньше $(i-1)$ точек. На рис. 2 показаны $1\frac{1}{2}$ (сплошной линией –

граница региона глубины 1), $2\frac{1}{2}$ (штрихпунктирной линией – граница региона глубины 2) и $3\frac{1}{2}$ (штрихпунктирной линией с двумя точками – граница региона глубины 3) полигоны.

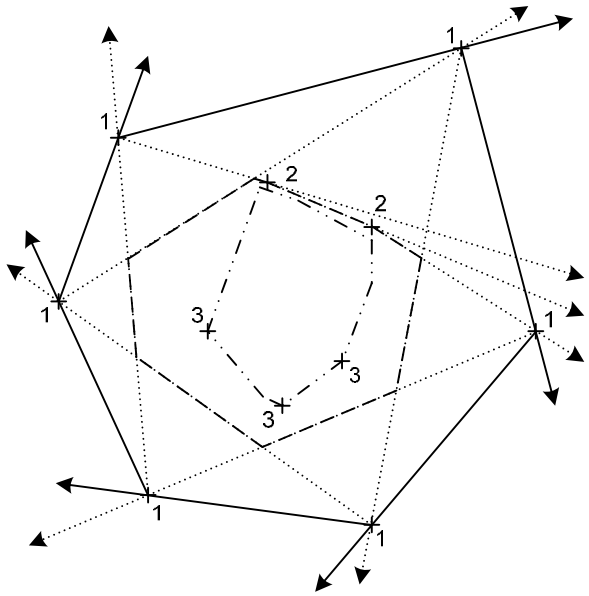

Рис. 2 – Полигоны $1\frac{1}{2}$, $2\frac{1}{2}$ и $3\frac{1}{2}$ (границы регионов глубины 1, 2 и 3)

Теперь обобщим понятие глубины на выборку в d-мерном пространстве. Произведя разделение распределения в R^d гиперплоскостью на два полупространства, обозначим через H то из них, которое содержит точку x, а через $P(H)$ – вероятность принадлежности ему случайной точки. Тогда полупространственная глубина некоторой точки $x \in R^d$ определяется как инфимум вероятности принадлежности случайной точки замкнутому полупространству, содержащему x [6]:

$$HD(x;P) = \inf\left\{P(H) : H \text{ замкнутое полупространство}, x \in H\right\}. \qquad (2)$$

На экспериментальных выборках естественно заменить вероятность порцией точек от всей выборки. Тогда для эмпирической выборки $\{x_1, x_2, ..., x_n\} \in R^d$ глубина Тюки определяется как наименьшая доля точек, которые необходимо отбросить, чтобы выпуклая оболочка оставшихся не включала x:

$$HD_n(x) = \min_{u \in R^d} \#\left\{i : \langle x_i, u\rangle \geq \langle x, u\rangle\right\}, \qquad (3)$$

где # обозначает долю точек от всей выборки. Иначе глубина Тюки определяется как наименьшая доля точек, отсекаемая от выборки гиперплоскостью, проходящей через x (рис. 3).

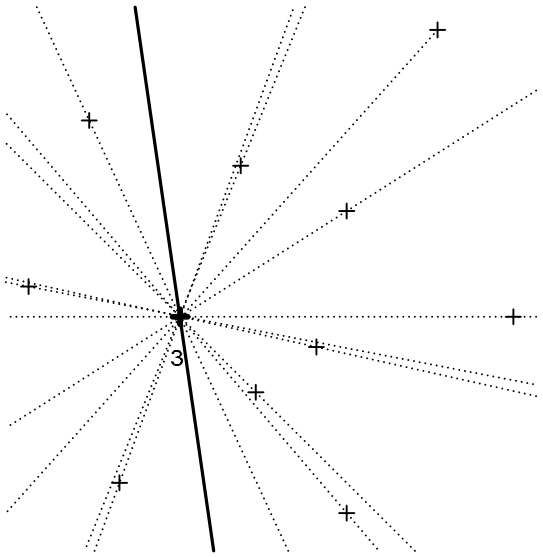

Рис. 3 – Способ определения глубины Тюки

Глубина Тюки выдерживает все требования, предъявляемые к глубине: она инвариантна к аффинному преобразованию, максимальна в точке угловой симметрии (если такая точка существует), убывает при движении от наиболее глубокой точки, стремясь к нулю в бесконечности. Максимально достижимая глубина на абсолютно непрерывных распределениях равна $\max\{HD(x,P)\}=\frac{1}{2}$ и она достигается в точке симметрии А-симметричного распределения [7].

Выше были изложены соображения, используемые при определении глубинно-упорядоченных регионов Тюки на плоскости, и способ их построения. Обобщив их на случай любой размерности, приведем формальное определение: α-регионом по Тюки является пересечение всех замкнутых полупространств, вероятность принадлежности случайной точки которым больше $1-\alpha$:

$$HD_{\alpha}(P)=\bigcap\{H: H \text{ замкнутое полупространство, } P(H)>1-\alpha\}. \qquad (4)$$

Регионы Тюки выдерживают требования аффинной эквивариантности, вложенности, компактности и выпуклости, но в общем случае не удовлетворяют требованиям субаддитивности и монотонности.

4 Классификация функций глубины

В [1] предлагают 4 общие структуры для задания функции глубины в зависимости от способа ее определения.

Глубина класса А определяется как:

$$D(x,P)=E\left[h(x;x_1,...,x_r)\right], \qquad (5)$$

где $x_1,...,x_r$ – случайный набор из P, $h(x;x_1,...,x_r)$ – ограниченная неотрицательная функция, определяющая в каком-то смысле близость точки x к набору точек $x_1,...,x_r$ а $E[\cdot]$ – оператор математического ожидания.

Для эмпирической выборки P_n глубина $D(x, P_n)$ отражается U-статистикой или V-статистикой.

Глубина класса B определяется как:

$$D(x,P)=\frac{1}{1+E\left[h(x;x_1,...,x_r)\right]}, \tag{6}$$

где $x_1,...,x_r$ – случайный набор из P, а $h(x;x_1,...,x_r)$ – неограниченная неотрицательная функция, определяющая в каком-то смысле расстояние от точки x до набора точек $x_1,...,x_r$. Близким, но не эквивалентным, определением глубины типа B является:

$$D(x,P)=E\left[\frac{1}{1+h(x;x_1,...,x_r)}\right]. \tag{7}$$

Глубина типа C определяется как:

$$D(x,P)=\frac{1}{1+h(x;P)}, \tag{8}$$

где $h(x;P)$ – мера отклонения точки $x\in R^d$ от центральной или наиболее глубокой точки распределения P, обычно неограниченная функция.

Глубина типа D определяется как:

$$D(x;P,b)\equiv\inf_{C}\left\{P(C)\middle|x\in C\in b\right\}, \tag{9}$$

где b – класс замкнутых подпространств в R^d, P – распределение в R^d, и отражает степень принадлежности хвосту распределения. Тогда b-глубина точки $x\in R^d$ по отношению к распределению P в R^d определяется как минимальная доля вероятности, охваченная множеством C, принадлежащим классу b и включающем x.

5 Симплексная глубина

Симплексная глубина [2] может быть получена как функция глубины типа A в случае $r=d+1$ и $h(x;x_1,...,x_r)=I\left(x\in S\left[x_1,...,x_{d+1}\right]\right)$, где $S\left[x_1,...,x_{d+1}\right]$ – симплекс, построенный на точках $x_1,...,x_{d+1}$ в R^d, $I(\cdot)$ – индикаторная функция, равная единице, если высказывание в скобках верно, и нулю в противном случае.

Рассмотрим одномерный случай, где симплексная глубина точки x определяется как вероятность ее принадлежности отрезку, соединяющему две случайные точки выборки. Заменив вероятность долей отрезков, которым принадлежит точка x, среди соединяющих каждую пару различных точек, для выборки из n точек $x_1, x_2,..., x_n$ имеем:

$$SD_n(x)=\frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} I\left(x\in\overline{x_i x_j}\right), \tag{10}$$

где $\overline{x_i x_j}$ – отрезок, соединяющий точки x_i и x_j (рис. 4, на выборке из шести точек $SD_n(x)=\frac{8}{15}$).

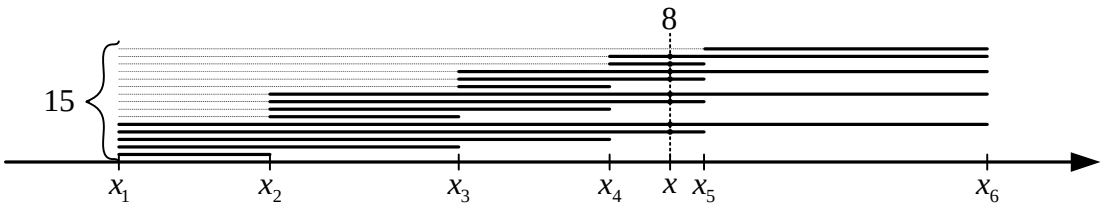

Рис. 4 – Способ определения симплексной глубины для точки x на выборке из 6 точек на прямой

На плоскости симплексом является треугольник, и глубина точки x определяется как вероятность ее принадлежности замкнутому треугольнику, построенному на трех случайных точках выборки (т. е. доля всевозможных замкнутых треугольников, построенных на точках выборки, которые включают точку x):

$$SD_n(x)=\frac{1}{\binom{n}{3}}\sum_{1\le i<j<k\le n} I\left(x\in\Delta\left(x_i,x_j,x_k\right)\right), \qquad (11)$$

где $\Delta\left(x_i,x_j,x_k\right)$ – замкнутый треугольник, построенный на точках x_i, x_j и x_k. Пусть каждый треугольник вырезан из полупрозрачной серой пленки толщиной $\binom{n}{3}^{-1}$. Разместим их один над одним слоями в соответствии с точками, на которых они построены, на белой плоскости, чтобы получить очертания регионов и иметь возможность определить глубину каждой точки как суммарную толщину пленки, ее перекрывающей (рис. 5).

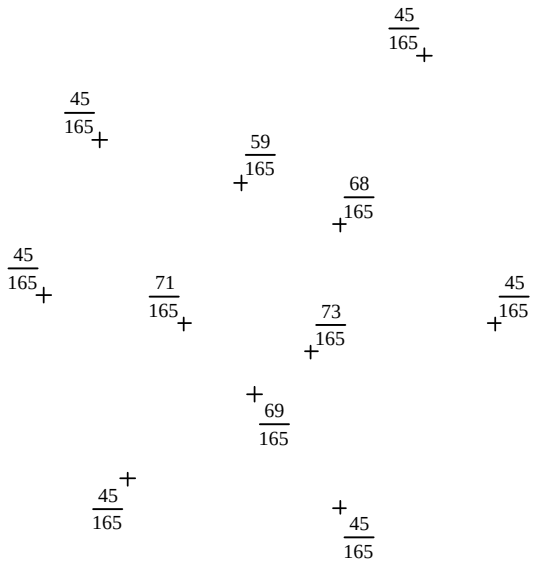

Рис. 5 – Способ определения симплексной глубины для выборки из 11 точек на плоскости

Обобщением треугольника в d-мерном евклидовом пространстве является симплекс или d-мерный тетраэдр, обозначенный нами выше как $S[x_1,...,x_{d+1}]$. Тогда симплексная глубина некоторой точки $x \in R^d$ относительно распределения определяется как вероятность принадлежности x произвольному замкнутому симплексу в R^d, построенному на точках распределения:

$$SD(x;P) = P\left(x \in S[x_1,...,x_{d+1}]\right). \quad (12)$$

Для эмпирической выборки $\{x_1, x_2,..., x_n\} \in R^d$ симплексная глубина определяется U-статистикой как доля замкнутых симплексов, построенных на точках выборки, включающих x:

$$SD_n(x) = \frac{1}{\binom{n}{d+1}} \sum_{1 \le i_1 < ... < i_{d+1} \le n} I\left(x \in S\left[x_{i_1},...,x_{i_{d+1}}\right]\right), \quad (13)$$

Точка x принадлежит симплексу $S\left[x_{i_1},...,x_{i_{d+1}}\right]$, если существует единственное решение системы из $d+1$ уравнений и все корни $\alpha_{i_1} \ge 0, \alpha_{i_2} \ge 0,..., \alpha_{i_{d+1}} \ge 0$:

$$\begin{cases} x = x_{i_1}\alpha_{i_1} + x_{i_2}\alpha_{i_2} + ... + x_{i_{d+1}}\alpha_{i_{d+1}} \\ \alpha_{i_1} + \alpha_{i_2} + ... + \alpha_{i_{d+1}} = 1 \end{cases}. \quad (14)$$

Из предъявляемых симплексная глубина выдерживает требования аффинной инвариантности, максимальности в точке угловой симметрии, стремления к нулю в бесконечности, однако монотонно убывает от точки угловой симметрии только на непрерывных распределениях. Для дискретных H-симметричных распределений симплексная глубина может не удовлетворять требованию максимальности в центре, а для C-симметричных распределений – даже требованию монотонности [1].

Построенные с помощью симплексной глубины упорядоченные регионы являются аффинно эквивариантными, вложенными, связными для абсолютно непрерывных распределений, но, в общем случае, не являются выпуклыми. В случае не абсолютно непрерывного распределения упорядоченные регионы симплексной глубины могут не удовлетворять даже требованию связности.

Индикаторный тип функции глубины, основанной на U-статистике, находит свое отражение в специфической форме эквиглубинных зон, сохраняющих размерность пространства, объясняет факт неудовлетворения требований выпуклости или даже связности глубинно-упорядоченных регионов на конечных выборках. Функция глубины может не убывать монотонно от центра, быть мультимодальной, будет иметь разрывы, и, являясь функцией с ограниченным изменением, представляет собой сумму функции скачков и сингулярной функции [8]. По этой же причине изменение координат векторов выборки в некоторых пределах не вызывает изменения значения функции

глубины в определенной точке, что отличает ее от нотаций, рассмотренных ниже.

6 Глубина Ойя

Представителем глубины типа B является глубина Ойя [9], основанная на расстоянии от точки до эмпирического облака (расстоние Ойя). Расстояние Ойя – результат попытки обобщения известных в классической статистике свойств распределения (моментов) на многомерный случай с использованием симплексов, работа с которыми подробно описана в предыдущем разделе.

Для точки $x \in R^d$ C-симметричных распределений расстояние Ойя определяется как средний d-мерный объем симплекса в R^d, построенного на d точках выборки и точке x:

$$D_n(x) = \frac{1}{\binom{n}{d}} \sum_{1 \le i_1 < \ldots < i_d \le n} v\left(x, x_{i_1}, \ldots, x_{i_d}\right), \tag{15}$$

где $v\left(x, x_{i_1}, \ldots, x_{i_d}\right)$ – площадь (в общем случае d-мерный объем) симплекса, построенного на точках $x, x_{i_1}, \ldots, x_{i_d}$ (рис. 6). На рисунке показан способ построения симплексов (треугольников) для точки, обозначенной жирным крестиком, а также подписаны расстояние Ойя (вверху) и глубина (внизу) для каждой точки.

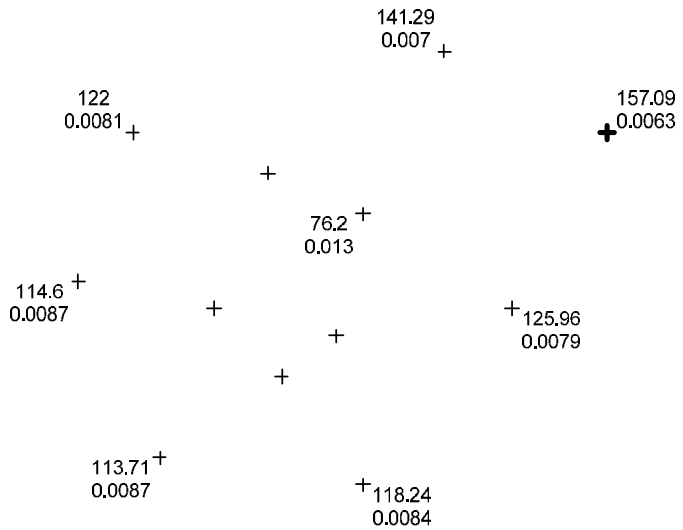

Рис. 6 – Способ построения симплексов (треугольников) для определения глубины Ойя на выборке из 11 точек на плоскости

d-мерный объем симплекса может быть вычислен по формуле:

$$v\left(x_1, x_2, \ldots, x_{d+1}\right) = abs\left(\frac{1}{d!}\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{2,1} & \cdots & x_{d+1,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{d+1,2} \\ \vdots & \vdots & & \vdots \\ x_{1,d} & x_{2,d} & \cdots & x_{d+1,d} \end{vmatrix}\right), \tag{15}$$

где $x_{i,j}$ – значение j-й координаты i-й точки, abs обозначает модуль числа, введен для отличия от обозначения определителя чтобы избежать путаницы.
Привлекательная особенность глубины – связь между медианой и центром гравитации распределения [10]. В [1] исследовали свойства глубины Ойя и доказали ряд теорем для глубин типа B и C.

7 Глубина Махаланобиса

Типичным представителем глубины типа C является глубина Махаланобиса, введенная в [1], основанная на расстоянии Махаланобиса, являющемся мерой отклонения точки от центра эмпирического облака и учитывающем его статистические свойства (математическое ожидание, матрица ковариации).

Расстояние Махаланобиса [11] между точками x и y по отношению к положительно определенной матрице M размерности $d \times d$ определяется следующим образом:

$$d_M^2(x,y) = (x-y)' \cdot M^{-1} \cdot (x-y). \qquad (16)$$

Тогда непосредственно глубина Махаланобиса на основании расстояния определяется как:

$$MHD(x;P) = \frac{1}{1+d_{\sum P}^2\left(x,\mu(P)\right)}, \qquad (17)$$

где $\mu(P)$ – среднее значение распределения P, для эмпирической выборки определяется по формуле $\mu = \sum_{i=1}^{n} x_i$, $\sum P$ – матрица ковариации распределения P, определяемая как:

$$\sum = \begin{bmatrix} E\left[\left(x_{(1)}-\mu_{(1)}\right)\left(x_{(1)}-\mu_{(1)}\right)\right] & E\left[\left(x_{(1)}-\mu_{(1)}\right)\left(x_{(2)}-\mu_{(2)}\right)\right] & \cdots & E\left[\left(x_{(1)}-\mu_{(1)}\right)\left(x_{(n)}-\mu_{(n)}\right)\right] \\ E\left[\left(x_{(2)}-\mu_{(2)}\right)\left(x_{(1)}-\mu_{(1)}\right)\right] & E\left[\left(x_{(2)}-\mu_{(2)}\right)\left(x_{(2)}-\mu_{(2)}\right)\right] & \cdots & E\left[\left(x_{(2)}-\mu_{(2)}\right)\left(x_{(n)}-\mu_{(n)}\right)\right] \\ \vdots & \vdots & & \vdots \\ E\left[\left(x_{(n)}-\mu_{(n)}\right)\left(x_{(1)}-\mu_{(1)}\right)\right] & E\left[\left(x_{(n)}-\mu_{(n)}\right)\left(x_{(2)}-\mu_{(2)}\right)\right] & \cdots & E\left[\left(x_{(n)}-\mu_{(n)}\right)\left(x_{(n)}-\mu_{(n)}\right)\right] \end{bmatrix},$$

где $x_{(j)}$ и $\mu_{(j)}$ – значение j-й координаты произвольного вектора и среднего значения выборки соответственно.

На рис. 7 приведены контуры глубинно-упорядоченных регионов, построенных при помощи глубины Махаланобиса, для трех разных распределений.

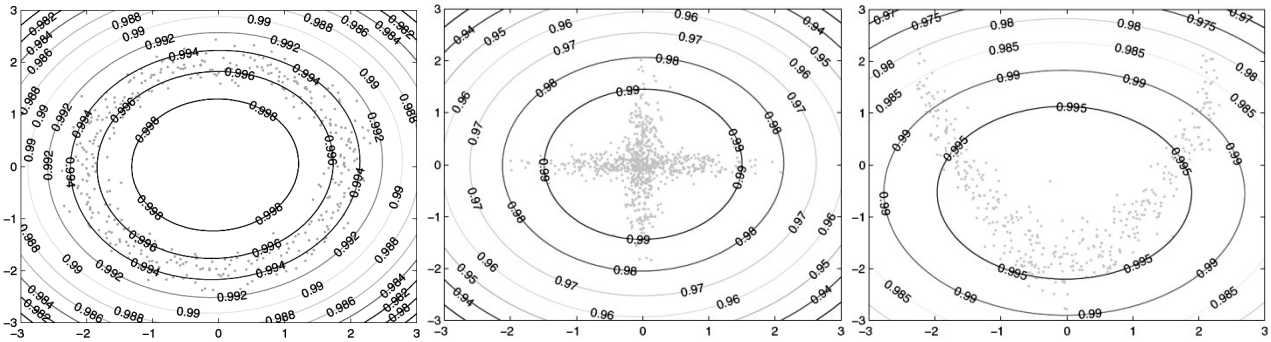

Рис. 7 – Контуры глубинно-упорядоченных регионов, построенных при помощи глубины Махаланобиса для трех разных распределений на плоскости

Привлекательной особенностью глубины Махаланобиса считается простота вычисления, однако ее использование ограничено применимостью лишь для эллиптических, или, по крайней мере, выпуклых распределений из-за эллиптической формы регионов. Глубина Махаланобиса не робастна и может не достигать максимума в центре А-симметричных распределений. Указанных недостатков можно избежать применением обобщенной функции глубины Махаланобиса [12].

8 Глубина зоноида

Наиболее совершенной формой функции глубины и глубинно-упорядоченных регионов можно назвать глубину зоноида, представленную рядом трудов ученых Кёльнского университета, подытоженных в книге [13]. Регионом (зоноидом) глубины α для распределений с конечным математическим ожиданием называется [14]:

$$D_\alpha(P)=\left\{\int\limits_{R^d} x\cdot g(x)dP(x) : g:R^d\to\left[0,\frac{1}{\alpha}\right] \text{ измерима и } \int\limits_{R^d} g(x)dP(x)=1\right\}, \qquad (19)$$

а для $\alpha=0$:

$$D_0(P)=cl\left(\bigcup_{\alpha\in(0,1]} D_\alpha(P)\right), \qquad (20)$$

где cl – замыкание.

Для эмпирической выборки можно записать (именно эта формула используется для вычисления глубины):

$$D_{n_\alpha}=\left\{\sum_{i=1}^{n}\lambda_i\cdot x_i : \sum_{i=1}^{n}\lambda_i=1, 0\le\lambda_i, \alpha\cdot\lambda_i\le\frac{1}{n} \text{ для всех } i\right\}, \qquad (21)$$

или, для $\alpha\in\left[\frac{k}{n},\frac{k+1}{n}\right], k=1,...,n-1$, получаем:

$$D_{n_\alpha}=conv\left\{\frac{1}{\alpha\cdot n}\sum_{j=1}^{k}x_{i_j}+\left(1-\frac{k}{\alpha\cdot n}\right)x_{i_{k+1}} : \{i_1,...,i_{k+1}\}\subset N\right\}, \qquad (22)$$

где $N=\{1,...,n\}$, а для $\alpha\in\left[0,\frac{1}{n}\right]$:

$$D_\alpha(P) = conv\{x_1,...,x_n\} = \bigcap\left\{h \in H^d : P(h) = 1\right\}, \quad (23)$$

где H^d – множество всех замкнутых пространств в R^d.

Очевидно регионом глубины $\alpha = 1$ является единственная точка – математическое ожидание; регионом для $\alpha \in \left[0, \frac{1}{n}\right]$ считается выпуклая оболочка выборки.

Важной особенностью регионов является их порождение лифт-зоноидом – геометрической фигурой, тесно связанной с кривой Лоренца и однозначно характеризующей распределение. Кривая Лоренца это графическое изображение кумулятивной функции распределения, изначально используемое как показатель неравенства в доходах населения и характеризующее зависимость между долями дохода и численности населения. Каждая точка кривой показывает процент дохода, сосредоточенный в руках определенной доли наименее обеспеченного населения.

Лифт-зоноид для эмпирической выборки $x_1, x_2, ..., x_n$ определяется как математическое ожидание отрезка $\left[(0,0),(1,x)\right]$, т. е. сегмента линии, соединяющей начало координат и точку выборки, к которой добавлена одна координата со значением 1 (лифтинг). Тогда лифт-зоноид определяется как:

$$\hat{Z}(P) = \sum_{i=1}^{n}\left[(0,0),\left(\frac{1}{n},\frac{x_i}{n}\right)\right]. \quad (24)$$

Очевидно, кривая Лоренца соответствует верхней части лифт-зоноида одномерной выборки при условии, что абсцисса масштабирована посредством умножения на $\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1}$. Зоноид – это масштабированная проекция лифт-зоноида, спроектированная на уровне α дополнительной координаты:

$$D_\alpha(P) = \frac{1}{\alpha} proj_\alpha\left(\hat{Z}(P)\right). \quad (25)$$

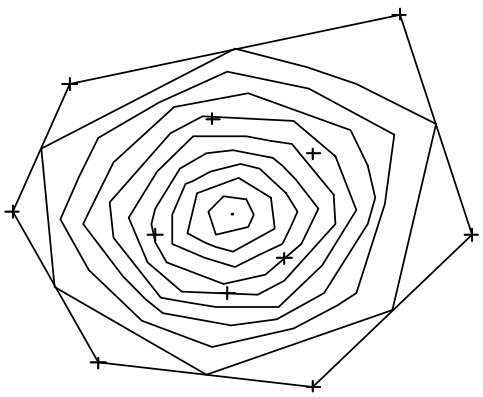

Рис. 8 – Примеры зоноидов глубины $\alpha = \frac{1}{11}, \frac{2}{11}, \frac{3}{11}, \frac{4}{11}, \frac{5}{11}, \frac{6}{11}, \frac{7}{11}, \frac{8}{11}, \frac{9}{11}, \frac{10}{11}, 1$ для выборки из 11 точек на плоскости

Примеры регионов глубины $\alpha = \frac{1}{11}, \frac{2}{11}, \frac{3}{11}, \frac{4}{11}, \frac{5}{11}, \frac{6}{11}, \frac{7}{11}, \frac{8}{11}, \frac{9}{11}, \frac{10}{11}, 1$ приведены на рис. 8. Зоноиды полностью удовлетворяют всем требованиям глубины, более того, в ходе исследований были доказаны некоторые дополнительные свойства:

7) Уникальность. Семейство зоноидов однозначно характеризует распределение: пусть $D_\alpha(P_X) = D_\alpha(P_Y)$ для всех $\alpha \in (0,1]$, тогда $P_X = P_Y$.

8) Симметричность. Зоноид половинной глубины $D_{0.5}(P)$ симметричен относительно математического ожидания $E[P]$.

9) Непрерывность по α: Пусть для некоторого распределения P существует последовательность α_n на $\alpha \in (0,1]$, сходящаяся к α: $\alpha_n \to \alpha$, $\alpha > 0$. Тогда $D_{\alpha_n}(P) \xrightarrow{H} D_\alpha(P)$, где $\xrightarrow{H}$ – сходимость по расстоянию Хаусдорфа.

10) Непрерывность по распределению. Пусть существует $\alpha \in (0,1]$ и последовательность распределений $P^{(n)}$ равномерно интегрируемая и слабо сходящаяся к P. Тогда $D_\alpha\left(P^{(n)}\right) \xrightarrow{H} D_\alpha(P)$.

На основании зоноидов определим глубину точки как:

$$D(x,P) = \begin{cases} \sup\{\alpha : x \in D_\alpha(P)\}, & \text{если } \exists \alpha : \mathrm{x} \in \mathrm{D}_\alpha(P), \\ 0, & \text{в остальных случаях.} \end{cases} . \qquad (26)$$

Свойства функции глубины дополняются следующими:

2) Максимальность (равенство единице) в точке математического ожидания.

5) Непрерывность по x:
Функция $x \to D(x,P)$ непрерывна для всех $x \in conv(P)$.

6) Непрерывность по распределению:
Функция $P \to D(x,P)$ непрерывна для всех $(x_1,...,x_n) \in \operatorname{int} conv\{P\}$.

7) Монотонность по отношению к растяжению:
$D(x,P_X) \le D(x,P_Y)$ если P_Y является растяжением P_X.

8.1 Вычисление глубины зоноида

Глубина может быть вычислена как результат решения задачи минимизации линейного программирования согласно формуле (21) [15]. Тогда задача линейного программирования имеет следующую постановку.

Для матрицы $X = (x_1,...,x_n)$, в которой колонками выступают точки $x_1,...,x_n$ пространства R^d, векторов-столбцов $\lambda = (\lambda_1,...,\lambda_n)'$, $1 = (1,...,1)'$, $0 = (0,...,0)'$ формулировка задачи линейного программирования для отыскания глубины точки $y \in R^d$ выглядит следующим образом:

$$\left.\begin{array}{ll} \text{Minimize } \gamma & \\ \text{subject to} & X\lambda = y \\ & \lambda' 1 = 1 \\ & \gamma 1 - \lambda \geq 0 \\ & \lambda \geq 0 \end{array}\right\}, \tag{27}$$

где $\lambda_1,...,\lambda_n$ и γ – переменные. При минимальном γ^* глубина определяется по формуле:

$$D\left(y \middle| x_1,...,x_n\right) = \frac{1}{n \cdot \gamma^*}. \tag{28}$$

Задача довольно просто решается при помощи симплекс метода. Отсутствие решения означает, что $y \notin conv\{x_1,...,x_n\}$.

На выборках большого объема целесообразно ускорить расчет глубины, задав специальную структуру ограничений для декомпозиции Данцига-Вульфа и переписав задачу линейного программирования в виде:

$$\left.\begin{array}{ll} \text{Minimize } \gamma & \\ \text{subject to} & X\lambda = y \\ & \left(\lambda_1,...,\lambda_n,\gamma\right)' \in S \end{array}\right\}, \tag{29}$$

где $S = \left\{\left(\lambda_1,...,\lambda_n,\gamma\right)' \in R^{n+1} : \sum_{i=1}^{n} \lambda_i = 1, 0 \leq \lambda_i \leq \gamma \leq 1, \text{ для всех } i\right\}$.

Так как S является выпуклым многогранником, то каждая принадлежащая ему точка может быть подана в виде выпуклой комбинации экстремальных точек, которые однозначно определены и принадлежат множеству V:

$$V = \left\{\frac{1}{|I|}\left(\delta_I, 1\right)' : 0 \neq I \subset \{1,...,n\}\right\}, \tag{30}$$

где $\delta_I = \left(\delta_I(1),...,\delta_I(n)\right)$, $\delta_I(k) = \begin{cases} 1, \text{ если } k \in I, \\ 0, \text{ если } k \notin I. \end{cases}$.

8.2 Вычисление зоноида

Так как зоноид является выпуклым многогранником, он однозначно задается набором вершин (граней). Рассмотрим подробно вычисление координат вершин зоноида на плоскости, то есть для двухмерной выборки [16].

Функция поддержки $h_C : S^{d-1} \to R$ компактного выпуклого множества $C \subset R^d$ определяется как $h_C(p) = \max\{p'x : x \in C\}$. Обозначим через H множество направлений p, для которых существуют такие $x_i \neq x_j$, что $p'x_i = p'x_j$. Тогда для направления $p \in S^{d-1} \setminus H$ через $r_p(x_i)$ будем обозначать позицию точки x_i в

упорядоченной по величине $p'x_j$, $j=1,...,n$ последовательности. Пусть $x_{p,\alpha}$ точка:

$$x_{p,\alpha}=\frac{1}{n\alpha}\sum_{i=1}^{n}\lambda_i^p x_i\,, \tag{31}$$

где $\lambda_i^p=\begin{cases}1, & \text{если } r_p(x_i)>n-[n\alpha],\\ n\alpha-[n\alpha], & \text{если } r_p(x_i)=n-[n\alpha],\\ 0, & \text{если } r_p(x_i)<n-[n\alpha].\end{cases}$

Тогда следующая теорема позволяет определить множество экстремальных точек:

Теорема. Пусть $x_1,...,x_n\in R^d$ попарно различны. Тогда множеством экстремальных точек зоноида глубины α будет:

$$\left\{x_{p,\alpha}: p\in S^{d-1}\setminus H\right\}. \tag{32}$$

Для последовательности длины n каждое направление $p\in S^{d-1}\setminus H$ задает перестановку π_p точек x_i, $i=1,...,n$, для которой:

$$p'x_{\pi_p(1)}<p'x_{\pi_p(2)}<...<p'x_{\pi_p(n)}. \tag{33}$$

Эта перестановка изменяется только в том случае, если направление p, вращаясь, проходит через точки из множества H. Для некоторых направлений множество H может содержать несколько наборов коллинеарных точек, т. е. точек, имеющих одинаковую проекцию на p. Когда p проходит через такие наборы, порядок точек каждого из них в перестановке изменяется на противоположный, как показано на рис. 9 для 5 точек выборки. На рис. 9 для направления p_1 точки в перестановке упорядочены следующим образом: x_1,x_2,x_3,x_4,x_5, для направления p_2 все 5 точек входят в множество H, а для направления p_3 наблюдается изменение порядка в наборах на противоположный: x_2,x_1,x_5,x_4,x_3.

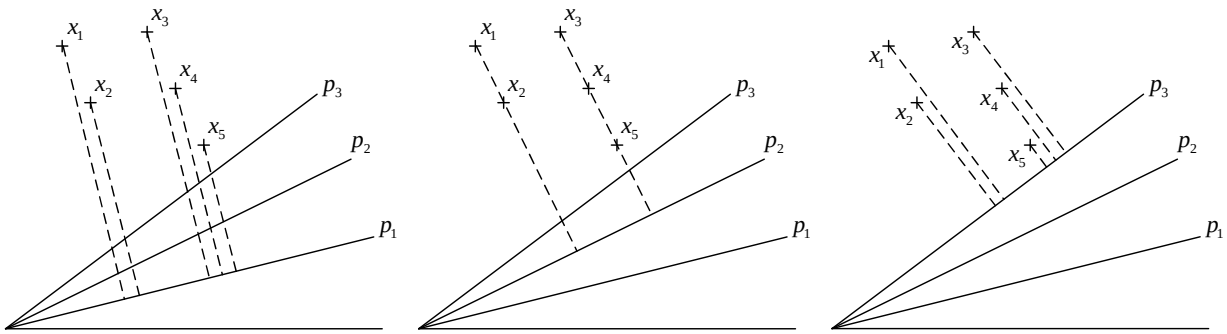

Рис. 9 – Перестановки для 5 точек выборки для направлений p_1, p_2 и p_3

Приведем точный алгоритм для вычисления вершин зоноида глубины α на двухмерной выборке. Алгоритм, по сути, проходит все направления вектора p, которые находятся под углом от 0 до π к «первой» координатной оси, но с

целью ускорения перебора останавливается только на тех направлениях, для которых множество H не пустое.

1) Сохраняем все точки в массиве, отсортированном по принципу:

$$\left(x_{i,1}, x_{i,2}\right) < \left(x_{j,1}, x_{j,2}\right) \Leftrightarrow \left(x_{i,1} < x_{j,1}\right) \text{ или } \left(x_{i,1} = x_{j,1} \text{ и } x_{i,2} > x_{j,2}\right).$$

Определяем два массива *ORDER* и *RANKS*, хранящие текущие значения $\pi_p(i)$ и $r_p(i)$, инициализированные значениями $1,\ldots,n$. Именно с ними мы и будем работать, изменяя порядок наборов коллинеарных точек, принадлежащих множеству H, на противоположный при прохождении p соответствующего угла. Условное вращение p будем осуществлять в направлении против часовой стрелки относительно «первой» координатной оси. Это объясняет, почему изначально точки отсортированы по возрастанию относительно этой оси. Сортировка относительно «второй» координатной оси объясняется следующим образом: если будут наборы коллинеарных точек, входящих в H для направления p, совпадающего с «первой» координатной осью, то при дальнейшем вращении p порядок точек в них должен измениться на противоположный, выстроив их по возрастанию относительно «второй» координатной оси, а значит, изначально внутри наборов точки должны быть отсортированы относительно этой оси по убыванию.

2) Для каждой пары точек x_i, x_j вычисляем угол между «первой» координатной осью и направлением, нормальным к отрезку, их соединяющему, всегда выбирая только угол в пределах $[0,\pi)$. Сохраняем результаты вычислений в массиве *ANGLE* записей, содержащих значение угла и индексы точек i и j, отсортированному по значению угла. По сути, каждая запись представляет собой угол, при прохождении которого изменяется порядок следования точек в перестановке (не исключены повторения, т. е. записи с одинаковым значением угла).

3) Из записей массива *ANGLE* для каждого угла формируем наборы коллинеарных точек, руководствуясь тем фактом, что точки x_i и x_j попадают в один такой набор тогда и только тогда, когда их индексы i и j находятся в одной записи массива *ANGLE*. Для каждого такого набора изменяем порядок индексов в массивах *ORDER* и *RANKS* на противоположный. Набор представляет собой множество точек, входящих в H для данного угла p и изменяющих номер по порядку в перестановке при прохождении через это направление.

4) Исходя из формулы (31) вершина зоноида определяется крайними $[n\alpha]$ или $[n\alpha]+1$ точками в перестановке для вектора p. При изменении направления p изменяют свое положение и точки в перестановке, а значит, может измениться и набор точек, определяющих координаты вершины. Изменение положения точек внутри набора не должно повлечь добавления вершины к зоноиду, тогда как изменение положения точек на границе набора

изменит и сам набор, а значит, к зоноиду должна быть добавлена новая вершина.

Пусть $k_- = \lceil n\alpha \rceil + 1$, $k_+ = n - \lceil n\alpha \rceil$ и $\delta = \left(n\alpha - \lceil n\alpha \rceil\right)$. Проверяем, содержит ли одна из перестановок, сделанных на предыдущем шаге, индекс k_+, если $n\alpha$ не целое, или индексы k_+ и $k_+ + 1$, если $n\alpha$ целое. В таком случае вычисляем новую вершину зоноида:

$$x_{p,\alpha} = \frac{1}{n\alpha}\left(\sum_{i=k_+ +1}^{n} x_{\pi_p(i)} + \delta x_{\pi_p(k_+)} \right). \tag{34}$$

Аналогично, если одна из перестановок содержит индекс k_-, если $n\alpha$ не целое, или индексы k_- и $k_- - 1$, если $n\alpha$ целое, то вычисляем следующую вершину зоноида:

$$x_{-p,\alpha} = \frac{1}{n\alpha}\left(\sum_{i=1}^{k_- -1} x_{\pi_p(i)} + \delta x_{\pi_p(k_-)} \right). \tag{35}$$

5) Продолжаем выполнение шагов 3 – 5 пока в массиве *ANGLE* еще содержатся необработанные записи.

Таким образом, сгенерировано две последовательности точек $x_{p,\alpha}$ и $x_{-p,\alpha}$ в порядке обхода против часовой стрелки. Объединив их, получаем последовательность всех вершин зоноида.

Основным недостатком алгоритма можно считать способ перебора по круговой последовательности, который не имеет явного обобщения на случаи высшей размерности. Расширение алгоритма на пространство размерности $d \geq 3$ приведено в [17].

9 Выводы

В данной статье изложены понятия глубины и глубинно-упорядоченных регионов, приведены их основные свойства и классификация. Изложенный математический аппарат, отражающий положение, дисперсию и форму распределения, представляет собой мощный инструмент для непараметрического анализа данных. Сфера его применения охватывает кластерный анализ, проверку критерия согласия, оценки положения и разброса, а также оценку рисков в эконометрии, где особенный интерес представляют субаддитивные регионы, применимые для работы с когерентными мерами риска.

Первой попыткой упорядочивания многомерных данных считается глубина Тюки, порождающая обобщающий порядок на евклидовом пространстве как более удобный способ представления данных. В статье рассмотрены характерные функции глубины разных типов и способов построения регионов (на основе полупространств, симплексов, средневзвешенного среднего), проведено сравнение на предмет выполнения предъявляемых требований.

Наиболее совершенной и привлекательной в плане приложений и поддерживаемых свойств можно считать глубину зоноида. Зоноиды обладают ценными свойствами непрерывности, субаддитивности и монотонности, хотя глубина зоноида не робастна и может не удовлетворять требованию максимальности в центре для A- и H-симметричных распределений. Преимущества в вычислительном смысле дают точные алгоритмы вычисления глубины и построения зоноидов в пространстве любой размерности.

Литература

1. Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. Ann. Statist. 28, 461-482.
2. Liu, R. (1990). On a notion of data depth based on random simplices. Ann. Statist. 18, 405-414.
3. Zuo, Y. and Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. Ann. Statist. 28, 483-499.
4. Cascos, I. Data depth: multivariate statistics and geometry. Chapter 1, 1-24.
5. Tukey, J.W. (1975). Mathematics and the picturing of data. Proc. Int. Cong. Math., Vancouver, 1974, Vol. 2, 523-531.
6. Rousseeuw, P.J. and Ruts, I. (1999). The depth function of a population distribution. Metrika 49, 213-244.
7. Rousseeuw, P.J. and Struyf, A. (2004). Characterizing angular symmetry and regression symmetry. J. Stat. Plann. Inference 122, 161-173.
8. Колмогоров А.Н., Фомин С.В. Элементты теории функций и функционального анализа, 4-е изд., М.: Наука, 1976
9. Oja, H. (1983). Descriptive statistics for multivariate distributions. Statist. Probab. Lett. 1, 327-332.
10. Chen, D., Devillers, O., Iacono, J., Langerman, S., Morin P. (2010). Oja Medians and Centers of Gravity. CCCG 2010, 147-150.
11. Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proc. Nat. Acad. Sci. India 12 49–55.
12. Hu, Y., Wang, Y., Wu, Y, Li, Q., Hou, C. (2009). Generalized Mahalanobis depth in the reproducing kernel Hilbert space. Statistical Papers, Springer-Verlag 2009.
13. Mosler, K. (2002). Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach. Springer, New York.
14. Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. Anals of Statistics, 25, 1998-2017.
15. Dyckerhoff, R., Koshevoy, G., Mosler, K. (1996). Zonoid data depth: Theory and computation. In: Pratt, A. (Ed.), Proceedings in Computational Statistics. COMPSTAT 1996. Physica, Heidelberg, pp. 235-240.
16. Dyckerhoff, R. (2000). Computing zonoid trimmed regions of bivariate data sets. In: Bethlehem, J., van der Heijden, P. (Eds.), Proceedings in Computational Statistics. COMPSTAT 2000. Physica, Heidelberg, pp. 295-300.
17. Mosler, K., Lange, T., Bazovkin, P. (2009). Computing zonoid trimmed regions of dimension d>2. Computational Statistics and Data Analysis. 53, 2500-2510.