

1 УДК 621.391 : 519.218.5

2 © 20?? г. В.А. Любецкий, К.Ю. Горбунов, С.А. Пирогов, Г.А. Хазиев, А.И.

3 Агламазова

4 **КЛАСТЕРИЗАЦИЯ ТОЧЕК МНОГОМЕРНОГО ПРОСТРАНСТВА**  
5 **НА ОСНОВЕ ИДЕОЛОГИИ SEURAT<sup>1</sup>**

6 В статье на математическом уровне обсуждаются современные прикладные  
7 задачи дискретной оптимизации, и для одной из них, кластеризации точек в  
8 многомерном вещественном пространстве, подробно рассмотрен алгоритм её  
9 решения. Характерной особенностью этих задач является большой размер ис-  
10 ходных данных, часто задаваемых матрицами с десятками тысяч строк и десят-  
11 ками тысяч столбцов. Это приводит к проблемам обработки больших данных  
12 в условиях сложного вычислительного алгоритма; при этом нередко требуется  
13 быстро получить достаточно точное решение задачи, поэтому вопрос о слож-  
14 ности алгоритма имеет принципиальное значение. Решение упомянутой зада-  
15 чи кластеризации приводит к трудным математическим проблемам: удалению  
16 скрытых параметров; выделению значимых признаков; переходу к оптималь-  
17 ным и информационно значимым координатам точек; специфическом представ-  
18 лении (картами многообразия) вершин нагруженного графа; выбор функции,  
19 зависящей от текущей кластеризации вершин, максимизация которой приводит  
20 к искомой кластеризации; для каждого кластера выбор признаков, которые ин-  
21 дивидуально характеризуют его; и, наконец, понижения размерности исходных  
22 данных.

23 *Ключевые слова:* оптимальное преобразование графов, эволюция строки вдоль  
24 дерева, оптимальная дискретная кластеризация.

25 **DOI:** 10.31857/S05552923??, **EDN:** ??

26 **§ 1. Введение**

27 В этой статье на математическом уровне изложены постановки ряда современных  
28 (но уже ставших классическими) прикладных задач и для одной из них, кластериза-  
29 ции множества точек в многомерном вещественном пространстве, описан алгоритм

<sup>1</sup> Исследование выполнено за счет гранта Российского научного фонда № 24-44-00099 (<https://rscf.ru/project/24-44-00099/>).

30 решения на основе идеологии Seurat. Этот алгоритм представляет собой пример  
31 современной обработки данных, направленный на решение сложной вычислитель-  
32 ной задачи. Решения поставленных задач широко востребованы и используются в  
33 разных прикладных областях, но мы не касаемся здесь собственно прикладных ас-  
34 пектов. Хотя эвристические алгоритмы для решения рассмотренной задачи класте-  
35 ризации широко применяются, авторам неизвестны цельные математические изло-  
36 жения её решения; в какой-то мере алгоритмы могут быть извлечены из прикладных  
37 работ, однако, там изложения строятся в контексте соответствующих прикладных  
38 областей, что затруднит математически ориентированного читателя.

39 Для таких задач кроме собственно математического исследования требуется най-  
40 ти эффективный алгоритм и доказательство того, что алгоритм действительно на-  
41 ходит заранее математически описанное решение; кроме того, нужно найти оценку  
42 времени работы (сложность) алгоритма; и, в равной мере, найти эффективную ком-  
43 пьютерную реализацию этого алгоритма. Поскольку эти задачи возникают в при-  
44 кладном аспекте, не всегда осознаётся, что компьютерной программе предшествует  
45 математическая постановка и, по меньшей мере, математическое осмысление задачи.  
46 Иными словами, случается, что прикладник пишет эвристическую компьютерную  
47 программу до математических постановки и исследования задачи.

48 В заключение заметим, что наша научно-методическая статья ориентирована на  
49 широкий круг читателей; хотя лучше, если они знакомы с материалом первого курса  
50 математического факультета.

51 Наконец, общим местом является замечание, что алгоритмическая и компьютер-  
52 ная обработка больших данных – универсальное направление исследований и метод  
53 работы буквально во всех областях естественных, инженерных и гуманитарных на-  
54 ук. Такая обработка опирается на методы современной математики (от алгоритмов  
55 – алгебры – анализа до геометрии) и также опирается на современные приёмы эф-  
56 фективного программирования.

## 57 § 2. Примеры задач

58 Начнём с трёх примеров математических задач, хотя методы решения первых  
59 двух из них здесь не обсуждаются (иначе пришлось бы значительно увеличить раз-  
60 мер текста, и возникла тематическая разбросанность), но их математическое обсуж-  
61 дение доступно по ссылкам.

62 1. Даны ориентированные графы с именами рёбер, без повторения имён или с их по-  
63 вторениями. Такие графы иногда называют *структурами*; в сущности, даны две  
64 картинки, показанные на рис. 1. Фиксированы шесть *операций*, преобразующие  
65 одну структуру в другую; каждой операции сопоставлено строго положительное  
66 рациональное (в общем случае, вещественное) число, которое называется *ценой*  
67 данной операции. Цена показывает, сколь редко используется данная операция в

68 процессе, который сам описывается как минимум суммарной цены всех использо-  
 69 зуемых в нём операций (с их повторениями): высокая частота использования  
 70 означает низкую цену операции, а низкая частота использования – её высокую  
 71 цену. Таким образом, обеспечивается адекватность этого процесса. Упомянутые  
 72 шесть операций над структурами хорошо известны и не приводятся здесь, см.  
 73 например, [1]. Итак, задача состоит в том, чтобы найти алгоритм, который по  
 74 двум данным структурам  $a$  и  $b$  и данным ценам операций выдаёт одну из цепочек  
 75 операций, которые последовательно преобразуют  $a$  в  $b$ , и имеет минимальную  
 76 суммарную цену операций.

77 В [2] получен алгоритм, по времени работы линейный от суммарного числа рёбер  
 78 в данных структурах  $a$  и  $b$ , который находит такую, минимальную при данных  
 79 ценах, цепочку операций. Вообще, линейная сложность алгоритма в реальных  
 прикладных задачах – большая редкость.

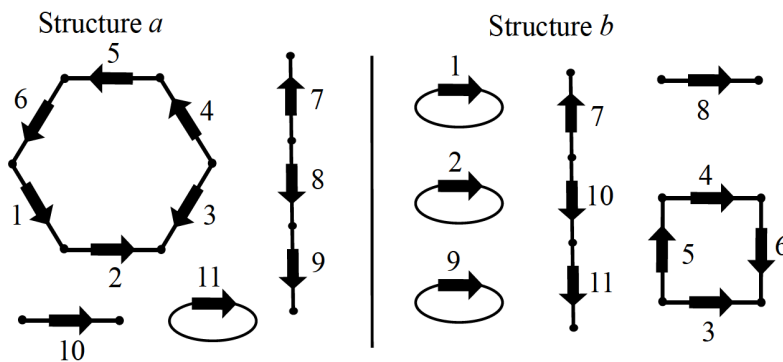


Рис. 1. Даны две структуры, нагруженные ориентированные графы  $a$  и  $b$ , и также цены каждой из шести операций над структурами. Требуется найти цепочку (последовательность) операций с их повторениями, суммарная цена которой минимальна по сравнению со всеми возможными цепочками, начинающимися с  $a$  и заканчивающимися на  $b$ .

80  
 81 2. В этом разделе речь идёт, скорее, о классе задач, для которых известно много эв-  
 82 ристических алгоритмов решения. Дано корневое дерево (иногда, ациклическая  
 83 сеть), для которого заданы длины рёбер, они соответствуют «дискретному вре-  
 84 мени» перехода от «предка» в начале ребра к «потомку» в конце ребра («начало»  
 85 расположено ближе к корню дерева, а «конец» – дальше от него). В листьях даны  
 86 современные однотипные данные из предметной области. Нелистовым вершинам  
 87 однозначно сопоставлены объекты того же типа, какой представлен в листьях, –  
 88 такое сопоставление называется *расстановкой* объектов по нелистовым верши-  
 89 нам; поскольку листьям объекты заранее сопоставлены, расстановка определена  
 90 на всех вершинах. В пространстве всех расстановок задан функционал, который

91 зависит от переменной (в нелистовых вершинах) расстановки и от длин рёбер.  
92 Задача состоит в минимизации функционала. Иными словами, задача состоит в  
93 описании дискретной эволюции «прапредка» – объекта в корне дерева вдоль все-  
94 го дерева вплоть до современных данных в листьях. Примерами объектов служат  
95 последовательности в фиксированном алфавите, которыми, как известно, можно  
96 кодировать любой конечный объект; часто графы заданного типа. Иногда задача  
97 ставится так, что и само дерево является аргументом функционала или, наоборот,  
98 дано дерево и все рёбра считаются одинаковой длины, т.е. длины отсутствуют.  
99 Нами найдены эффективные компьютерные решения ряда задач такого рода и  
100 доказаны соответствующие теоремы, хотя в целом этот важный круг задач далёк  
101 от математического освещения, см., например, [3].

102 Один из важных подходов к определению такой эволюции состоит в следую-  
103 щем. Для данных точек в  $n$ -мерном пространстве задано ещё поле их скоростей,  
104 которое интерпретируется как потенциал развития точки-клетки (см. начало сле-  
105 дующего пункта 3). Поле скоростей преобразуется в марковскую (или близкую  
106 к ней) цепь переходных вероятностей развития точек-клеток. По цепи образуют-  
107 ся нечёткие множества клеток, которые образуют начальные, промежуточные  
108 и финальные макросостояния, и определяются вероятности перехода из одного  
109 макросостояния в другое. Отсюда находится эволюция макросостояний, от на-  
110 чального в финальные, которые интерпретируются как кластеры клеток [4].

111 3. Перейдём к задаче кластеризации точек в многомерном вещественном простран-  
112 стве, решение которой составит оставшуюся часть статьи. Дана числовая мат-  
113 рица, строки которой называются *признаками* (features), а столбцы – *клетками*,  
114 не имея в виду именно биологическую клетку, а любую «изолированную часть  
115 мира». Далее столбцы будут рассматриваться как точки соответствующего про-  
116 странства. Элемент матрицы – вещественное число, которое называется *экспрес-  
117 сией* (выраженностью) признака-строки  $i$  в клетке-столбце  $j$ . *Кластеризацией*  
118 столбцов матрицы называется разбиение её столбцов на непересекающиеся мно-  
119 жества, каждое из которых называется *кластером*. Требуется найти кластери-  
120 зацию, для которой столбцы внутри любого кластера наиболее похожи друг на  
121 друга в сравнении с похожестью столбцов между разными кластерами. «Похо-  
122 жество» двух столбцов определяется функционалом, который не так просто сфор-  
123 мулировать; это будет сделано по ходу решения задачи, что, увы, отступает от  
124 принципа, который пропагандировался во Введении.

125 Многочисленные эвристические решения этой задачи весьма сложны, в качестве  
126 примера мы сошлёмся на её решение в контексте прикладной биоинформатики,  
127 [5]. Мы изложим ниже в некоторых пунктах оригинальное решение, для которого  
128 разработали используемую нами эффективную компьютерную программу.

129 В заключение отметим, что в разделах 3-10 мы следуем общему плану метода Seurat-  
130 Louvain-Leiden. Наша цель – привлечь внимание математически ориентированного

131 читателя к многочисленным математическим проблемам, которые возникают в этом  
132 методе, далёком от математического не только обоснования, но даже отчасти и по-  
133 нимания. Это также относится и к задачам, сформулированным в разделах 1-2.

134 Следующий раздел 3 подробно излагается в разделах 4-10; читателю менее зна-  
135 комому с этим материалом можно вначале пропустить раздел 3.

### 136 § 3. План решение задачи кластеризации столбцов числовой матрицы.

137 Итак, клетка – точка в  $N$ -мерном пространстве вещественных чисел  $\mathbb{R}^N$ . Ины-  
138 ми словами, *координатами* клетки  $j$  называется столбец чисел в данной матрице,  
139 который удобно представить себе “расположенным под клеткой”  $j$ . Число столбцов  
140 в этой и последующих матрицах не меняется и равно  $M$ . Приведём план решения  
141 более или менее общий для разных подходов к этой задаче. Математическое обос-  
142 нование и оценка сложности предлагаемого ниже алгоритма далеки от решения,  
143 хотя активно изучаются. При желании можно прочесть пункты этого плана после  
144 остальных разделов статьи.

145 1. В данной числовой матрице  $X = (x_{ij})$ , размера  $N \times M$ , каждый её элемент  $x_{ij}$   
146 заменим его долей относительно всего столбца  $j$ , так полученную матрицу обо-  
147 значим также  $X$ . Для каждой строки  $i$  матрицы  $X$  вычислим среднее  $x_i$  и выбо-  
148 рочную дисперсию  $\sigma_i^2$ , и поэлементно преобразуем матрицу  $X$  в новую матрицу  
149  $Y$ , в которой  $\sigma_i^2$  уже слабо зависит от  $x_i$ .

150 Напомним, что для любой матрицы размера  $N \times M$  среднее  $i$ -ой строки равно

$$x_i = \frac{1}{M} \sum_{j=1}^M x_{ij},$$

151 а дисперсия строки равна

$$\sigma_i^2 = \frac{1}{M-1} \sum_{j=1}^M (x_{ij} - x_i)^2$$

152 и стандартное отклонение строки равно  $\sigma_i$ .

153 2. По матрице  $Y$  образуем новую матрицу  $Z$ . Для этого по каждой строке  $i \in$   
154  $Y$  построим регрессию  $f$  для множества  $\{(y_l, \sigma_l^2)\}$  точек на плоскости: точка –  
155 дисперсия  $\sigma_l^2$  вместе со средним  $y_l$ , строки  $l$ , где  $y_l$  берётся только из заданной  
156 окрестности  $I_i$  среднего  $y_i$ . Тогда матрица  $Z$  определяется как

$$Z = \frac{y_{ij} - y_i}{f(y_i)}.$$

157 В  $Z$  выберем заданное число строк с наибольшей новой дисперсией  $\delta = \frac{\sigma_i^2}{f(i)}$ .  
158 Обозначим  $\tilde{Z}$  часть матрицы  $Z$ , содержащую только эти строки.

- 159 3. Для столбцов матрицы  $\tilde{Z}$  найдём лучшие координаты – координаты по базису из  
160 левых сингулярных векторов этой матрицы. Затем спроектируем столбцы этой  
161 матрицы на подпространство из  $n$  наиболее информативных таких векторов, где  
162 информативность зависит от соответствующих сингулярных чисел. Полученную  
163 матрицу обозначим  $Z^*$ , её размер  $n \times M$ . Само  $n$  определяется по доле необъ-  
164 яснённой дисперсии для  $\tilde{Z}$ . Заметим, что квадраты сингулярных чисел и левые  
165 сингулярные векторы совпадают с собственными числами и собственными векто-  
166 рами, соответственно, ковариационной матрицы  $\frac{1}{M-1} \tilde{Z}^T \cdot \tilde{Z}$  признаков, размера  
167  $M \times M$ .
- 168 4. Столбцы матрицы  $Z^*$  одновременно являются точками в  $\mathbb{R}^n$  (столбцов-точек  $M$   
169 штук) в  $n$ -ом вещественном пространстве  $\mathbb{R}^n$  и одновременно ещё они будут *вер-*  
170 *шинами* следующего графа  $G$ . Для каждой из этих точек  $j$  по евклидову рас-  
171 стоянию до других точек, найдём список из  $k$  ближайших соседей точки  $j$  (этот  
172  $j$ -список начинается с самой  $j$ ), и рассмотрим *ранги* точек из  $j$ -списка по возрас-  
173 танию расстояния от  $j$ .
- 174 5. В графе  $G$  проведём ребро между вершинами  $j$  и  $l$ , если их списки имеют об-  
175 щего  $k$ -соседа. Рёбрам припишем веса, что закончит определение нагруженного  
176 неориентированного графа  $G$ . Выберем некоторую начальную кластеризацию  $C_0$   
177 вершин в  $G$ .
- 178 6. Итеративно найдём *максимум функции модулярности*, аргументом которой яв-  
179 ляется текущая кластеризация  $C$  вершин в  $G$ ; в результате найдём финальную  
180 кластеризацию вершин в  $G$  (вершины – столбцы в матрице  $Z^*$ ).
- 181 7. Вернёмся к матрице  $Y$ , которая имеет  $n$  и  $M$  исходных строк и столбцов, но  
182 теперь ещё и кластеризацию её клеток (назовём её *матрицей кластеризации*).  
183 В каждом кластере найдём *дифференциально-экспрессированные строки* (DE-  
184 строки или, то же самое, *DE-признаки*) с помощью Mann-Whitney  $U$ -теста и  
185 процедуры Венжамини-Нохберг.
- 186 8. Граф  $G$  (с координатами в  $n$ -ом пространстве) можно приблизить другим гра-  
187 фом  $G_1$ , вершины которого находятся в  $K$ -ом пространстве, где  $K$  много меньше,  
188 чем  $n$ , например,  $K = 2$ . Это делается с помощью нелинейной функции, называ-  
189 емой *УМАР*; *граф  $G_1$*  в  $K$ -ом пространстве находится как минимум *дивергенции*  
190 *Кульбака-Лейблера*. Решаемая здесь задача понижения размерности данных (то  
191 есть размерности  $n$  до размерности  $K$ ) сама по себе – одна из центральных при  
192 работе с большими данными. Вообще говоря, такое понижение размерности мож-  
193 но выполнить от любой размерности, например, от  $N$  к  $K$ .

#### 194 § 4. лог-Преобразование матрицы $X$

195 В прикладных работах часто каждая строка  $i$  исходной матрицы  $X$  рассматри-  
196 вается как выборка своей случайной величины  $X_i$  с математическим ожиданием  $\mu_i$

197 и дисперсией  $v_i$ , причём  $v_i = v(u_i)$ , а  $v(u)$  – положительный квадратичный *полином*,  
 198 не зависящий от  $i$ . Тогда стремятся найти такую функцию  $g(x)$ , где  $x$  – любой эле-  
 199 мент матрицы  $X$ , что матрица  $Y = \{y_{ij}\}$ , где  $y_{ij} = g(x_{ij})$ , для каждой строки  $i$  имеет  
 200 эмпирическую дисперсию  $\sigma_i^2$  (сколь возможно) слабее зависящую от её среднего  $y_i$ ,  
 201 а в пределе не зависящую от него. Эта постановка, как и её последующее решение,  
 202 являются эвристическими, хотя и широко распространёнными, как это обсуждалось  
 203 во Введении.

204 Итак,  $i$ -ым строкам в матрицах  $X$  и  $Y$  соответствуют случайные величины  $\mathbf{X}_i$   
 205 и  $\mathbf{Y}_i = g(\mathbf{X}_i)$ , у которых далее индекс  $i$  опустим для краткости записи. Разложим  
 206  $\mathbf{Y}$  по степеням  $\mathbf{X} - u$ :  $\mathbf{Y} = g(u) + g'(u) \cdot (\mathbf{X} - u) + g''(u) \cdot (\mathbf{X} - u)^2 + \dots$ , и в  
 207 разложении оставим только аффинную часть; получим  $\mathbf{E}(\mathbf{Y}) \approx g(u)$  и  $\text{Var}(\mathbf{Y}) \approx$   
 208  $g'(u)^2 \cdot v$ . Таким образом придём к дифференциальному уравнению  $g'(u)^2 \cdot v(u) = 1$ .  
 209 Полученную функцию  $g$  применим к каждому элементу из каждой строки в  $X$  и  
 210 получим искомую матрицу  $Y$ . Для монотонной функции  $g$  разные средние строк в  
 211 преобразуются в разные средние строк в  $Y$ , которые в  $Y$  соответствуют примерно  
 212 одинаковым дисперсиям строк; в этом смысле новая дисперсия слабо зависит от  
 213 нового среднего.

214 В биохимических и других приложениях, к которым относится и эксперимен-  
 215 тальное определение экспрессий генов в клетках, часто дисперсия  $v$  является таким  
 216 положительным квадратичным полиномом от матожидания  $u$ . Например, измеряе-  
 217 мая величина может иметь вид  $\alpha + \mu e^\eta + \varepsilon$ , где  $\mu$  – её точное (истинное, но неизвест-  
 218 ное нам) значение,  $\alpha$  – сдвиг данных, а  $\varepsilon$  и  $\eta$  – аддитивная и мультипликативная  
 219 погрешности измерения, которые являются независимыми случайными величина-  
 220 ми и, например, нормальными с нулевым матожиданием. Тогда дисперсии  $v$  такого  
 221 измерения зависит от матожидания  $u$  именно как упомянутый полином

$$v = \frac{\text{Var}(e^\eta)}{\mathbf{E}(e^\eta)^2} \cdot (u - \alpha)^2 + \text{Var}(\varepsilon) = (c_1 u + c_2)^2 + c_3.$$

222 Итак, для этого (и любого другого упомянутого вида) полинома указанное урав-  
 223 нение приводит к гиперболо-арксинусному преобразованию в роли  $g$ . Явный вид  $g$   
 224 для элемента матрицы и так записанного полинома таков:

$$x_{ij} \rightarrow y_{ij} = \begin{cases} \frac{1}{c_1} \cdot \text{arsinh}\left(\frac{c_2}{\sqrt{c_3}} + \frac{c_1}{\sqrt{c_3}} \cdot x_{ij}\right), c_3 > 0 \\ \frac{1}{c_1} \cdot \ln(c_2 + c_1 \cdot x_{ij}), c_3 = 0, \end{cases}$$

225 где гиперболический арсинус определяется как

$$\operatorname{arsinh}(z) = \ln(z + \sqrt{z^2 + 1}) = \int_0^z \frac{1}{\sqrt{t^2 + 1}} dt,$$

226 а константы  $c_1, c_2, c_3$  подбираются по исходным данным. Точнее, эти константы, по  
227 возможности, подбираются из условия минимума по ним корреляции между дис-  
228 персиями и средними строк в матрице  $Y$ .

229 Заметим, что при  $c_3 = 0$  получается функция, близкая ко всеми принятой функ-  
230 ции  $\log_2(1+x)$  (или, как часто пишут в прикладных публикациях,  $\log_2(x)$ ). Конечно,  
231 нет основания полагать, что для исходных данных всегда выполняется  $c_3 = 0$ .

## 232 § 5. Выделение строк с большой дисперсией

233 Для матрицы  $Y$  и каждой её строки  $i$  вычислим среднее  $y_i$  и дисперсию  $\sigma_i^2$ .  
234 Построим полиномиальную низкой степени регрессию  $f(\cdot)$  для множества точек  
235  $\{(y_i, \sigma_i^2)\}$ , где  $y_i$  в середине интервала  $\mathcal{O}$ , а интервалы  $\{\mathcal{O}\}$  выбраны так, что число  
236 точек в каждом из них примерно одинаково, рис. 2. Положим  $f(i) = f(y_i)$  – значение  
237 регрессии  $f(\cdot)$  в  $y_i$ . Регрессия относится к интервалу  $\mathcal{O}$ , и в этом смысле называется  
238 локальной, а число  $f(i)$  можно назвать *регрессионной дисперсией* строки  $i$  (иногда  
239 её называют средне-дисперсионным отношением). Она характеризует «дисперсию»  
240 строки  $i$  без учёта её индивидуальности. Иными словами, нас будут интересовать  
241 строки, для которых дисперсия  $\sigma_i^2$  не подчиняется общему правилу, заданному ре-  
242 грессией  $f(\cdot)$ , то есть  $\sigma_i^2$  и  $f(i)$  значительно отличаются,  $f(i)$  существенно меньше  
243  $\sigma_i^2$ .

244 Образуем матрицу  $Z = \frac{y_{ij} - y_i}{f(i)}$ . Для её строк  $i$  среднее равно  $z_i = 0$ , а дисперсия  
245 равна  $\delta_i = \frac{\sigma_i^2}{f(i)}$ . В  $Z$  слишком большие дисперсии рассматриваются как артефакт  
246 исходных данных; поэтому строки, у которых  $\delta_i^2 > \sqrt{M}$ , удаляются из  $Z$ . В так  
247 полученной матрице  $Z$  оставим только строки с наибольшей регрессионной диспер-  
248 сией  $\delta_i$  в заданном количестве. Обозначим  $\tilde{Z}$  часть матрицы  $Z$ , содержащую только  
249 выбранные строки (например, 2000); тогда её размер  $2000 \times M$ .

250 Итак, строки матрицы  $\tilde{Z}$  слабо зависят от среднего и не подчиняются регрес-  
251 сионному правилу, имея большую дисперсию. Это позволяет отчётливо различать  
252 столбцы по их координатам.

253 В результате каждая клетка-столбец получает 2000 координат, представляющих  
254 точку в 2000-ом пространстве вещественных чисел  $\mathbb{R}^{2000}$ . Получим множество  $\mathcal{M}$  из  
255  $M$  точек в этом пространстве.

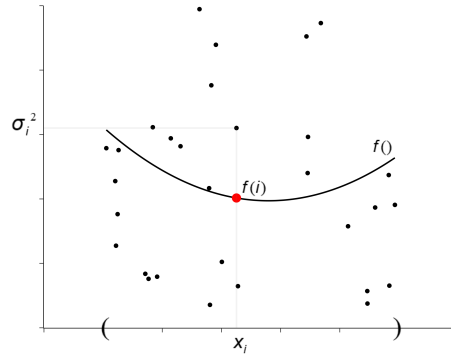


Рис. 2. Показана квадратичная регрессия, локальная на интервале  $O$ , с серединой  $y_i$ , которая определяет регрессионную дисперсию  $f(i)$  строки  $i$  матрицы  $Y$ . В этом интервале точки плоскости определяются как среднее  $y_i$  (абсцисса) строки  $l$  и её дисперсия  $\sigma_i^2$  (ордината).

256

## § 6. Выбор лучших координат для точек из множества $\mathcal{M}$

257 Выберем ещё лучшие, *новые координаты* клетки  $j$ , которые будут линейными  
 258 комбинациями её старых координат, столбцов матрицы  $\tilde{Z}$ . Это выполняется в два  
 259 шага.

260 1-ый шаг. Для матрицы  $\tilde{Z}$  вычислим сингулярные числа и левые сингулярные  
 261 векторы  $u_1, u_2, \dots$  единичной длины, и перейдём к *ортонормированному* базису  
 262 в  $\mathbb{R}^{2000}$ , состоящему из этих векторов, рис. 3. Удобно считать, что сингулярные  
 263 числа расположены по убыванию  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{2000}$ , как и соответствующие им  
 264 строки матрицы и новые координатные оси. Новые координаты клетки  $j$  (её старые  
 265 координаты – элементы столбца  $\tilde{Z}_j$ ) равны  $(u_k, \tilde{Z}_j)$ , где указана  $k$ -я координата,  
 266  $1 \leq k \leq 2000$ . Получен результат 1-го шага – матрица размера  $2000 \times M$ , которую  
 267 обозначим  $\hat{Z}$ .

268 2-ой шаг. Оставим в  $\hat{Z}$  только  $n$  новых и наиболее информативных координат,  
 269 которые назовём *сокращёнными* (или наиболее информативными); они образуют  
 270 матрицу  $Z^*$  размера  $n \times M$ , которая получена из матрицы  $\hat{Z}$  1-го шага размера  
 271  $2000 \times M$ .

272 Выбор числа  $n$  основан на следующем соображении информативности новых ко-  
 273 ординат. Для краткости будем далее опускать знаки  $^t$ . Замена старых координат  $Z_j$   
 274 клетки  $j$  на её проекцию на первые  $n$  координат  $Z_j \rightarrow Z_j^* = \sum_{k=1}^n (u_k, Z_j) \cdot u_k$  вносит  
 275 *средний квадрат ошибки* в расчёте на одну клетку, равный  $\frac{1}{M} \sum_{j=1}^M \|Z_j - \tilde{Z}_j^*\|^2 =$   
 276  $\sum_{k=n+1}^N \lambda_k$ , где  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{2000}$ . Эта сумма называется *остаточной дисперси-*  
 277 *ей*.

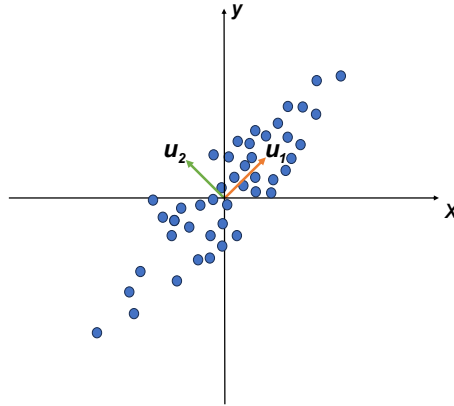


Рис. 3. Для клеток  $j$  показан переход от их старых координат – столбцов матрицы  $\tilde{Z}$  к их новым координатам – столбцам матрицы  $\tilde{Z}'$ .

278 Величина  $\frac{1}{M} \sum_{j=1}^M \|\sum_{k=1}^n (u_k, Z_j)^2 \cdot u_k\|^2 = \frac{1}{M} \sum_{j=1}^M \sum_{k=1}^n (u_k, Z_j)^2 = \sum_{k=1}^n \lambda_k$  на-  
 279 зывается *объяснённой дисперсией*. *Относительная ошибка*  $\delta_n$  определяется как от-  
 280 ношение остаточной дисперсии к выборочной дисперсии; эта доля называется *необъ-*  
 281 *яснённой дисперсией*:  $\delta_n = \frac{\lambda_{n+1} + \dots + \lambda_{2000}}{\lambda_1 + \lambda_2 + \dots + \lambda_{2000}}$ .

282 Выберем  $n$  из условия  $\delta_n \leq \varepsilon_0$ , где  $\varepsilon_0$  – заданный порог.

### 283 § 7. Кластеризация клеток по их новым сокращённым координатам (столбцам в 284 матрице $Z^*$ )

285 образуем граф  $G$ : его вершины – клетки-точки  $j$  с их  $n$  сокращёнными и новыми  
 286 (наиболее информативными) координатами – столбцами  $Z_j^*$  матрицы  $Z^*$ . Этот граф  
 287 называется  $kNN$ -графом, а координаты клетки-точки-вершины  $j$  называются её  
 288 *РСА-координатами*.

289 Для каждой вершины  $j$  по евклидову расстоянию между точками в  $n$ -мерном  
 290 пространстве составим список  $k$  ближайших к ней вершин ( $k$ -соседей, сама  $j$  первая  
 291 в этом  $j$ -списке, где  $k$  параметр).  $j$ -*Окрестностью* в  $G$  называется  $j$ -список вместе  
 292 со всеми рёбрами, соединяющими любые две вершины из него. Заметим, что таким  
 293 образом мы переходим к новым окрестностям для всех вершин  $j$ : к картам, которые  
 294 представлены одномерными симплексами на многообразии всех клеток.

295 Список нумеруем подряд натуральными числами от 1 (которые называют *ран-*  
 296 *гами*) по возрастанию евклидова расстояния. Обозначим  $\text{rank}(v, j)$  – ранг вершины  
 297  $v$  в  $j$ -списке, который начинается с вершины  $j$ . Вершины, которые находятся на  
 298 одинаковом расстоянии от  $j$ , нумеруются дробными числами: например, вершины,  
 299 находящиеся на расстоянии 1, 2, 7, 7, 8, ... имеют ранги 1, 2, 3,5, 3,5, 4, ...

300 Вершины  $j$  и  $l$  соединим ребром, если  $j \neq l$ , а  $l$  является некоторым соседом в  
 301  $j$ -списке и  $j$  является некоторым соседом в  $l$ -списке.

302 Вес  $A_{jl}$  ребра  $e = (j, l)$  положим равным максимуму по всем общим  $k$ -соседям  $v$   
 303 для  $j$  и  $l$ :  $A_{jl} = \max\{k - 0, 5(\text{rank}(v, j) + \text{rank}(v, l))|v\}$ .

304 Найдём некоторую начальную кластеризацию  $\mathcal{C}_0$  вершин графа  $G$  (разбиение его  
 305 вершин). Выбор  $\mathcal{C}_0$  важен и существенно влияет на окончательный результат кла-  
 306 стеризации, известно несколько способов такого выбора. Полученные в результате  
 307 кластеры могут допускать рёбра между вершинами из разных кластеров; о таких  
 308 рёбрах говорят, что они “*между кластерами*”; наоборот, о рёбрах, у которых оба  
 309 края принадлежат одному кластеру, говорят, что они “*в кластере*”.

310 Начиная с начальной кластеризации  $\mathcal{C}_0$ , итеративно найдём локальный макси-  
 311 мум функции  $H(\mathcal{C})$ , которая называется *модулярностью* кластеризации  $\mathcal{C} = \mathcal{C}(G)$   
 312 графа  $G$  и характеризует качество разбиения вершин в  $G$ . Функция модулярности  
 313 определятся как

$$H(\mathcal{C}) = \frac{1}{2m} \sum_{j,l} \left( A_{jl}(e) - \gamma \cdot \frac{k_j \cdot k_l}{2m} \right) \rightarrow \max.$$

314 Здесь  $A_{jl}$  – вес ребра  $e$ , которое пробегает *все рёбра* во всех кластерах текущего  
 315 разбиения  $\mathcal{C} = \mathcal{C}(G)$ , а  $m$  – сумма весов всех рёбер в  $G$ , и  $k_j$  сумма весов всех рёбер  
 316 в  $G$ , инцидентных вершине  $j$ ;  $\gamma$  – параметр, называемый *гранулярностью*: минимум  
 317 суммарного веса рёбер в кластере (относительно суммарного веса всех рёбер) по  
 318 всем кластерам, делённый на максимум аналогичной доли суммарного веса рёбер  
 319 между всеми парами разных кластеров.

320 О параметре  $\gamma$  заметим следующее. При его увеличении число отрицательных  
 321 рёбер  $e$  может увеличиться, и становится выгодным разбить кластер на несколь-  
 322 ко кластеров, чтобы отрицательные рёбра сосредоточились между кластерами и,  
 323 тем самым, были исключены из  $H$ . Тогда число кластеров увеличится, а сами они  
 324 уменьшатся. Наоборот, при уменьшении  $\gamma$  число положительных рёбер может уве-  
 325 личиться, поэтому становится выгодным объединить несколько кластеров в один,  
 326 чтобы включить в  $H$  межкластерные рёбра, ставшие положительными. Тогда число  
 327 кластеров уменьшится, и они увеличатся.

328 Нетривиальный выбор вида функции  $H(\mathcal{C}(G))$  принципиально влияет на резуль-  
 329 тат кластеризации; отметим следующие соображения по поводу выбора и обоснова-  
 330 ния функции модулярности. Обозначим  $r \in \mathcal{C}$  – кластер данной кластеризации  $\mathcal{C}$ .  
 331 Веса  $A_{jl}$  рёбер в  $G$  можно интерпретировать как мультиграф без весов: вершины  $j$  и  
 332  $l$  соединяются рёбрами в количестве  $A_{jl}$ . Обозначим той же буквой  $A_{jl}$  симметрич-  
 333 ную матрицу инцидентности полученного *мультиграфа* без петель, который так же  
 334 обозначим  $G$  (на диагонали находятся нули, остальные числа равны весу  $\frac{A_{jl}}{2}$ ; как  
 335 обычно, у петли вес удваивается). Тогда  $A_{jl} = A_{lj}$  и суммирование по всем рёбрам  
 336  $j, l \in r$  приводит к удвоенной сумме числа рёбер в  $r$ . В формуле для  $H$  перейдём

337 от суммирования по рёбрам  $e = (j, l) \in G$  к суммированию по кластерам  $r \in G$  и  
 338 получим

$$H(\mathcal{C}(G)) = \frac{1}{m} \sum_r \left[ \left( \frac{1}{2} \cdot \sum_{(j,l) \in r} A(j,l) \right) = \frac{k(r)^2}{4m} \right],$$

339 где  $m$  – число рёбер в  $G$  и  $\sum_{(j,l) \in r} A(j,l)$  – удвоенное число рёбер внутри кластера  $r$ .  
 340 Здесь  $k(j)$  – степень инцидентности вершины  $j$  в графе  $G$ , вычисляемая по матрице  
 341 инцидентности  $A_{jl}$ , а  $k(r) = \sum_{j \in r} k(j)$  – степень инцидентности кластера  $r$ . Легко  
 342 видеть, что  $\sum_{j \in G} k(j) = 2m$ . Итак, уменьшаемое – число рёбер в  $r \in \mathcal{C}(G)$ , и мы  
 343 хотим интерпретировать вычитаемое.

344 Это можно сделать как среднее число рёбер в семействе случайных графов с те-  
 345 ми же вершинами (и параметрами), что у  $G$  и той же кластеризацией этих вершин.  
 346 Однако приведём более простое комбинаторное пояснение. На множестве вершин  
 347 мультиграфа  $G$  с его кластеризацией рассмотрим “идеальный” (не мульти, обычный)  
 348 граф  $G'$  с той же кластеризацией и теми же степенями инцидентности  $k(j)$ , что у  
 349 вершин мультиграфа  $G$ . Напомним, что в  $G$  и  $G'$  степень инцидентности определя-  
 350 ется как сумма весов всех инцидентных ей рёбер. А именно,  $G'$  полный с петлями  
 351 и в нём вес ребра  $(j, l)$  полагаем равным  $\frac{k(j)k(l)}{2m}$ , если  $j \neq l$ , и  $\frac{k(j)k(l)}{4m}$ , если  $j = l$ . В  
 352  $G'$  для любой вершины  $j$  степень инцидентности равна  $\sum_{l \neq j} \frac{k(j)k(l)}{2m} + \frac{2k(j)^2}{4m} = k(j)$ .  
 353 И в  $G'$  сумма весов всех рёбер в  $r$  равна  $\sum_{j \neq l, \{j,l\} \in r} \frac{k(j)k(l)}{2m} + \sum_j \frac{k(j)^2}{4m} = \frac{k(r)^2}{2m}$ . По-  
 354 следнее проверяется прямым раскрытием скобок в  $k(r)$ . Повторим, что графы  $G$  и  
 355  $G'$  имеют в качестве общих параметров саму кластеризацию  $\mathcal{C}$ , количества вершин  
 356 и рёбер, и степени инцидентности  $\{k(j)\}$  для всех вершин  $j$ . Величины  $k(j)$  можно  
 357 рассматривать как характеристики неоднородности графа, которые одинаковы у  $G$   
 358 и  $G'$ : у них в одних и тех же областях плотность рёбер повышена или понижена по  
 359 сравнению со средней плотностью.

360 **7.1. Итеративный алгоритм кластеризации Leiden.** Изложенный выше алгоритм  
 361 уточняется следующим образом. Для исходного графа  $G$  выбирается начальная кла-  
 362 стеризация  $\mathcal{C}_0$  и максимизируется функция модулярности  $H(\mathcal{C}(G))$ . По полученному  
 363 графу строим новый граф  $G_1$ , в котором все вершины одного кластера в  $G$  пред-  
 364 ставлены одной новой вершиной в  $G_1$ . Все рёбра внутри кластера в  $G$  заменяются  
 365 петлёй в  $G_1$  у этой новой вершины, а все рёбра между двумя разными кластерами  
 366 в  $G$  заменяются одним новым ребром в  $G_1$  между соответствующими новыми вер-  
 367 шинами. Вес петли полагаем равным суммарному весу исходных соответствующих  
 368 внутрикластерных рёбер, а вес нового ребра – суммарному весу исходных соответ-  
 369 ствующих межкластерных рёбер. После этого максимизируем функцию  $H(\mathcal{C}(G_1))$ .  
 370 Таким образом процесс продолжается, пока максимум  $H(\mathcal{C}(G_k))$  на очередном гра-  
 371 фе  $G_k$  увеличивается, рис. 4. Затем последовательными объединениями вернёмся к

372 исходному графу  $G$ . А именно, если вершины в  $G_{k-1}$  входят в один и тот же кластер  
 373 в  $G_k$ , то соответствующие им кластеры объединяются в один кластер в  $G_{k-1}$ . И об-  
 374 ратным ходом придём к кластеризации исходного графа  $G$ . Получим кластеризацию  
 375 вершин в исходном  $G$ . При каждом переходе от  $k - 1$  к  $k$  качество соответствующей  
 376 кластеризации графа  $G_k$  не уменьшается.

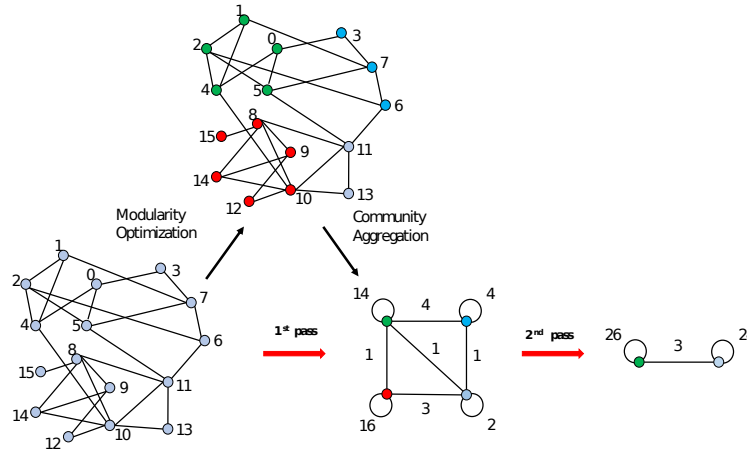


Рис. 4. Веса всех рёбер единичные. Показаны шаги итеративного алгоритма Leiden. Пример взят из [6].

377 **7.2. К обоснованию итеративного алгоритма кластеризации Louvain-Leiden.** В  
 378 качестве начальной кластеризации можно выбрать *синглетонную кластеризацию*,  
 379 которая состоит из всех вершин графа  $G_k$ , рассматриваемых как одноэлементные  
 380 множества (“синглетоны”). Но обычно в качестве начальной кластеризации берут бо-  
 381 лее крупную кластеризацию, называемую *окрестностной*; обозначим её  $C_0$ . Граф  
 382  $G_k$  далее обозначим  $G$ , например, это граф  $G$ , определённый в начале этого разде-  
 383 ла, с которого и начинает работу итеративный алгоритм. Для построения этой  $C_0$   
 384 используются описанные в начале этого раздела  $j$ -окрестности всех вершин в  $G$ , и  
 385 веса рёбер графа  $G$ . Опишем кластер  $c \in C_0$ , каждый кластер образуется по неко-  
 386 торой вершине  $j \in G$  и в этом смысле обозначается  $c_j$  (кластер с “центром” в  $j$ ).  
 387 Кластеры  $c_j$  образуются итеративно, каждый начиная с  $j$ -окрестности. В подграфе  
 388  $c_j$  обозначим  $d(v)$  сумму весов рёбер в  $c_j$ , инцидентных  $v$ ; это – степень вершины  
 389  $v$ , отнесённая к  $c_j$ . Проредим каждое  $c_j$ , удаляя из него вершины  $v \in c_j$  вместе со  
 390 всеми инцидентными им рёбрами, если  $\frac{d(v)}{|c_j|} < r$ , где  $r \in (0, 1]$  – порог и  $|c_j|$  – сумма  
 391 весов всех рёбер в  $c_j$ . Удаление таких вершин из  $c_j$  зависит от порядка просмотра  
 392 вершин в  $c_j$ ; продолжаем удаление, пока такая вершина находится. Удаления вы-  
 393 полняются независимо для каждого  $j$ . Если  $|c_j| = 0$ , то вершину  $j$  сохраним; если  $v$   
 394 удаляется из всех  $j$ -окрестностей в  $G$ , то оставим  $v$  в качестве синглтона.

395 После этого в полученном наборе подграфов  $c_j$  устраним их пересечения по вер-  
 396 шинам: пусть  $v$  принадлежит нескольким  $c_j$ . Для каждого  $c_j$  вычислим *средний вес*  
 397 рёбер, соединяющих  $v$  со всеми вершинами в  $c_j$ . Вершину  $v$  сохраним в той  $c_j$ , для  
 398 которой средний вес максимален, а из остальных  $c_j$  удалим её. Итоговые  $c_j$  образуют  
 399 кластеры начальной кластеризации  $C_0$  в графе  $G$ .

400 Наметим доказательство того, что при каждом переходе от  $k$  к  $k + 1$  качество  
 401 соответствующей кластеризации графа  $G_k$  не уменьшается.

402 Для краткости докажем здесь более высокое качество кластеризации  $C_0$  по срав-  
 403 нению с синглетонной в упрощённом случае, когда учитываются только  $k$  рёбер при  
 404 каждой вершине, а вершины из окрестностей не удаляются и ближайший сосед двух  
 405 вершин из  $j$ -окрестности – сама точка  $j$ , и ранги – целые попарно различные числа.  
 406 Пусть  $G$  – полный граф с петлями: так будет, если отсутствующие рёбра положить  
 407 с весом 0. Напомним, что значение  $H$  на любой кластеризации в  $G$  (с точностью до  
 408 мультипликативной константы) равно сумме разностей весов рёбер, лежащих внут-  
 409 ри кластера в  $G$  и в идеальном графе  $G'$ . Петли лежат внутри кластера при любой  
 410 кластеризации и, следовательно, приносят одинаковый вклад в  $H$  для синглетонной  
 411 и окрестностной кластеризаций. Поэтому нам требуется оценить суммы весов в  $G$  и  
 412  $G'$  рёбер, которые лежат внутри кластера в окрестностной кластеризации, но не в  
 413 синглетонной. Это все рёбра, не являющиеся петлями, в кластерах  $c_j \in C_0$ .

414 Поэтому в  $G$  сумма весов рёбер, которые не петли, из  $c_j \in C_0$  равна  $\frac{k^2(k-1)}{2} -$   
 415  $\frac{1}{2} \sum_{i \neq l} (i+l) = \frac{k^2(k-1)}{2} - \frac{(k-1)k(k+1)}{4} = \frac{k(k-1)^2}{4}$ , здесь  $i$  и  $l$  – ранги вершин в  $c_j$ , которые  
 416 принимают значения от 1 до  $k$ . Оценим сумму  $k_i$  весов рёбер, инцидентных вершине  
 417 ранга  $i$  из  $c_j \in C_0$ :

$$k_i = k(k-1) - \frac{1}{2} \sum_{l \neq i} (i+l) = \frac{3k^2 - 5k - 2ik + 4k}{4} \leq \frac{3k(k-1)}{4}.$$

418 При  $\gamma = 1$  в идеальном графе  $G'$  сумма весов рёбер из  $c_j$  равна  $\frac{1}{2m} \sum_{i \neq l \in c_j} k_i k_l \leq$   
 419  $\frac{1}{2m} \left(\frac{3}{4}k(k-1)\right)^2 \frac{k(k-1)}{2} = \frac{9}{64m} k^3 (k-1)^3$ . Итак, качество окрестностной кластеризации  
 420 больше качества синглетонной, если  $\frac{k(k-1)^2}{4} > \frac{9}{64m} k^3 (k-1)^3$ , то есть  $m > \frac{9}{16} k^2 (k-1)$ .  
 421 Обычно обрабатываются графы, у которых  $m > 15000$ , а  $k$  не больше 30, так что  
 422 это неравенство выполняется.

423 **Теорема 1.** Пусть граф  $G$  полный с петлями и  $C_k$  – итоговая кластеризация  
 424  $k$ -го уровня, а  $C_{k+1}$  – кластеризация  $(k+1)$ -го уровня, в которой все кластеры –  
 425 синглетоны. Тогда  $H(C_k) \leq H(C_{k+1})$ .

426 **Доказательство.** Назовём кластеры кластеризации в  $C_{k+1}$   $(k+1)$ -го уровня круп-  
 427 ными, кластеры кластеризации в  $C_k$   $k$ -го уровня – средними, кластеры кластериза-  
 428 ции в  $C_{k-1}$   $(k-1)$ -го уровня – мелкими. Аналогично назовём и сами кластеризации  
 429 в этих графах.

430 Из определения веса ребра между кластерами следует, что суммарный вес  $m$  всех  
 431 рёбер между кластерами не меняется при переходе к кластеризации следующего  
 432 уровня (независимо от того, являются ли кластеры следующего уровня синглтонами).  
 433 И также следует: если крупные кластеры – синглтоны, то суммарный вес всех  
 434 рёбер, лежащих внутри такого синглтона (а это вес петли при соответствующем  
 435 среднем кластере) равен суммарному весу рёбер между мелкими кластерами, ле-  
 436 жащими внутри среднего. Суммируя этот вес по синглтонам, получим, что сумма  
 437  $\sum A_{jl}$  из формулы для  $H$  для крупной кластеризации равна этой сумме для средней.

438 Осталось рассмотреть сумму произведений  $k_j \cdot k_l$  для пар  $(j, l)$  вершин, лежащих  
 439 внутри кластера данной кластеризации. Любой крупный кластер-синглтон  $a$  состо-  
 440 ит из одного среднего кластера, который обозначим  $a'$ . Для крупной кластеризации  
 441 слагаемое в упомянутой сумме – квадрат веса петли при  $a'$ , назовём её первой сум-  
 442 мой. Для средней кластеризации это – сумма по всем парам  $(j, l)$  мелких кластеров  
 443 (в  $a'$ ) произведений  $k_j \cdot k_l$ ; назовём её второй суммой.

444 Первая сумма не больше второй. Действительно, в первой сумме при раскрытии  
 445 скобок возникают произведения  $w(e_1) \cdot w(e_2)$  весов рёбер  $e_1$  и  $e_2$  между мелкими  
 446 кластерами, лежащими внутри  $a'$ , причём каждое произведение встретится два ра-  
 447 за, если  $e_1 \neq e_2$  и один раз, если иначе. Во второй сумме при раскрытии скобок  
 448 также возникнут произведения  $w(e_1) \cdot w(e_2)$  весов рёбер  $e_1$  и  $e_2$ , уже не обязательно  
 449 лежащих внутри  $a'$ . Однако каждое такое произведение для рёбер, лежащих внутри  
 450  $a'$ , встретится не меньше двух раз, если  $e_1 \neq e_2$  и не меньше одного раза, если иначе.  
 451 Поэтому часть суммы  $\sum k_j k_l$  из формулы для  $H$ , относящаяся к одному кластеру-  
 452 синглтону  $a$ , для крупной кластеризации не превосходит аналогичную часть суммы  
 453 (относящуюся к  $a'$ ) для средней кластеризации. Суммируя по всем синглтонам и  
 454 учитывая, что сумма  $\sum k_j k_l$  входит в  $H$  со знаком минус, получим утверждение. ▲

455 Однако нет гарантии, что функция  $H$  не уменьшится на обратном ходе проце-  
 456 дуры. Поэтому, мы начинали этот ход с итоговой кластеризации каждого, не обяза-  
 457 тельно заключительного, уровня и из всех таким образом полученных кластериза-  
 458 ций в  $G$  выбирали кластеризацию с максимальным значением  $H$ .

## 459 § 8. Дифференциально-экспрессируемые признаки.

460 Уже получена кластеризация (разбиение)  $\{G_s\}$  клеток-столбцов исходной мат-  
 461 рицы  $Y$  размера  $N \times M$ , которая прошла log-нормирование, но не шкалирование-  
 462 центрирование; матрица состоит из старых координат, до перехода к PCA-координатам  
 463 и сокращения менее информативных из них; здесь  $s$  пробегает все кластеры. Для  
 464 каждого кластера  $G$  клеток и каждой строки  $i$  определим дифференциально экс-  
 465 прессируемые признаки-features. Такие признаки-строки сокращённо называются  
 466 *DE-признаками* (или *DE-строками*). Говоря интуитивно, признак-строка  $i$  назы-  
 467 вается *дифференциально-экспрессируемой*, если экспрессии в клетках  $j \in G$  стати-

468 стически значимо отличаются от экспрессий во всех остальных клетках  $l \notin G$  той  
 469 же строки  $i$ . Это понимание *DE-строк* для данного кластера  $G$  ниже уточняется:  
 470 сначала с помощью Mann-Whitney  $U$ -теста [7–9], затем с помощью процедуры  
 471 Benjamini-Hochberg [10], а также ещё ряда условий.

472 Итак, фиксирована строка  $i$  (её имя и соответствующий индекс можно не упоми-  
 473 нать) и фиксирован кластер  $G$ , который индуцирует разбиение строки на два мно-  
 474 жества, которые будем обозначать теми же буквами  $G$  и  $\neg G$ , по-прежнему называя  
 475 их кластерами. Расположим все экспрессии строки  $i$  в порядке их возрастания; и  
 476 присвоим числам в полученной последовательности экспрессий *ранги* – натураль-  
 477 ные числа также по их возрастанию, начиная с 1. Одинаковым числам присвоим  
 478 средний по ним ранг. Пусть  $R_1$  – сумма рангов экспрессий в данном кластере  $G$ , и  
 479  $R_2$  – сумма рангов экспрессий вне этого кластера, т.е. в  $\neg G$ ; пусть  $m_1$  и  $m_2$  – мощно-  
 480 сти множеств  $G$  и  $\bar{G}$ , соответственно;  $m_1 + m_2 = M$ . Поскольку  $R_1 + R_2 = M \cdot \frac{1+M}{2}$ ,  
 481 можно оперировать только с  $R_1$  или только с  $R_2$ .

482 Предполагается, что  $G$  и  $\bar{G}$  – *независимые* выборки из двух непрерывных рас-  
 483 пределений  $\mathfrak{F}$  и  $\mathfrak{L}$ , которые *совпадают* или *не совпадают*. Для экспериментально-  
 484 математических выборок  $G$  и  $\neg G$  вопрос об их независимости трудно разрешим,  
 485 так что он остаётся на усмотрение вычислителя; в нашей ситуации это является  
 486 предположением. Если  $\mathfrak{F} = \mathfrak{L}$ , то для экспрессий  $y_j \in G$  и  $y_l \in \neg G$  выполняется  
 487  $\mathbf{P}(y_j < y_l) = 1/2$ . Это свойство называется отсутствием “доминирования”  $G$  и  
 488  $\neg G$  друг над другом. Обратное неверно: например, если  $\mathfrak{F}$  и  $\mathfrak{L}$  – два нормальных  
 489 распределения с одинаковыми математическими ожиданиями, но разными диспер-  
 490 сиями, то доминирование отсутствует:  $\mathbf{P}(y_j - y_l < 0) = \frac{1}{2}$ , так как разность этих  
 491 распределений имеет нулевое математическое ожидание.

492 Итак, мы хотим проверить, выполняется ли *0-ая гипотеза*, которая состоит в  
 493 “отсутствии доминирования  $G$  и  $\neg G$  друг над другом”. В случае отрицания (говорят:  
 494 отвержения) 0-ой гипотезы, строка  $i$  называется *DE-признаком*. Если 0-ая гипотеза  
 495 отвергается, то на том же уровне достоверности выполняется  $\mathfrak{F} \neq \mathfrak{L}$ .

496 **8.1. Критерий Манна-Уитни-Вилкоксона (Wilcoxon rank-sum test, Mann-Whitney**  
 497 **U test и определение  $p$ -value; сокращённо МУВ).** По строке  $i$  матрицы  $Y$  образу-  
 498 ем случайную величину  $U = R_2 - \frac{m_2(m_2+1)}{2} = \sum_{j=1}^{m_1} \sum_{l=1}^{m_2} I(x_j < x_l)$ , где  $x_j$  и  $x_l$  –  
 499 экспрессии в кластере  $G$  и вне него, то есть в  $\neg G$ , соответственно; а  $I(\cdot)$  – харак-  
 500 теристическая функция условия, которое приведено в скобках. Если 0-ая гипотеза  
 501 *выполняется*, то её математическое ожидание и дисперсия равны  $\mathbf{E}(U) = \frac{m_1 \cdot m_2}{2}$  и  
 502  $\sigma(U)^2 = m_1 \cdot m_2 \frac{M+1}{12}$ .

503 В разделе 6 выполнено центрирование и нормирование матрицы  $Y$ . Аналогично  
 504 этот общий приём применяется и к  $U$ , чтобы получить статистику (случайную вели-  
 505 чину)  $Z = \frac{U - m_1 m_2 / 2}{\sqrt{m_1 m_2 (M+1) / 12}}$ . Если в последовательности рангов наблюдается асиммет-

506 рия (относительно линейного порядка чисел) экспрессий в пользу  $G$  или, наоборот, в  
 507 пользу  $\neg G$ , то  $Z$  отклоняется вправо или влево от нуля, больше или меньше. Если  
 508 это отклонение  $Z$  превосходит заданный порог  $Z_0$ , то строку-признак  $i$  считаем  $DE$ ;  
 509 уточним этот подход.

510 Повторим, пусть  $G$  и  $\neg G$  – выборки двух некоторых случайных величин. Повто-  
 511 рим, 0-я гипотеза состоит в том, что  $\mathbf{P}(y_j < y_l) = \frac{1}{2}$  для  $y_j \in G$  и  $y_l \in \bar{G}$  (говорят:  
 512 « $y_j$  не доминирует  $y_l$ , а  $y_l$  не доминирует  $y_j$ »). Эта гипотеза выражает “отсутствие  
 513 порядковой асимметрии” между  $G$  и  $\bar{G}$ . Заметим, что распределение для  $Z$  зада-  
 514 ётся мерой на  $\mathbb{R}$ , которая определяется как  $Z^{-1}$  от меры на  $M$ -ом пространстве,  
 515 индуцированной распределениями  $\mathfrak{F}$  и  $\mathfrak{L}$ .

516 При выполнении 0-й гипотезы и возрастающих  $m_1, m_2 \rightarrow \infty$ , распределение ста-  
 517 тистики  $Z$  стремится к нормальному 0-1 распределению, обозначаемому  $N(0, 1)$ . В  
 518 силу этого приближённо случайная величина  $Z$  имеет распределение  $N(0, 1)$ , то есть  
 519 вместо обычно неизвестных распределений  $\mathfrak{F}$  и  $\mathfrak{L}$  вероятность, далее везде обозна-  
 520 чаемую  $\mathbf{P}$ , будем вычислять по нормальному распределению  $N(0, 1)$ .

521 Хотя это дальше не используется, заметим: если экспрессия в  $G$  доминирует  
 522 экспрессию в  $\neg G$ , то есть  $\mathbf{P}(y_j < y_l) > 1/2$ , то  $Z$  стремится к  $+\infty$ , а если экспрессия  
 523 в  $G$  в том же смысле доминирует экспрессию в  $\neg G$ , то  $Z$  стремится к  $-\infty$  при также  
 524 возрастающих  $m_1, m_2$ .

525 Выберем  $Z_0$  из условия  $2\mathbf{P}(x > Z_0) = \alpha$ , тогда  $Z_0$  называется *квантилем уровня*  
 526 *значимости*  $\alpha$ ; само  $\alpha$  задаётся вычислителем, например,  $\alpha = 0, 1$ .

527 Значение  $p$ -value для данной строки  $i$ , обозначаемое  $p_{\text{val}}(i)$ , определяется по рас-  
 528 пределению  $N(0, 1)$  как удвоенная площадь под распределением правее числа  $|Z|$ ,  
 529 то есть  $p_{\text{val}}(i) = 2\mathbf{P}(x > |Z|)$ ; а эмпирическое значение  $Z = Z(i)$  вычисляется по  
 530 данной строке  $i$ .

531 Заметим  $|Z(i)| > Z_0 \Leftrightarrow p_{\text{val}}(i) < \alpha$ . Число  $p_{\text{val}}(i)$  называется *текущим уровнем*  
 532 *значимости* с уровнем значимости  $\alpha$ ; например,  $\alpha = 0, 05$ .

533 Если  $p_{\text{val}}(i) < \alpha$ , то строка-признак  $i$  включается в *предварительный список*  
 534 *DE-признаков* данного кластера  $G$  клеток (и 0-я гипотеза отвергается). Иначе для  
 535  $i$  принимается 0-я гипотеза.

536 Дальнейшее сокращение *предварительного списка DE-признаков* связано с воз-  
 537 можной случайностью самого значения  $p_{\text{val}}(i)$ .

## 538 8.2. Дополнение МУВ процедурой Беньямини-Хохберга. Определение $p_{\text{adj}}$ -значения.

539 Переставим строки-features  $i$ ,  $1 \leq i \leq N$ , в матрице  $Y$  по возрастанию самих чисел  
 540  $p_{\text{val}}(i)$ :  $p_{\text{val}}(1) \leq \dots \leq p_{\text{val}}(n)$ , где бывшая строка  $i$  получает какой-то свой номер  $k$ .

541 По определению:

$$p_{\text{adj}}(k) = \min \left\{ \frac{n}{k} \cdot p_{\text{val}}(k), p_{\text{adj}}(k+1) \right\}, p_{\text{adj}}(n) = p_{\text{val}}(k).$$

542 Тогда  $p_{\text{adj}}(k) \leq p_{\text{adj}}(k+1)$  и  $0 < p_{\text{val}}(k) \leq p_{\text{adj}}(k) \leq 1$ .

543 Итак, *DE-признаком* ещё раз предварительно называется строка  $i$  (с новым номе-  
544 ром  $k$ ), которая удовлетворяет в качестве 1-го шага в пункте 8.1 условию  $p_{\text{adj}}(k) < \alpha$ .  
545 Очень слабым вариантом является  $\alpha = 0, 1$ .

546 Обратной индукцией легко проверить: для строки  $k$ , если выполняется  $p_{\text{val}}(k) >$   
547  $\alpha$  (то есть строка  $k$  не-DE-признак относительно  $p_{\text{val}}$ ), то  $p_{\text{adj}}(k) > \alpha$  (эта же строка  
548 – не-DE-признак относительно  $p_{\text{adj}}$ ).

549 На так отобранные DE-признаки накладываются ещё три дополнительных усло-  
550 вия соответственно с параметрами  $q, q_1, q_2$  (например,  $q = 2, q_1 = q_2 = 0, 25$  или  
551 очень слабый вариант параметров  $q = 0, 1, q_1 = 0, 01, q_2 = 0$ ).

552 Первое условие на DE-признаки:

$$\log_2 \text{FC}(i) = \left| \log_2 \left( \frac{1}{m_1} \cdot \sum_{j \in G} x_j \right) - \log_2 \left( \frac{1}{m_2} \cdot \sum_{l \notin G} x_l \right) \right| > q.$$

553 Второе и третье условия на DE-признаки: для признака  $i$  обозначим pct.1 и pct.2  
554 доли клеток с ненулевой экспрессией в кластере  $G$  и в его дополнении  $\neg G$ , соответ-  
555 ственно; и положим:  $\max\{\text{pct.1}, \text{pct.2}\} \geq q_1, |\text{pct.2} - \text{pct.1}| \geq q_2$ .

556 Окончательное определение DE-признака состоит в выполнении всех перечис-  
557 ленных условий.

558 Маркером называется DE-признак, для которого  $p_{\text{adj}}(i) < \beta$  с особо малым значе-  
559 нием  $\delta \ll \alpha$  (например,  $\beta = 10^{-20}$  или меньше); значение  $\beta$  подбирается по данным.

## 560 § 9. Пояснение к процедуре Бенъямини-Хохберга.

561 До конца этого пункта предположим 0-ую гипотезу, обозначив её (\*).

562 Напомним, что распределение для  $Z$  лишь приближается к нормальному распре-  
563 делению; а  $\mathbf{P}(x)$  здесь удобнее понимать как  $\mathbf{P}(x) = \int_x^\infty e^{-\frac{t^2}{2\pi}} dt$ .

564 Тогда вероятность *отвержения* 0-ой гипотезы равна  $\mathbf{P}(p_{\text{val}}(i) < \alpha) = \mathbf{P}(x >$   
565  $\mathbf{P}^{-1}(\alpha/2))$ , где  $\mathbf{P}^{-1}(\alpha/2) = Z_0$  и  $\mathbf{P}^{-1}$  – обратная функция к  $\mathbf{P}$ . Итак, если 0-  
566 ая гипотеза верна, то  $\mathbf{P}(p_{\text{val}}(i) < \alpha) = \alpha$ . Матрица  $Y$  имеет  $N$  строк, поэтому  
567  $\mathbf{P}$ (“для всех строк 0-ая гипотеза принимается”) =  $(1-\alpha)^N$ , а  $\mathbf{P}$ (“для некоторых строк  
568 0-ая гипотеза отвергается”) =  $1 - (1-\alpha)^N$ , что при большом  $N$  кажется парадок-  
569 сальным в предположении (\*). Упомянутая в пункте 8.2 перенумерация строк (от  $i$   
570 на  $k$ ) не влияет на эту кажущуюся парадоксальность.

571 Для процедуры Бенъямини-Хохберга доказано (здесь не приводится), что при  
572 замене условия  $p_{\text{val}}(i) < \alpha$  на условие  $p_{\text{adj}}(i) < \alpha$  получим:  $\mathbf{P}(\{k | p_{\text{adj}}(i) < \alpha\} = \emptyset) \geq$   
573  $1 - \alpha$  что уменьшает соединение маловероятного события с предположением (\*).  
574 Заметим, что замена  $p_{\text{val}}$  на  $p_{\text{adj}}$ , с одной стороны, удаляет ложно дифференциально-

575 экспрессируемые признаки (“ложно-положительные”), а с другой стороны, может  
 576 привести к потере реально дифференциально-экспрессируемых признаков (“ложно-  
 577 отрицательных”).

## 578 § 10. Понижение размерности данных.

579 Если кластеризованное множество точек (данные) в пространстве  $\mathbb{R}^n$  большой  
 580 размерности  $n$  отобразить в пространство  $\mathbb{R}^K$ , где  $K < n$ , например, на плоскость  
 581 ( $K = 2$ ), обычной (линейной) проекцией, то проекции кластеров в  $\mathbb{R}^n$ , т.е. кластеры  
 582 в  $\mathbb{R}^K$ , могут перекрывать друг друга или находиться в проекции ближе, чем по ев-  
 583 клидову расстоянию в  $\mathbb{R}^n$ . Такой неадекватной ситуации стремится избежать нели-  
 584 нейная «проекция», называемая UMAP, рис. 5. Обычно данные в  $\mathbb{R}^n$  и  $\mathbb{R}^K$  представ-  
 585 ляются графами  $G$  и  $G_1$ , вершины которых однозначно соответствуют друг другу и  
 586 точкам соответствующих пространств. Итак, задача понижения размерности состо-  
 587 ит в переходе от данных в  $\mathbb{R}^n$  к соответствующим данным в  $\mathbb{R}^K$ . А точнее, от графа  
 588  $G$  к графу  $G_1$ , у которых одинаковые клетки-вершины.

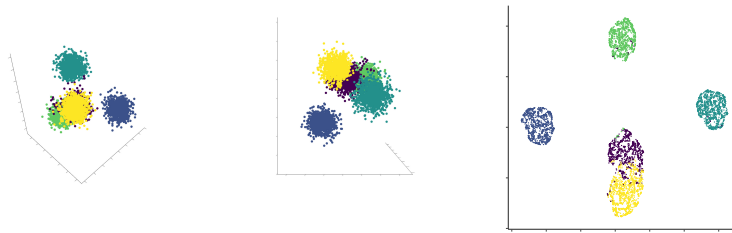


Рис. 5. Показан пример нелинейной проекции UMAP данных (здесь не приводятся)  
 в большой размерности на плоскость ( $K = 2$ ).

589 Повторим точнее, даны клетки-точки  $j$  в  $n$ -ом пространстве с их PCA-координатами  
 590 (и одновременно  $j$  – вершины графа  $G$ ). Как и в пункте 7, для каждой точки-клетки  
 591  $j$  строится список  $k$ -соседей и  $\rho(j, l)$  – евклидово расстояние от  $j$  до её  $k$ -соседа  $l$ .  
 592 Пусть  $\rho_j$  – наименьшее ненулевое расстояние от  $j$  до её ближайшего  $k$ -соседа. Для  
 593 данного  $j$  вычислим нормировочное число  $\sigma_j$ , заданное уравнением

$$\sum_l \exp\left(-\frac{\rho(j, l) - \rho_j}{\sigma_j}\right) = \log_2 k,$$

594 где  $l$  пробегает всех  $k$ -соседей точки-клетки  $j$ .

595 Строится новый граф  $G_1$ , который имеет такие же вершины-клетки, рёбра и  
 596 кластеризацию, какие у графа  $G$ ; но вершины у  $G_1$  однозначно соответствуют точ-

597 кам  $K$ -мерного пространства. Тем самым, точки-вершины у  $G$  переходят в точки-  
 598 вершины у  $G_1$ ; это отображение и называется УМАР, [11].

599 Перейдём к построению *весов* рёбер и *координат* вершин у графа  $G_1$ . Но сначала  
 600 изменим веса в графе  $G$ , и полученный граф обозначим  $G_{01}$ . В нём вершины-клетки  
 601  $j$  и  $l$ , соединяются неориентированным ребром, если они являются  $k$ -соседями друг  
 602 друга, как и было в пункте 7. Но вес ребра  $e = (j, l)$  в  $G_{01}$  положим равным  $W(j, l) =$   
 603  $w(j \rightarrow l) + w(l \rightarrow j) - w(j \rightarrow l) \cdot w(l \rightarrow j), 0 < W(j, l) < 1$ , где вспомогательный  
 604 односторонний вес равен

$$w(j \rightarrow l) = \exp\left(-\frac{\rho(j, l) - \rho_j}{\sigma_j}\right).$$

605 Вес  $W(j, l)$  можно представить себе как вероятность существования хотя бы од-  
 606 ного ребра между вершинами  $j, l$ .

607 Таким образом, получен новый в части весов неориентированный нагруженный  
 608 граф  $G_{01}$  в  $n$ -мерном пространстве “большой размерности”, где те же самые клетки  
 609  $j$  являются вершинами с теми же рёбрами  $e$  и новыми весами  $W(e)$ . Пусть для  
 610 краткости  $K = 2$ .

611 Перейдём к построению *самого графа*  $G_1$  на плоскости: он имеет те же вершины  
 612  $j$  и рёбра  $e$ , что у графов  $G$  и  $G_{01}$ , но опять новые веса. Обозначим  $x(j)$  координаты  
 613 на плоскости клетки-вершины  $j \in G_1$ . Определим в  $G_1$  вес того же, что в  $G_{01}$ , ребра  
 614  $e = (j, l)$ , полагая  $S_{jl} = (1 + a\|x(j) - x(l)\|_2^b)^{-1}, 0 < S_{jl} < 1$ , где  $a$  и  $b$  – параметры  
 615 (“УМАР-веса”). Веса  $S_{jl}$  в  $G_1$  близки к расстоянию, обратному к евклидову между  
 616  $x(j)$  и  $x(l)$ .

617 Осталось вычислить координаты  $x(j)$  всех вершин  $j$  так, чтобы  $G_{01}$  и  $G_1$  бы-  
 618 ли в некотором смысле близки друг к другу. Задача УМАР состоит в том, чтобы  
 619 найти все координаты  $x(j)$  из условия близости двух распределений на одних и тех  
 620 же рёбрах в  $G_{01}$  и  $G_1$ . А именно, найти их из условия минимума следующей функ-  
 621 ции, называемой дивергенцией Кульбака-Лейблера:  $L(x) = \sum_e L(x(j_1), \dots, x(j_M)) =$   
 622  $\sum_e \left[ W(e) \log\left(\frac{W(e)}{S(e)}\right) + (1 - W(e)) \log\left(\frac{1 - W(e)}{1 - S(e)}\right) \right] \rightarrow \min$ , минимизируя по координатам  
 623  $x = x_1, x_2; \dots; x_{M-1}, x_M$  всех вершин в  $G_1$ .

624 Результат минимизации определит искомый 2-мерный граф  $G_1$  на плоскости.  
 625 Повторим: на него переносится кластеризация с графа  $G$  в  $n$ -мерном пространстве.

626 Как обычно, численная минимизация существенно зависит от выбора начальной  
 627 точки. Начальными координатами вершин графа  $G_1$  можно выбрать две первые,  
 628 наиболее информативные координаты точек в  $n$ -мерном пространстве.

629 Минимизируемый функционал перепишем:

$$L(x) = \sum_j \sum_{l \neq j} \left( W_{jl} \cdot \log \frac{W_{jl}}{S_{jl}} + (1 - W_{jl}) \cdot \log \frac{1 - W_{jl}}{1 - S_{jl}} \right) \rightarrow \min$$

630 Получим задачу минимизации:

$$\min_x L(x) = \min_x \left( - \sum_j \sum_{i \neq j} (W_{ji} \cdot \log S_{ji} + (1 - W_{ji}) \cdot \log(1 - S_{ji})) \right) = \min_{x,y} \sum_j \sum_{l \neq j} (W_{jl} \cdot A_{jl}(S) + (1 - W_{jl}) \cdot R_{jl}(S)),$$

631 где

$$A_{jl} = -\log S_{jl}, R_{jl} = -\log(1 - S_{jl}).$$

632 Для минимизации по  $x(j)$  можно использовать алгоритмы градиентного спуска  
633 или стохастического градиентного спуска, в которых итерации заканчиваются, если  
634 минимизируемая функция начинает мало меняться.

### 635 СПИСОК ЛИТЕРАТУРЫ

- 636 1. *Gorbunov K. Yu, Lyubetsky V.A.* An almost exact linear complexity algorithm of the shortest  
637 transformation of chain-cycle graphs //Eprint, Apr 29 2020. arXiv:2004.14351 [math.CO].  
638 <https://doi.org/10.48550/arXiv.2004.14351>
- 639 2. *Gorbunov K. Yu, Lyubetsky V.A.* Linear time additively exact algorithm for transformation  
640 of chain-cycle graphs for arbitrary costs of deletions and insertions // Mathematics 2020,  
641 V.8. N.11. Art. 2001. P. 1-28. <https://doi.org/10.3390/math8112001>
- 642 3. *Gorbunov K. Yu, Lyubetsky V.A.* Algorithms for the reconstruction of genomic structures  
643 with proofs of their low polynomial complexity and high exactness // Mathematics. 2024.  
644 V. 12. N. 6, Art. 817. P. 1-26. <https://doi.org/10.3390/math12060817>
- 645 4. *Weiler P., Lange M., Klein M., Pe'er D. Theis F.* CellRank 2: unified fate mapping in  
646 multiview single-cell data //Nat. Methods. 2024. V. 21. Art. 1196-1205. P. 1-32 <https://doi.org/10.1038/s41592-024-02303-9>
- 648 5. Satija Lab. Seurat - Guided Clustering Tutorial. [https://satijalab.org/seurat/](https://satijalab.org/seurat/articles/pbmc3k_tutorial.html)  
649 [articles/pbmc3k\\_tutorial.html](https://satijalab.org/seurat/articles/pbmc3k_tutorial.html), 2023 (дата обращения: 13 июня 2025).
- 650 6. *Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E.* Fast unfolding of communities  
651 in large networks //Journal of Statistical Mechanics: Theory and Experiment, 2008. V.10.  
652 P10008. P.1-12. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
- 653 7. *Боровков А. А.* Математическая статистика. Дополнительные главы. М.: Наука 1984. –  
654 472 с.
- 655 8. *Гаек Я., Шудак Э.* Теория ранговых критериев. М.: Наука 1971. – 376 с.
- 656 9. *Чубисов Д. М.* Лекции по асимптотической теории ранговых критериев. Вып. 14 М.:  
657 МИАН, 2009. – 186 с.
- 658 10. *Benjamini Y., Hochberg Y.* Controlling the false discovery rate: a practical and powerful  
659 approach to multiple testing //Journal of the Royal Statistical Society. Series B  
660 (Methodological), 1995. V.57. I.1. P.289-300. <https://doi.org/10.2307/2346101>
- 661 11. *Hardle W. K., Sîmar L., Fengler M. R.* Applied Multivariate Statistical Analysis. 6th  
662 Edition. – Springer, 2024. – 637 p.

<i>Любецкий Василий Александрович</i>	Поступила в редакцию
<i>Горбунов Константин Юрьевич</i>	09.09.2025
<i>Пирогов Сергей Анатольевич</i>	После доработки
<i>Хазиев Георгий Андреевич</i>	26.09.2023
<i>Агламазова Анастасия Ильинична</i>	Принята к публикации
Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва	20.10.2023
lyubetsk@iitp.ru	
gorbunov@iitp.ru	
pirogov@iitp.ru	
khaziev@iitp.ru	
aglamazova.an@iitp.ru	