

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
Национальный исследовательский университет
«Высшая школа экономики»

ФАКУЛЬТЕТ МАТЕМАТИКИ

Есаян Армен Давидович

**Гомологические подмостки и их применение к
медицинским данным**

Курсовая работа студента 2 курса
образовательной программы бакалавриата «Математика»

Научный руководитель:
Кандидат физико-математических наук,
доцент
Чернышев Всеволод Леонидович

Москва 2022

Содержание

1 Введение	2
2 Основные математические понятия	3
2.1 Симплексиальные комплексы и гомологии	3
2.2 Фильтрация и устойчивые гомологии	5
3 Гомологические подмостки и минимальный базис	5
3.1 Построение фильтрации симплексиального комплекса Вьеториса-Рипса	5
3.2 Гомологические подмостки	6
3.3 Минимальные гомологические подмостки	8
4 Гомологические подмостки на данных фМРТ	9
4.1 Спектр нормализованного лапласиана	9
4.2 Спектральная кластеризация	10
4.3 Данные	10
4.4 Результаты	10
5 Диаграммы устойчивости на данных фМРТ	12
5.1 Построение диаграмм устойчивости	13
5.2 Расстояние Вассерштейна и расстояние бутылочного горлышка	14
5.3 Данные	15
5.4 Результаты	15
6 Выводы	18

Аннотация

Гомологические подмостки – относительно новый объект в топологическом анализе данных. По своей сути гомологические подмостки представляют из себя взвешенные графы, которые строятся с помощью устойчивых гомологий и содержат в себе информацию об эволюции гомологических базисов вдоль фильтрации. В терминах медицинских данных гомологические подмостки можно интерпретировать как скелетизацию матриц коннективности между областями мозга в заданном атласе. В данной курсовой работе исследуются различия в функциональных паттернах мозга пациентов из двух групп: здоровые пациенты (*controls*) и пациенты с болезнью Альцгеймера (*patients*). Рассматривается метод, включающий в себя построение минимальных нормализованных гомологических подмостков и в дальнейшем применение известных методов из теории сложных сетей таких, как спектральная кластеризация. Также в работе строятся диаграммы устойчивости пациентов на данных фМРТ, предобработанных в трех разных картах и вычисляется расстояние бутылочного горлышка (*bottleneck distance*) между диаграммами устойчивости. Результаты показывают, что спектральная кластеризация не выявляет особых различий между двумя группами.

1 Введение

Представление сложных систем с помощью сетей давно является одним из самых эффективных способов исследовать данные из различных областей науки, в том числе в нейронауках [1, 2], социальных сетях [3], а также эпидемиологии [4]. В данном подходе структура сетей изучается через призму статистических методов, метрик сетей таких, как степени, центральности, модулярность. Благодаря своей простоте интерпретации и элегантности, сетевой анализ был основой для работы с данными, имеющими структуру графов. Однако с развитием нейронаук появилась необходимость в понимании данных, полученных, например, из функциональной магнитно-резонансной томографии (фМРТ), которые по сути представляют из

себя четырехмерные изображения активности отделов мозга. Анализ данных и нейровизуализация фМРТ, электроэнцефалограмм (ЭЭГ), магнитоэнцефалограмм (МЭГ) показывает результаты в клинических исследованиях [5], поэтому встала задача анализа сложных сетей в динамике, учитывая течение времени.

С развитием теории топологического анализа данных (Topological Data Analysis, TDA [6, 7]) – относительно молодой области исследований в задачах анализа данных – начали появляться новые подходы к работе с данными, которые основываются на теоретических результатах из алгебраической топологии и гомологической алгебры. В TDA рассматриваются классические объекты такие, как симплексиальные гомологии, устойчивые гомологии, модули над кольцами и так далее.

В статье “Homological scaffolds of brain functional networks” [8] 2014 года был показан один из весомых результатов применения топологического анализа данных в нейронауке и нейровизуализации. В статье вводится определение гомологических подмостков – объект, способный представлять гомологические характеристики сети коннективности в виде графа с помощью поиска в устойчивых гомологиях базисных циклов. Это позволяет далее сводить задачу к стандартным сетевым методам. Применение такого инструмента было продемонстрировано на медицинских данных: сравнивались гомологические подмостки контрольной и тестовой группы из 15 человек (тестовой группе вводили внутривенно псилоцибин, контрольной – плацебо). Данные представляли из себя фМРТ мозга [8, Results from fMRI networks, с. 5]. Однако возникло несколько проблем, связанных с неоднозначностью выбора базисных циклов в одномерных устойчивых гомологиях. Из-за этого неоднозначно определялись и сами гомологические подмостки. В 2020 году в статье “Homological scaffold via minimal homology bases” [9] математически однозначно были описаны условия, при котором возможно определить гомологические подмостки однозначно: в качестве базисных циклов в гомологиях с помощью алгоритма Дея [10] выбирались те базисные циклы, которые имели наименьшую суммарную длину.

Данная курсовая работа представляет из себя реализацию гомологических подмостков на данных фМРТ пациентов, страдающих болезнью Альцгеймера, и их дальнейшее количественное сравнение с помощью кластеризации спектров лапласиана.

2 Основные математические понятия

2.1 Симплексиальные комплексы и гомологии

В этой части формально определим математические объекты, которые будут необходимы использоваться для построения гомологических подмостков.

Пусть дано дискретное множество $M = [1, 2, \dots, m]$ из m элементов. Введем одно из центральных понятий, с помощью которого мы будем описывать топологическую структуру парцелляций мозга.

Определение 1. *Симплексиальным комплексом* на множестве вершин M называется такая совокупность подмножеств $K \subset 2^M$, что выполняются следующий условия:

1. $I \in K$ и $J \subset I \implies J \in K$
2. $\emptyset \in K$

В таком случае будем называть множество M – вершинами симплексиального комплекса K , а $I \in K$ – его симплексами. Размерностью симплексиального комплекса $\dim K$ называют максимальную размерность симплекса $I \in K$.

Мы дали довольно формальное математическое определение в терминах множеств и подмножеств, но, говоря про симплексиальные комплексы, проще иметь в виду их геометрическую реализацию. По сути симплексиальный комплекс можно представлять как объединение k -симплексов, вложенных в евклидово пространство \mathbb{R}^n . Более формально: пусть e_1, e_2, \dots, e_n – базис в пространстве \mathbb{R}^n , $\Delta_I = \text{conv}(e_i, i \in I)$ – симплекс, выпуклая оболочка векторов

$e_i, i \in I$. Тогда будем называть топологическое пространство $|K| = \bigcup_{I \in K} \Delta_I$ геометрической реализацией симплексиального комплекса K (рис.1).

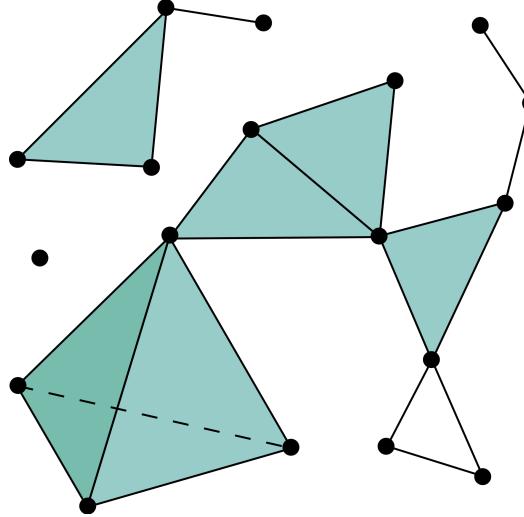


Рис. 1: Пример геометрической реализации симплексиального комплекса

Можно заметить, что симплексиальный комплекс представляет из себя естественное многомерное обобщение обычного двумерного графа. Данный объект является одним из важнейших в топологическом анализе данных и вычислительной топологии[11].

Перейдем к еще одному из центральных понятий из алгебраической топологии – симплексиальным гомологиям. В данной работе будут использоваться в основном одномерные гомологии. Пусть K – симплексиальный комплекс на конечном (упорядоченном) множестве $M = [m]$ с естественным порядком $1 < 2 < \dots < m$. Пусть R – кольцо (далее будем рассматривать только кольцо \mathbb{Z}_2).

Определение 2. Модулем j -цепей $C_j(K, R)$ над кольцом R называется свободный R -модуль, порожденный j -мерными симплексами из K . То есть для $j \geq 0$

$$C_j(K, R) = \{\sum a_I I, a_I \in R, I \in K, \dim I = j\}$$

Данный объект позволяет нам ввести операцию сложения между симплексами одинаковой размерности. Можно построить R -гомоморфизм между модулями j и $(j - 1)$ -мерных цепей $\delta : C_j(K, R) \rightarrow C_{j-1}(K, R)$. Для этого достаточно задать образы гомоморфизма на симплексах $I \in K$ в силу свободности группы, а именно для $I = \{i_0, i_1, \dots, i_j\}$

$$\delta I = \sum_{l=0}^j (-1)^l \{i_0, \dots, \hat{i}_s, \dots, i_j\}$$

Такой гомоморфизм называют **дифференциалом**, или **границным гомоморфизмом** [13, с. 10-14]

В одномерном случае будем рассматривать пространство $C_1(K, \mathbb{Z}_2)$, состоящее из одномерных цепей. Будем называть $\sigma \in C_1(K, \mathbb{Z}_2)$ **1-циклом**, если количество инцидентных рёбер этого цикла (1-симплексов) чётно. Из линейности легко заметить, что циклы образуют модуль $Z_1(K, \mathbb{Z}_2)$, называемое **пространством 1-циклов** симплексиального комплекса.

Определим векторное пространство $B_1(K, \mathbb{Z}_2) \subset Z_1(K, \mathbb{Z}_2)$, порожденное всеми δI , где $I \in K, \dim I = 2$ – двумерные симплексы, а δ – дифференциал. Это пространство называют **пространством 1-границ** симплексиального комплекса K .

Определение 3. Фактор-пространство

$$H_1(K, \mathbb{Z}_2) = \frac{Z_1(K, \mathbb{Z}_2)}{B_1(K, \mathbb{Z}_2)}$$

называют **векторным пространством 1-мерных гомологий** симплексиального комплекса K над кольцом \mathbb{Z}_2

Грубо говоря, наличие ненулевых гомологий даёт нам информацию о наличии “дырок” в симплексиальном комплексе. Действительно, по определению, ненулевой класс эквивалентности симплексиальных гомологий – это циклы, не являющиеся границами, что и интуитивно называют “дырками”.

2.2 Фильтрация и устойчивые гомологии

Ограничение, которое имеют обычные симплексиальные гомологии, заключается в том, что мы работаем с одним симплексиальным комплексом. При работе с реальными данными нужно уметь обрабатывать встречающийся шум и погрешности, поэтому необходимо работать сразу с большим количеством измерений. Одними из основных инструментов топологического анализа данных, которые позволяют решить вышеописанную проблему, являются устойчивые гомологии. Они представляют из себя последовательность вложенных друг в друга симплексиальных комплексов и соответствующие гомоморфизмы между ними. Более формально:

Определение 4. Последовательность

$$\emptyset \subset K_1 \subset K_2 \subset \dots \subset K_n = K$$

вложенных симплексиальных комплексов будем называть **фильтрацией** симплексиального комплекса K .

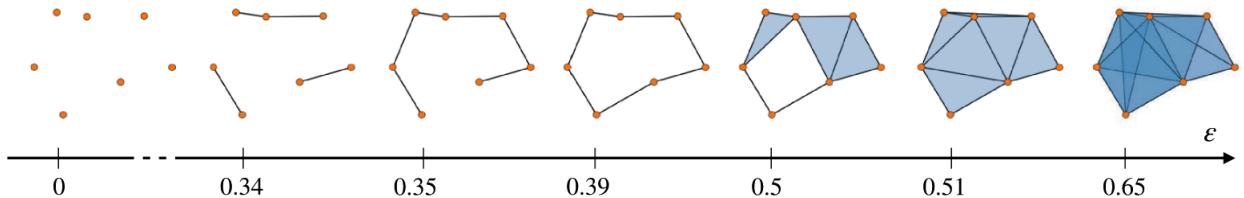


Рис. 2: Параметризованная фильтрация симплексиального комплекса ([9], fig.1a)

Фильтрацию симплексиальных комплексов (рис.2) можно запараметризовать какой-то непрерывной величиной. В таком случае параметр можно интерпретировать как “время” и с помощью него отслеживать “эволюцию” представителей циклов в одномерных гомологиях, таким образом отсеивая шум и различные погрешности в каждом отдельно выбранном комплексе. Математически это можно описать следующим образом: для каждой пары $p < q$ симплексиальных комплексов $K_p \subset K_q$ можно задать естественный гомоморфизм между их симплексиальными гомологиями, то есть гомоморфизм $H_1(K_p) \rightarrow H_1(K_q)$. Такое свойство устойчивых гомологий называют функциональностью, и оно также выполняется и для больших размерностей.

3 Гомологические подмостки и минимальный базис

3.1 Построение фильтрации симплексиального комплекса Вьеториса-Рипса

Чаще всего реальные данные не сразу представлены в виде симплексиальных комплексов. Для того, чтобы построить фильтрацию, есть множество различных способов, которые зависят от того, в каком виде представлены данные. Например, данные могут выглядеть в виде

облака точек в евклидовом пространстве \mathbb{R}^n . В нашем случае мы будем рассматривать взвешенный граф $G = (V, E, w : E \rightarrow \mathbb{R}_{\geq 0})$, где V – множество вершин, E – множество рёбер, w – неотрицательная функция весов. Самый простой способ наделить взвешенный граф симплексиальной структурой и “отфильтровать” его – рассматривать фильтрацию симплексиального комплекса Вьеториса-Рипса. Опишем такой способ построения симплексиального комплекса из взвешенного графа.

Зафиксируем параметр $\varepsilon > 0$. Будем соединять пару вершин ребром в том случае, если расстояние между ними не более, чем ε . Таким образом, мы получим подграф $G^\varepsilon = (V, E^{new}) \subset G$, где $E^{new} = \{e \in E, w(e) \leq \varepsilon\}$.

Для того, чтобы перейти от графа к симплексиальному комплексу, рассмотрим ещё один математический объект. Будем называть симплексиальный комплекс **флаговым**, если в нём любой набор вершин, попарно соединённых рёбрами, образует симплекс[12]. Одно из очевидных свойств флагового комплекса заключается в том, что он полностью определяется своим одномерным остовом. Объяснить это свойство для себя можно так: если у нас есть одномерный “каркас”, то дырки в этом каркасе можно заклеить как раз-таки симплексами различных размерностей, причем единственным способом. Такую операцию называют **флаговым замыканием графа** (обозначение – $FlagG$).

Теперь мы можем полноценно определить фильтрацию для нашего случая. Определим симплексиальный комплекс Вьеториса-Рипса следующим образом:

$$VR^\varepsilon = Flag(G^\varepsilon)$$

Меняя параметр ε , мы получим фильтрацию симплексиальных комплексов, так как очевидно, что $VR^{\varepsilon_1} \subset VR^{\varepsilon_2}$ при $\varepsilon_1 < \varepsilon_2$ (рис.3).

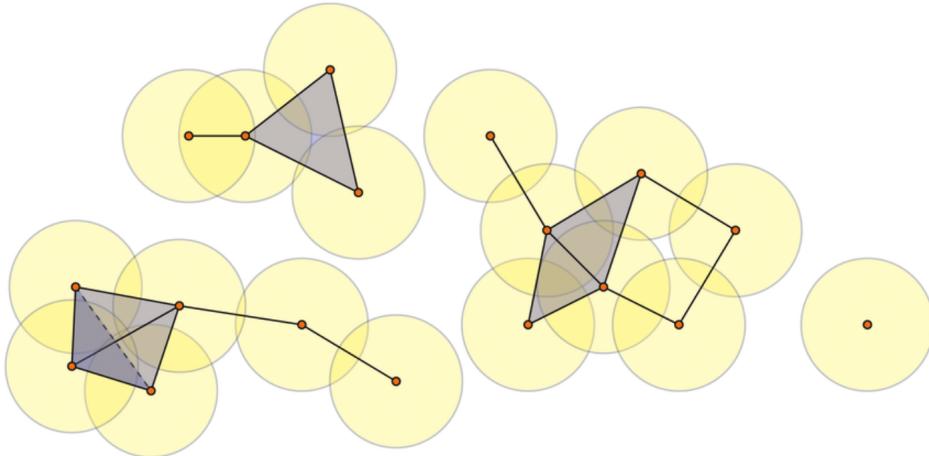


Рис. 3: Пример симплексиального комплекса Вьеториса-Рипса: соединяются все вершины, которые находятся не дальше определенного расстояния ([15], fig.2)

3.2 Гомологические подмостки

Гомологические подмостки представляет из себя способ скелетизации матриц коннективности, который позволяет извлечь основную информацию из фильтрации и выделить наиболее важные связи между вершинами, не обращая внимания на шум и другие погрешности. Дадим определение гомологических подмостков.

Пусть $\{b_i\}$ – множество из базисных циклов одномерных устойчивых гомологий вдоль фильтрации заданного графа G .

Определение 5. Пусть G – взвешенный граф, \mathcal{F} – фильтрация из вышеописанных флаговых комплексов. Определим новую весовую функцию $h_G : E \rightarrow \mathbb{R}^+$ следующим образом:

$$h_G(e) := \sum_i I_{e \in b_i}$$

где $I_{e \in b_i}$ – индикаторная функция, принимающая значение 1, если $e \in b_i$, и 0 иначе.

Тогда гомологическими подмостками взвешенного графа G будем называть взвешенный граф $\mathcal{H}(G) = (V, E_H, h_G : E_H \rightarrow \mathbb{R}^+)$, где $E_H = \{e \in E, h_G(e) > 0\}$.

Базисы $\{b_i\}$ одномерных гомологий играют важную роль в определении гомологических подмостков: гомологические циклы интуитивно описывают “дырки” в данных. Например, в гомологических подмостках будут отсутствовать те ребра, которые ни разу не встречались в базисных циклах устойчивых гомологий. Это можно интерпретировать следующим образом: рёбра, которые ни разу не были границами “дырок” за всё время эволюции, не должны содержать в себе информацию, которая присуща только границам этих “дырок”.

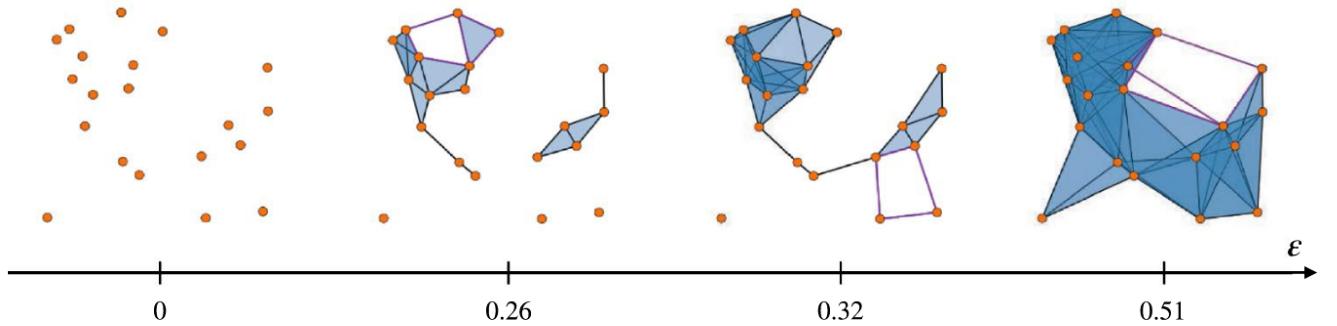


Рис. 4: Базисные циклы в одномерных устойчивых гомологиях ([9], fig.2a)

Однако одна из основных проблем данного определения в том, что вид гомологических подмостков зависит от выбора гомологического базиса. Рассмотрим данное определение внимательнее и заметим несколько его недочетов [9, с. 5-6, 8-10]:

- Базис в пространстве циклов в устойчивых гомологиях необязательно единственный. Как и в обычных векторных пространствах и модулях, всегда можно перейти от одного базиса к другому, это верно и для пространства циклов. Неоднозначность выбора базиса влечет за собой неоднозначность построения гомологических подмостков. Способ разрешить такую неопределенность – выбирать базисные циклы с минимальной суммарной длиной.
- Минимальность суммарной длины базисных циклов тоже не влечет однозначность, так как может быть несколько циклов с минимальной длиной. Эту проблему можно решить несколькими способами. Мы будем рассматривать нормализованные гомологические подмостки, которые решают эту проблему с помощью нормировки весов рёбер в зависимости от количества одинаковых по длине минимальных базисных циклов.
- Наборы возможных базисных циклов могут быть линейно зависимы

Последние две пункта в нашем случае будут не актуальны, так как наши данные будут представлять из себя облако точек, равномерно распределенное в области \mathbb{R}^n , поэтому вероятность того, что длины двух циклов будут равны, будет равна 0, то есть

$$\mathbb{P}[\mu(b_i) = \mu(b_j)] = 0$$

при $i \neq j$.

3.3 Минимальные гомологические подмостки

Рассмотрим взвешенный граф G и элемент из его фильтрации – симплексиальный комплекс K_ε . Будем выбирать в его первых гомологиях базис таким образом, чтобы суммарная длина циклов была минимальной:

$$\{b_1, \dots, b_{\beta_1}\} = \arg \min_{Span([b_i])=H_1} \sum_i \mu(b_i)$$

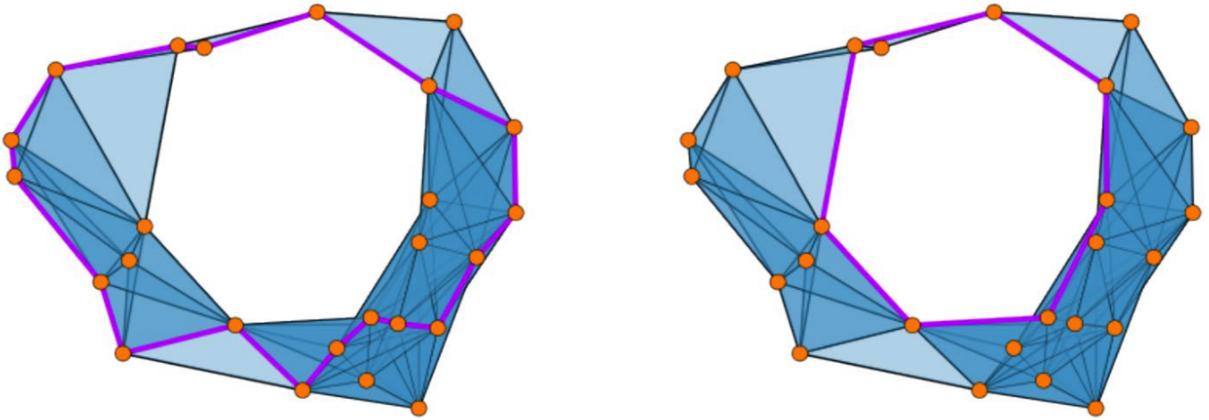


Рис. 5: Выбор минимального базиса в симплексиальном комплексе ([8], fig.3)

Как оказалось, нахождение минимального гомологического базиса – вычислительно трудная задача. Алгоритм нахождения минимального базиса в первых гомологиях с полиномиальной сложностью был впервые представлен в 2018 году и получил название **алгоритм Дея**[10].

В худшем случае данный алгоритм работает за $O(|K|^3)$, где $|K|$ – количество симплексов в симплексиальном комплексе K . Учитывая, что в худшем случае $O(|K|) = O(n^3)$, где n – количество вершин, и что сама фильтрация имеет сложность в худшем случае $O(n^2)$ (так как примерно n^2 рёбер), то построение минимальной гомологической подмостки для заданного взвешенного графа с n вершинами имеет алгоритмическую сложность $O(n^{11})$ [9, Computational complexity, с. 7].

Рассмотрим следующее множество, состоящее из минимальных базисов:

$$B := \bigcup_{i=1}^{\beta_1(K)} \{b_{i,1}, \dots, b_{i,n_i}\}$$

где $\beta_1(K)$ – 1-ое число Бетти (ранг первых гомологий симплексиального комплекса), b_{i,j_i} для $j_i \in \{1, \dots, n_i\}$ гомологичны (то есть являются представителями одного и того же гомологического класса) и для каждого j_i образуют минимальный гомологический базис. Назовем $V_i = \{b_{i,1}, \dots, b_{i,n_i}\}$ **вариантами** i -ого минимального базиса.

Переформулируем данное выше определение гомологических подмосток.

Определение 6. Пусть G – взвешенный граф, \mathcal{F} – фильтрация из симплексиальных комплексов K^ε , B^ε – множество из минимальных гомологических базисов в симплексиальном комплексе K^ε . Агрегируем все множества с минимальными базисами:

$$B^* = \sqcup_\varepsilon B^\varepsilon$$

Тогда определим новую весовую функцию $h_{G,min} : E \rightarrow \mathbb{R}^+$ следующим образом:

$$h_{G,min}(e) := \sum_{V \subset B^*} \frac{1}{|V|} \sum_{b \in V} I_{e \in b}$$

Тогда **нормализованными минимальными подмостками** взвешенного графа G будем называть взвешенный граф $\mathcal{H}_{min}(G) = (V, E_{Hmin}, h_{G,min} : E_H \rightarrow \mathbb{R}^+)$, где $E_{Hmin} = \{e \in E, h_{G,min}(e) > 0\}$.

Переформулировка определения гомологических подмостков помогает решить основную проблему, связанную с предыдущим определением: теперь итоговый взвешенный граф $\mathcal{H}_{min}(G)$ задается однозначно и не зависит от выбора гомологического базиса.

4 Гомологические подмостки на данных фМРТ

Гомологические подмостки позволяют представлять информацию о базисах одномерных устойчивых гомологий в виде репрезентативного взвешенного графа. В этой части рассмотрим задачу поиска отличий в функциональных паттернах мозга людей с помощью минимальных нормализованных гомологических подмостков. Методология исследования заключается в построении гомологических подмостков на медицинских данных (фМРТ мозга), спектральной кластеризации и дальнейшее оценивание качества кластеризации. Также сравниваются спектральные распределения испытуемых из двух групп людей. На рис.6 представлен пайплайн исследования с помощью вышеописанного метода.

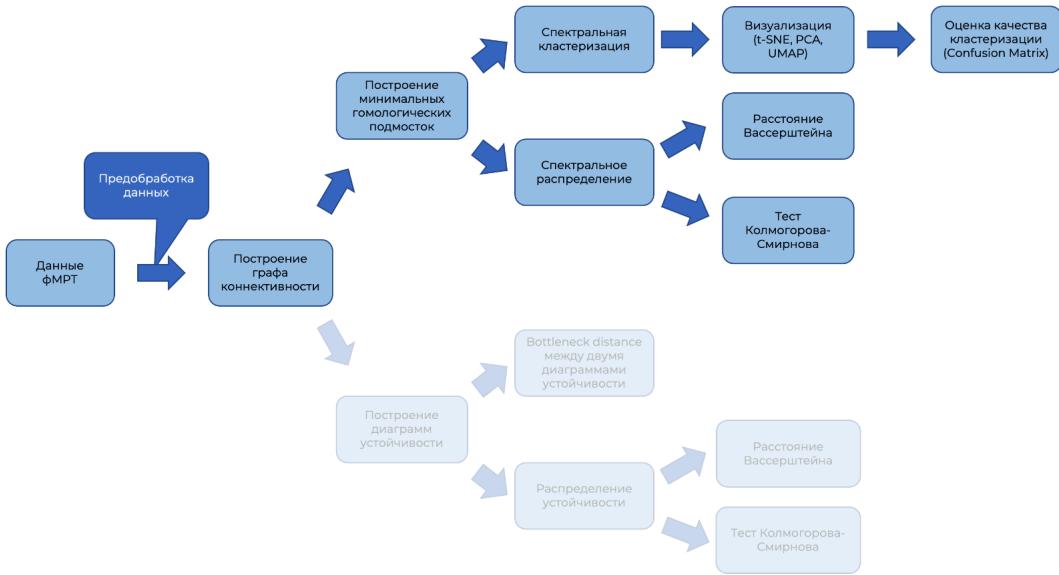


Рис. 6: Пайплайн распознавания пациентов с болезнью Альцгеймера с помощью гомологических подмостков

4.1 Спектр нормализованного лапласиана

Рассмотрим матрицу смежности $A = (a_{ij})$ графа G . Пусть D – диагональная матрица с значениями равными сумме весов инцидентных рёбер, то есть $d_{ij} = \sum_j a_{ij}$.

Определение 7. *Нормализованным лапласианом \mathcal{L} графа G будем называть*

$$\mathcal{L} = D^{-1/2}(D - A)D^{-1/2}$$

Данный объект представляет особый интерес по нескольким причинам. Во-первых, это самосопряженный оператор, а, значит, его матрица диагонализуема. Во-вторых, собственные значения этого оператора принимают вещественные значения от 0 до 2.

4.2 Спектральная кластеризация

Пусть $X = \{x_1, \dots, x_n\}$ – набор из спектров лапласиана гомологической подмостки для каждого испытуемого, $y = \{0, 1\}$ – целевые значения:

$$y = \begin{cases} 1, & \text{если испытуемый из выборки controls} \\ 0, & \text{если испытуемый из выборки patients} \end{cases}$$

В данной курсовой работе рассматривается задача кластеризации спектров (вектор спектра отсортирован по возрастанию и стандартизирован) здоровых и больных пациентов по заданной выборке с помощью алгоритма k -средних. Для визуализации данных спектров и результатов кластеризации используются различные алгоритмы машинного обучения: метод главных компонент (PCA) и стохастическое вложение соседей с t -распределением (t-SNE).

4.3 Данные

Данные для построения гомологических подмостков собраны из датасета OASIS-3 [14] – набора данных продольной нейровизуализации, клинических, когнитивных и биомаркерных данных для болезни Альцгеймера. Датасет состоит из предобработанных (стандартизация к Z -шкале) временных рядов активностей парцеляций мозга. Рассматриваются две группы испытуемых из контрольной (controls – здоровые пациенты) и тестовой группы (patients – больные Альцгеймером). Каждому испытуемому соответствует матрица с размерностью $T \times P$, где T – время, P – анатомические структуры, зависящие от атласа.

В качестве взвешенного графа, по которому будет строится фильтрация графа для определенного пациента, рассматривались два способа построения коннектома, отличающиеся весами рёбер. Вес ребра e , соединяющего i и j -ую вершину в изначальном графе $G = (V, E, w : E \rightarrow \mathbb{R}_{\geq 0})$ задаётся следующим образом:

$$w_1(e) = |cor(v_i(t), v_j(t)|$$

$$w_2(e) = 1 - |cor(v_i(t), v_j(t)|$$

где $v_i(t)$ и $v_j(t)$ – временные ряды активности i и j -ой областей мозга.

Для построения гомологических подмостков и спектрального анализа рассматривался данные из атласа MSDL (рис.7), в котором число областей головного мозга $P = 39$. Для коннектома с весовой функцией w_1 размеры контрольной и тестовой групп составили 176 и 181 человек соответственно, для w_2 – 21 и 20 соответственно.

Данные				
№ выборки	Атлас	Весовая функция	Объём выборки из controls	Объём выборки из patients
1	MSDL	w_1	176	181
2	MSDL	w_2	21	20

4.4 Результаты

Рассмотрим результаты кластеризации в разрезе двух разных весовых функций, которые мы использовали при построения графа коннективности.

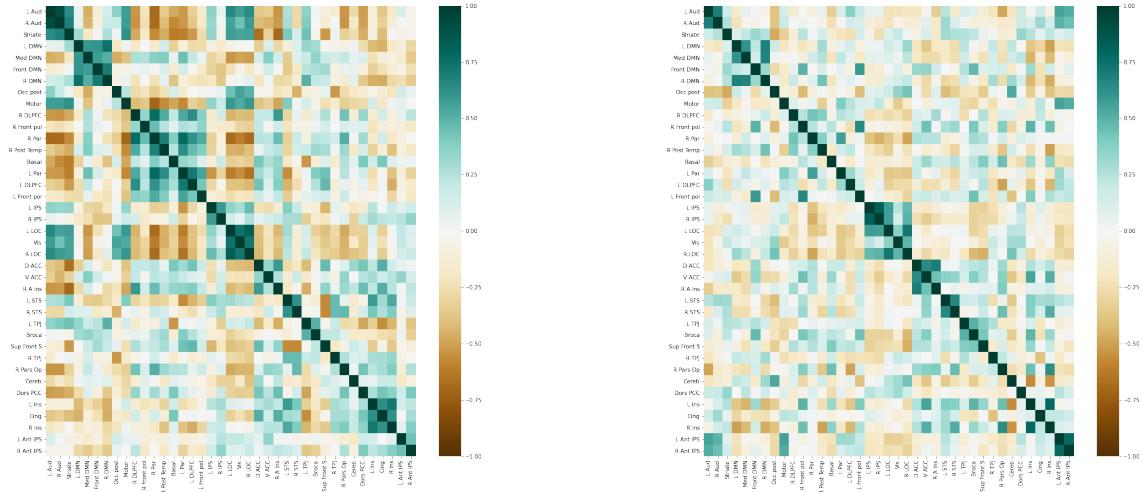


Рис. 7: Слева – матрица коннективности временных рядов здорового пациента, справа – пациента с болезнью Альцгеймера

$$1. w_1(e) = |cor(v_i(t), v_j(t)|$$

Для выборки №1 на левой картинке (рис.8) изображена t-SNE визуализация спектров лапласиана минимальных гомологических подмостков, справа – PCA визуализация результатов кластеризации алгоритмом k -средних. Алгоритм кластеризации относит к классу patients только трёх больных пациентов из 181, причём ошибается в них тоже. Из-за этого точность (Accuracy) предсказания будет очень мала (по сути Accuracy < 0.5 и Recall ≈ 0). Дальнейший подсчет метрик качества кластеризации показал, что такой подход не выявляет кластеров между двумя группами испытуемых.

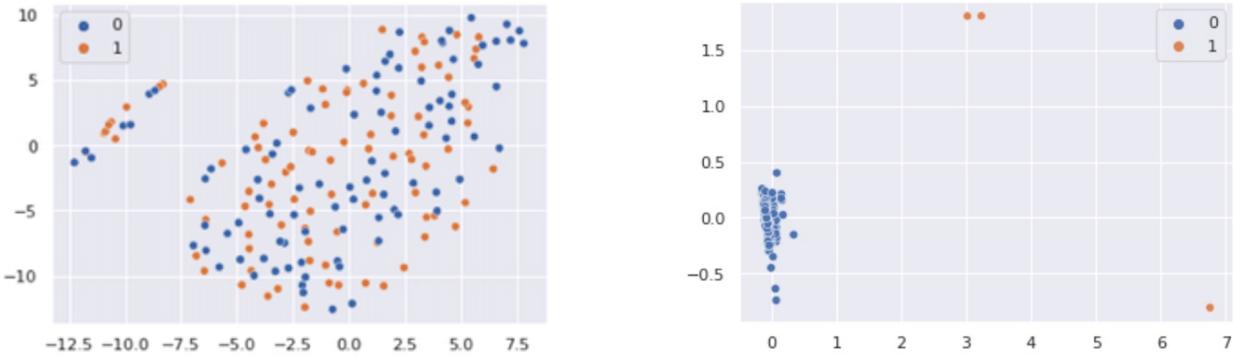


Рис. 8: Слева – t-SNE визуализация спектров лапласиана на выборке №1, справа – PCA визуализация кластеризации алгоритмом k -средних

$$2. w_2(e) = 1 - |cor(v_i(t), v_j(t)|$$

Для выборки №2 изображены (рис.9) изображены PCA визуализация спектров лапласиана, справа – PCA визуализация результатов кластеризации алгоритмом k -средних. Можно заметить, что для такой весовой функции алгоритм тоже не выявляет никаких хороших закономерностей и не способен выявить какие-либо различия между двумя группами из выборки №2. Из матрицы ошибок (рис.10) для полученной кластеризации вычислены следующие метрики: Accuracy = 0.463, Recall = 0.048, Precision = 0.333.

Несмотря на то, что кластеризация не выявляет различий в паттернах гомологических подмостков, спектральные распределения двух групп испытуемых (рис.9) по результатам

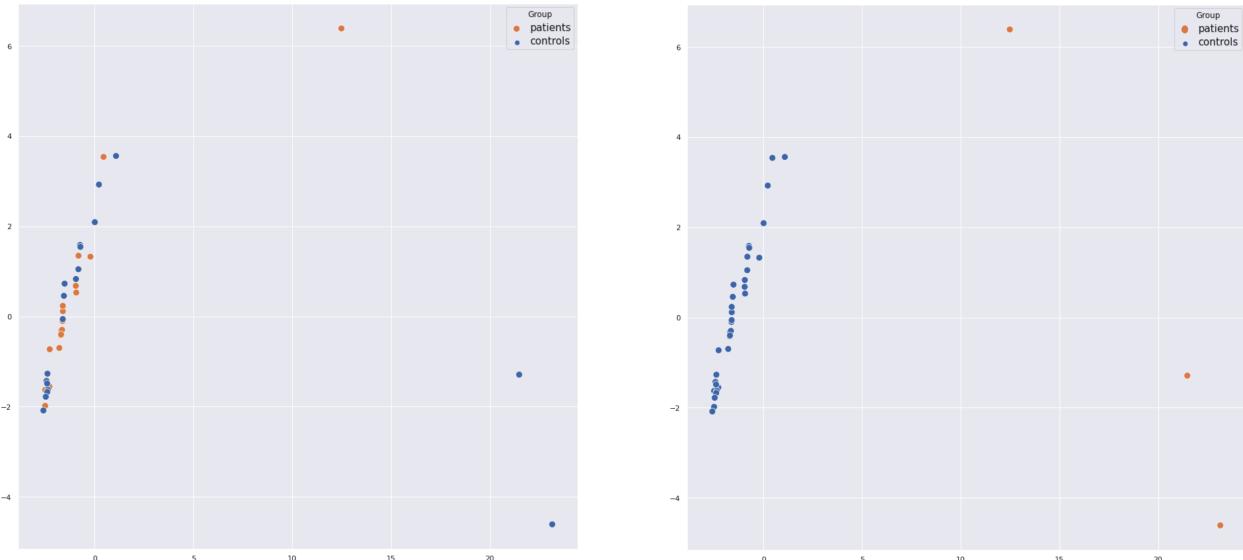


Рис. 9: Слева – PCA визуализация спектров лапласиана на выборке №2, справа – PCA визуализация кластеризации алгоритмом k-средних

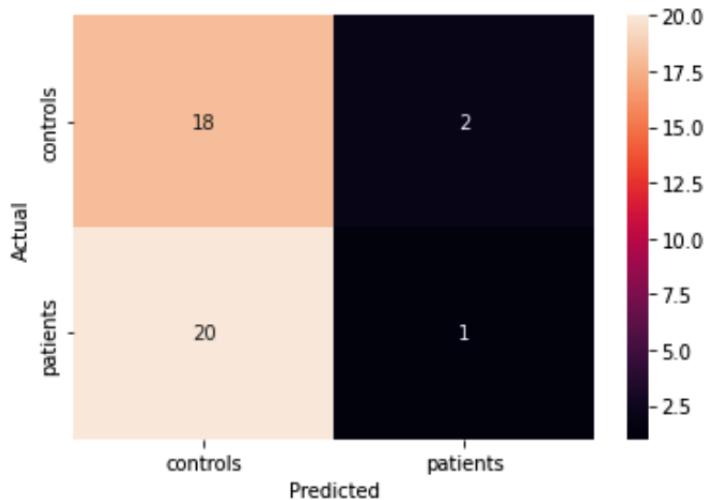


Рис. 10: Матрица ошибок кластеризации на выборке №2

двустороннего теста Колмогорова-Смирнова являются статистически различными на уровне значимости в 5% (KS statistic ≈ 0.074 , pvalue ≈ 0.024). Однако возможно такое, что на выборках с большим количеством измерений, pvalue будет больше уровня значимости, поэтому для конкретных выводов достаточно мало вычисленных данных.

5 Диаграммы устойчивости на данных фМРТ

В этой части рассматривается другой подход, с помощью которого можно изучать гомологическую структуру графов коннективности. Устойчивые гомологии позволяют следить за “эволюцией” гомологий вдоль фильтрации. Нетрудно заметить, что с течением времени структура симплициальных комплексов из фильтрации преобразуется: могут как появляться (рождаться) новые, так и исчезать (умирать) более старые симплексы. “Возраст” гомологии (разница между временем смерти и временем рождения определенного симплекса) принято называть устойчивостью. Одним из способов визуализировать такой эффект устойчивых гомологий – диаграммы устойчивости. На рис.12 представлен пайплайн исследования структур

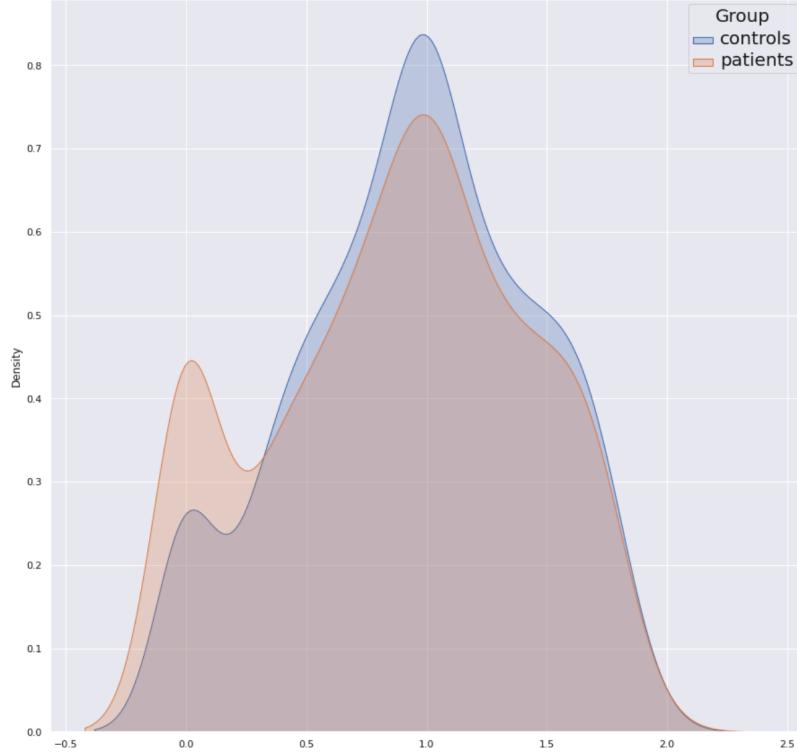


Рис. 11: Сравнение спектральных распределений на выборке №2. Графики представляют функции плотности вероятности, полученные из ядерной оценки плотности (KDE)

мозга с помощью диаграмм устойчивости.

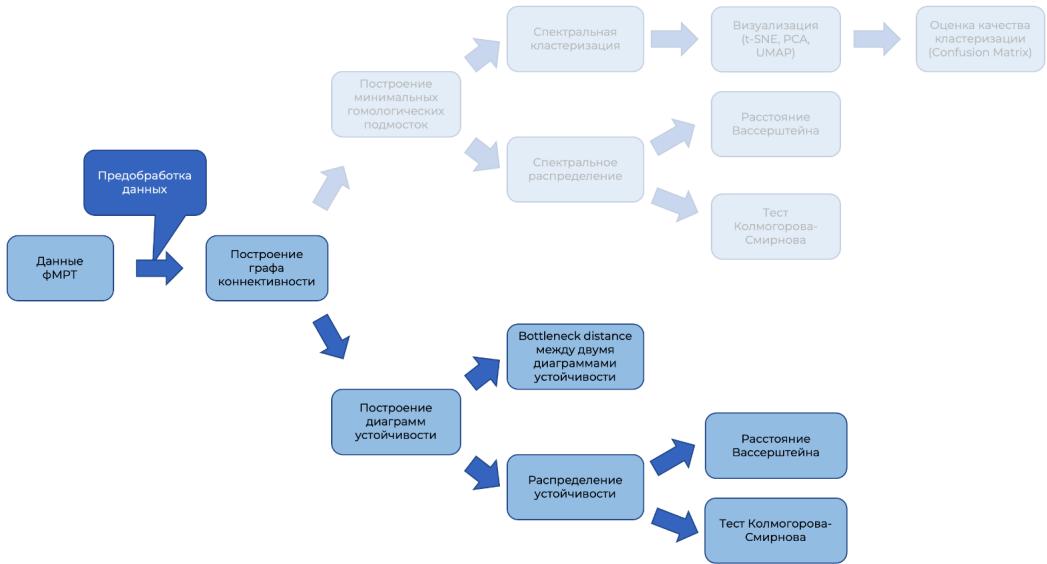


Рис. 12: Пайплайн распознавания пациентов с болезнью Альцгеймера с помощью диаграмм устойчивости

5.1 Построение диаграмм устойчивости

Рассмотрим фильтрацию графа G симплексальными комплексами Вьеториса-Рипса. $\{K^\varepsilon\}$ с параметром $\varepsilon > 0$. В нулевой момент времени $\varepsilon = 0$ всё, что мы имеем от нашего графа, – это вершины. С увеличением параметра ε начинают создаваться новые симплексы и, соответственно, новые гомологические циклы. В данном случае мы будем смотреть только

за одномерными гомологиями. Мы можем заметить, что в течении времени могут как возникать новые гомологические циклы, так и умирать старые. Пусть гомологический 1-цикл создается рождается в фильтрации в момент времени ε_1 и умирает в момент времени ε_2 . Тогда “жизнь” этого цикла можно описать координатами $(\varepsilon_1, \varepsilon_2)$ в плоскости \mathbb{R}^2 . Рассмотрим все такие гомологические циклы, которые когда-либо рождались в эволюции, и нанесём их всех на плоскость \mathbb{R}^2 . Полученное множество точек называют **диаграммой устойчивости** (рис.13).

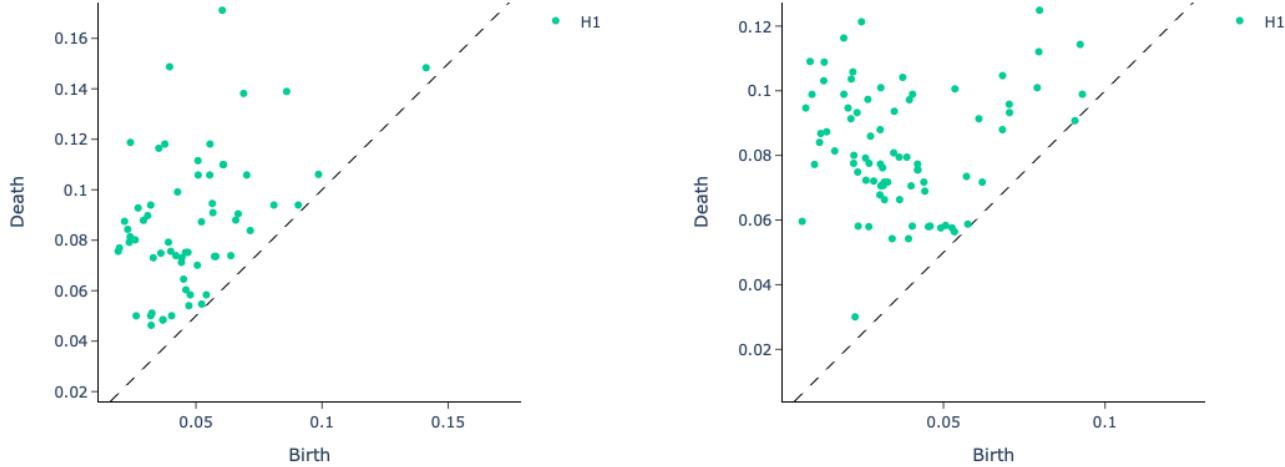


Рис. 13: Примеры диаграмм устойчивости фильтрации графа коннективности

Очевидно, что $\varepsilon_1 < \varepsilon_2$, поэтому диаграмма устойчивости содержит все свое точки над диагональю (так же некоторые точки могут лежать на диагонали в том случае, они погибли сразу же после рождения). Такой случай можно рассматривать как шум в данных, так как это означает, что гомологический цикл имел очень короткую незначительную жизнь и, скорее всего, не содержал в себе никакого глубокого топологического смысла.

5.2 Расстояние Вассерштейна и расстояние бутылочного горлышка

Меру разницы двух диаграмм устойчивости можно определять с помощью вычисления между ними расстояния Вассерштейна[15]. Пусть задана пара диаграмм устойчивости D_1 и D_2 , Δ – множество точек на бесконечности (диагонали).

Определение 8. *p-ым расстоянием Вассерштейна* между диаграммами устойчивости D_1 и D_2 называется

$$\inf_{\gamma} \left(\sum_{x \in D_1 \cup \Delta} \|x - \gamma(x)\|_{\infty}^p \right)^{1/p}$$

где $\gamma : D_1 \cup \Delta \rightarrow D_2 \cup \Delta$ – биекция между точками диаграмм устойчивости D_1 и D_2 и $\|\cdot\|_{\infty}$ – L_{∞} -норма, определенная как $\|x - y\|_{\infty} = \max(|x|, |y|)$

Устремляя $p \rightarrow \infty$ получаем, так называемое, **расстояние бутылочного горлышка** (bottleneck distance):

$$\inf_{\gamma} \left(\sup_{x \in D_1 \cup \Delta} \|x - \gamma(x)\|_{\infty} \right)$$

5.3 Данные

Данные для построения и сравнения диаграмм устойчивости взяты из того же источника, что и в главе 4.1. Построение диаграмм устойчивости и расчёт расстояния Вассерштейна вычислительно более простой метод, чем построение минимальных гомологических подмостков, поэтому для него удалось вычислить метрики на большем количестве данных. В данном методе рассматриваются графы коннективности только с весовой функцией w_1 .

Данные				
№ выборки	Атлас	Весовая функция	Объём выборки из controls	Объём выборки из patients
1	MSDL	w_1	1292	813
2	AAL	w_1	1295	815
3	BASC	w_1	1367	821

5.4 Результаты

Были вычислены расстояния бутылочного горлышка между общими диаграммами устойчивости между группами испытуемых (controls и patients). Результаты вычислений представлены в следующей таблице:

Расстояние Bottleneck Distance между диаграммами устойчивости		
№ выборки	Атлас	Bottleneck Distance
1	MSDL	0.09100
2	AAL	0.07356
3	BASC	0.18872

Можно заметить, что наибольшее различие между диаграммами устойчивости возникают в атласе BASC, что можно объяснить так: среди всех рассматриваемых атласов BASC является самым объёмным. В парцеляции в атласе BASC насчитывается 197 отделов мозга, поэтому такое большое количество парцеллятов может обнаруживать более чувствительные различия в гомологических структурах данных. Далее для визуализации и интерпретации результатов для каждого атласа были построены сами диаграммы устойчивости (рис.14-16)

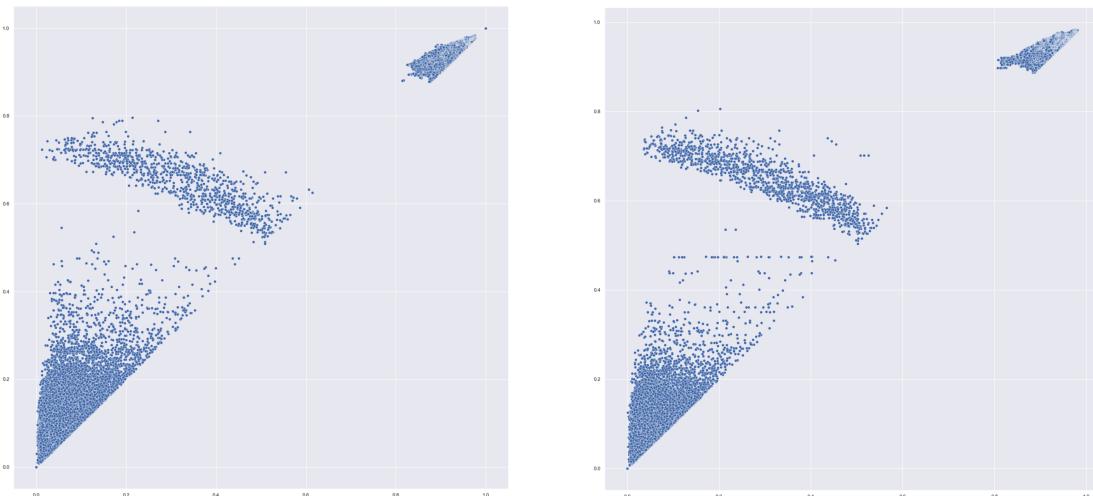


Рис. 14: Диаграммы устойчивости в атласе MSDL у здоровых пациентов (слева) и больных Альцгеймером (справа)

Опишем отличительные свойства диаграмм устойчивости, полученных из данных мозга.

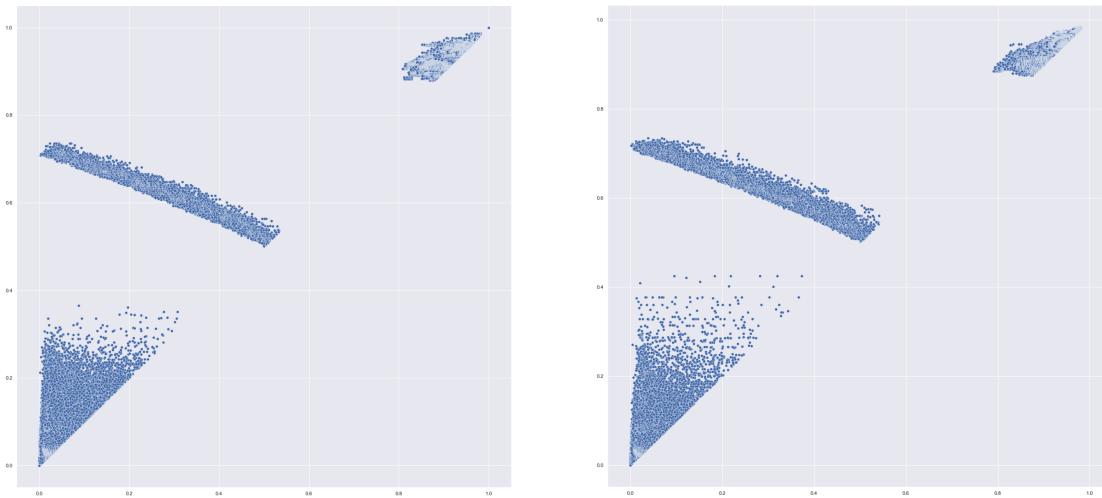


Рис. 15: Диаграммы устойчивости в атласе AAL у здоровых пациентов (слева) и больных Альцгеймером (справа)

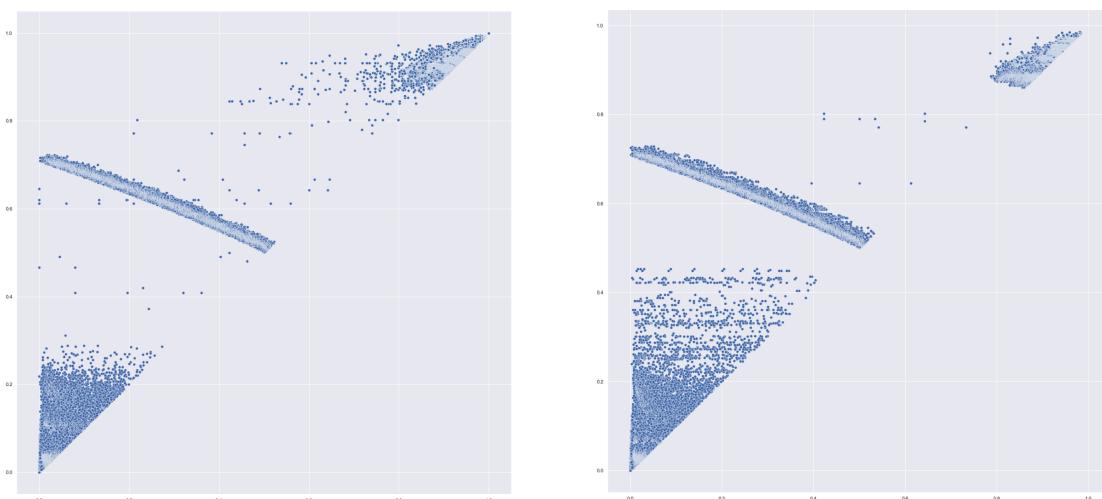


Рис. 16: Диаграммы устойчивости в атласе BASC у здоровых пациентов (слева) и больных Альцгеймером (справа)

- В устойчивых гомологиях графа коннективности отделов мозга гомологические циклы разделяются на три определенных временных промежутка – это следует из того, что диаграммы устойчивости состоят из трёх явно выраженных скопления точек. Для определённости пронумеруем эти скопления от 1 до 3, двигаясь вдоль диагонали.
- В каждом атласе диаграммы устойчивости групп с больными пациентами выражены характерными выбросами с левом нижнем углу.
- Наиболее интересные закономерности в диаграммах устойчивости выявляются в атласе BASC: у здоровых пациентов выбросы возникают в третьем скоплении, а у больных – в первом.

Также были построены распределения устойчивостей гомологических циклов (рис. 17), проведены статистические тесты с нулевой гипотезой о равенстве распределений с помощью двустороннего теста Колмогорова-Смирнова и рассчитаны расстояния Вассерштейна для $p = 1$.

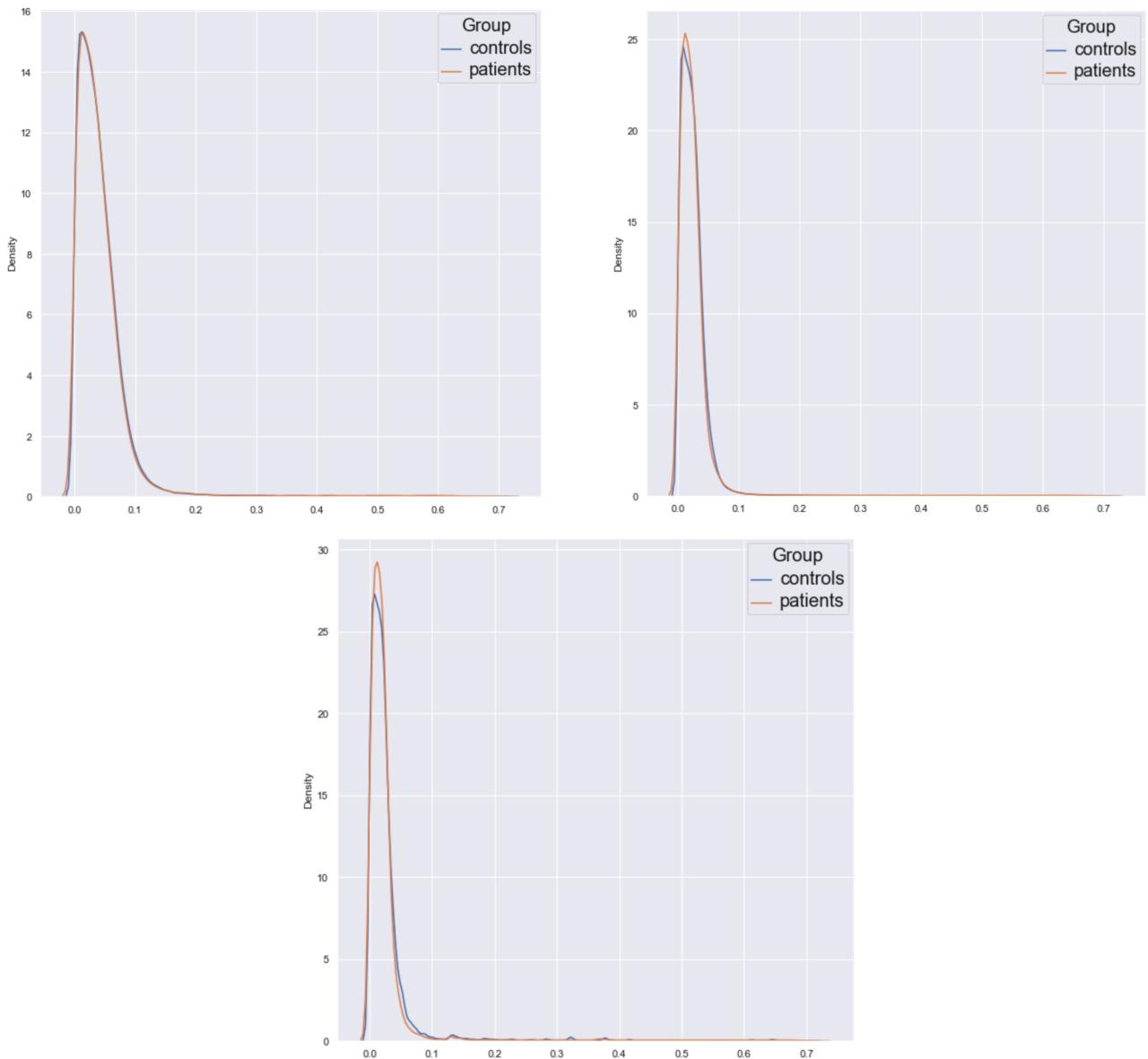


Рис. 17: Распределения устойчивостей гомологий для каждого атласа сравниваются между группами испытуемых (controls/patients). Против часовой стрелки с первой картинки – MSDL, AAL, BASC

Сравнение распределений устойчивости между группами controls и patients			
№ выборки	Атлас	KS-statistic	pvalue
1	MSDL	0.011458	3.73646e-05
2	AAL	0.02980	1.57846e-206
3	BASC	0.05051	0.0

Статистический двусторонний тест Колмогорова-Смирнова о равенстве распределений устойчивости даёт сильно меньший pvalue, чем заданный уровень значимости в 5% , из чего можно сделать вывод, что нулевая гипотеза неверна. Значит, распределения устойчивости действительно различаются. Результаты вычисления расстояний Вассерштейна дают похожие результаты, что и с диаграммами устойчивости: в атласе BASC распределения различаются больше всего.

6 Выводы

В данной работе были рассмотрены базовые определения из топологического анализа данных, даны определения гомологических подмостков, а также минимальных нормализованных гомологических подмостков. В исследовательской части курсовой работы считались гомологические подмостки, построенные из фильтрации графов коннективности отделов мозга (в данном методе рассматривался атлас MSDL). Несмотря на то, что для построения графа коннективности использовались две разные весовые функции w_1 и w_2 , выявить особые паттерны в функциональной активности у больных Альцгеймером не удалось в обоих случаях. Спектральная кластеризация не справляется с тем, чтобы выявлять особые кластеры, на которые можно было бы поделить группы из здоровых и больных пациентов. Значения метрик качества кластеризации слишком низкие, чтобы утверждать о хорошей предсказательной силе данного метода.

Также мы рассмотрели один из основных методов визуализации устойчивых гомологий – диаграммы устойчивости. На более объемной выборке были построены диаграммы устойчивости для трёх разных атласов (MSDL, AAL, BASC) и выявлены интересные закономерности, которые могут зависеть от гомологической структуры человеческого мозга.

Список литературы

- [1] Bassett, D. S. Sporns, O. Network neuroscience. *Nat. Neurosci.* 20, 353 (2017)
- [2] Bassett, D. S. Bullmore, E. Small-world brain networks. *Neuroscientist* 12, 512–523 (2006)
- [3] Vega-Redondo, F. *Complex Social Networks* (Cambridge University Press, Cambridge, 2007)
- [4] Pastor-Satorras, R., Castellano, C., Van Mieghem, P. Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* 87, 925 (2015)
- [5] Pandit AS, Expert P, Lambiotte R, Bonnelle V, Leech R, Turkheimer FE Sharp DJ. 2013 Traumatic brain injury impairs small-world topology. *Neurology* 80, 1826–1833
- [6] Ghrist, R. *Elementary Applied Topology* (Createspace, Seattle, 2014)
- [7] Patania, A., Vaccarino, F. Petri, G. Topological analysis of data. *EPJ Data Sci.* 6, 7 (2017)
- [8] Petri, G. et al. Homological scaffolds of brain functional networks. *J. R. Soc. Interface* 11, 20140873 (2014)
- [9] Guerra, M., De Gregorio, A., Fugacci, U., Petri, G., Vaccarino, F. (2020). Homological scaffold via minimal homology bases. arXiv:2004.11606. <https://doi.org/10.1038/s41598-021-84486-1>
- [10] Dey, T. K., Li, T. Wang, Y. Efficient algorithms for computing a minimal homology basis. In *Latin American Symposium on Theoretical Informatics* 376–398 (Springer, 2018)
- [11] Edelsbrunner H, Harer J. 2010 Computational topology: an introduction. Providence, RI: American Mathematical Society.
- [12] Evans TS. 2010 Clique graphs and overlapping communities. *JSTAT* 2010, P12037. ([doi:10.1088/1742-5468/2010/12/P12037](https://doi.org/10.1088/1742-5468/2010/12/P12037))

- [13] Прасолов В.В. Элементы теории гомологий. — М.: МЦНМО, 2006. — 448 с.: ил.ISBN 5-94057-245-6
- [14] <https://www.oasis-brains.org>
- [15] Chad M. Topaz, Ziegelmeier, L., Halverson, T. Topological Data Analysis of Biological Aggregation Models (2014)
- [16] Kerber, M., Morozov, D., Nigmetov, A. Geometry Helps to Compare Persistence Diagrams (2016)