**УДК 512.643.5:303.722.3**

# ВЫБОР ГЕОМЕТРИИ ВЛОЖЕНИЯ ДЛЯ АНАЛИЗА ДАННЫХ ПО СПЕКТРУ ОДНОЙ МАТРИЦЫ РАССТОЯНИЙ

**Козлов Д.А.** [1]
**Научный руководитель — кандидат физико-математических наук, доцент**
**Возианова А.В.** [1]
[1]Университет ИТМО
dima.dima0029@gmail.com

## Введение

Во многих задачах анализа данных объекты естественно задаются не признаками, а попарными расстояниями или мерой несходства: например, по метрикам сходства профилей, различиям временных рядов или агрегированным показателям. Для визуализации, поиска выбросов и интерпретации структуры данных часто применяют многомерное шкалирование (англ. multidimensional scaling - MDS) [1]. Классический MDS предполагает евклидову природу расстояний, при отклонении от этого предположения возникают искажения взаимного положения групп объектов и снижается информативность визуализации [1]. В последние годы активно развиваются вложения в пространства постоянной кривизны (гиперболические и сферические), которые могут лучше описывать структуру данных в задачах анализа расстояний [2,3]. При этом практической проблемой остаётся выбор подходящей геометрии и параметров кривизны (радиус/кривизна) по матрице расстояний. В прикладных задачах одна и та же совокупность объектов может порождать различные матрицы «расстояний» в зависимости от выбора метрики. Это делает актуальным исследование того, как выбор метрики влияет на спектральные свойства матриц, возникающих в MDS-подобных методах, и, как следствие, на выбор наиболее подходящей геометрии вложения.

## Основная часть

Предлагается подход спектрального выбора геометрии как этап предобработки и диагностики качества вложений для задач анализа данных. Из исходной матрицы расстояний строятся три матрицы-представления, каждая из которых в идеальном случае является Gram-матрицей соответствующей геометрии: евклидовой (через double-centering квадратов расстояний) [1], гиперболической (через преобразование типа гиперболического косинуса и лоренцеву Gram-структуру, как в HYDRA) [2], и сферической (через преобразование косинусом, соответствующее угловой близости на сфере) [3]. Для каждой модели вводится спектрально вычислимый показатель согласованности данных с геометрией, основанный на ошибке лучшей аппроксимации соответствующего класса Gram-матриц заданной «сложности» (ранг/размерность). Параметр кривизны (радиус/кривизна) трактуется как гиперпараметр модели и подбирается по сетке значений: спектральные индикаторы используются для быстрого скрининга, после чего выполняется локальная дооптимизация вложения и сравнение моделей по единой метрике качества восстановления расстояний. Для сферической модели рассматриваются практические итерационные процедуры (проекции/оптимизация на многообразии), обеспечивающие получение точек именно на сфере. В дальнейшем планируется провести анализ того, как спектральные характеристики (распределение собственных значений, инерция, величина отрицательной части спектра и спектральные разрывы) меняются при построении матриц «расстояний» с использованием различных метрик и норм, и как эти изменения коррелируют с качеством евклидовых/гиперболических/сферических вложений.

## Выводы

Предлагаемый подход даёт аналитически интерпретируемый и вычислительно эффективный инструмент для задач анализа данных, позволяющий:

1) диагностировать, когда евклидовая модель расстояний даёт систематические искажения;

2) выбирать более подходящую геометрию вложения (евклид/гиперболика/сфера) и параметр кривизны;

3) получать устойчивые инициализации для последующей дооптимизации и визуализации;

4) проводить сравнительный анализ альтернативных метрик/мер несходства через их спектральный «профиль» и влияние на качество вложения.

Практическая применимость иллюстрируется на данных, где расстояния строятся из нормированных реальных показателей (например, профили успеваемости студентов по нескольким активностям и посещаемости), а также может быть использована для кластеризации и выявления аномалий в любых задачах, где матрица расстояний является основной формой представления данных.

**Литература**

1. Borg I., Groenen P. J. F. Modern multidimensional scaling: Theory and applications. – New York, NY : Springer New York, 2005.

2. Keller-Ressel M., Nargang S. Hydra: a method for strain-minimizing hyperbolic embedding of network-and distance-based data //Journal of Complex Networks. – 2020. – Т. 8. – №. 1. – С. cnaa002.

3. Wilson R. C. et al. Spherical and hyperbolic embeddings of data //IEEE transactions on pattern analysis and machine intelligence. – 2014. – Т. 36. – №. 11. – С. 2255-2269.