MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 27, NO. 1, 2020

journal homepage: www.mais-journal.ru

COMPUTING METHODOLOGIES AND APPLICATIONS

Hierarchical Clustering as a Dimension Reduction Technique for Markowitz Portfolio Optimization

A. Y. Poletaev[1], E. M. Spiridonova[1]

DOI: 10.18255/1818-1015-2020-1-62-71

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya, Yaroslavl 150003, Russia.

MSC2020: 62-08
Research article
Full text in Russian

Received December 12, 2019
After revision February 14, 2020
Accepted February 28, 2020

Optimal portfolio selection is a common and important application of an optimization problem. Practical applications of an existing optimal portfolio selection methods is often difficult due to high data dimensionality (as a consequence of the large number of securities available for investment). In this paper, a method of dimension reduction based on hierarchical clustering is proposed. Clustering is widely used in computer science, a lot of algorithms and computational methods have been developed for it. As a measure of securities proximity for hierarchical clustering Pearson pair correlation coefficient is used. Further, the proposed method's influence on the quality of the optimal solution is investigated on several examples of optimal portfolio selection according to the Markowitz Model. The influence of hierarchical clustering parameters (intercluster distance metrics and clustering threshold) on the quality of the obtained optimal solution is also investigated. The dependence between the target return of the portfolio and the possibility of reducing the dimension using the proposed method is investigated too. For each considered example in the paper graphs and tables with the main results of the proposed method - application which are the decrease of the dimension and the drop of the yield (the decrease of the quality of the optimal solution) - for a portfolio constructed using the proposed method compared to a portfolio constructed without the proposed method are given. For the experiments the Python programming language and its libraries: scipy for clustering and cvxpy for solving the optimization problem (building an optimal portfolio) are used.

Keywords: clustering; optimization; Markowitz portfolio

INFORMATION ABOUT THE AUTHORS

Anatoliy Y. Poletaev correspondence author	orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru graduate student.
Elena M. Spiridonova	orcid.org/0000-0002-1089-7072. E-mail: lena@uniyar.ac.ru Sc.D.

For citation: A. Y. Poletaev and E. M. Spiridonova, "Hierarchical Clustering as a Dimension Reduction Technique for Markowitz Portfolio Optimization", *Modeling and analysis of information systems*, vol. 27, no. 1, pp. 62-71, 2020.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 27, № 1, 2020
сайт журнала: www.mais-journal.ru

Иерархическая кластеризация как метод снижения размерности в задаче оптимизации инвестиционного портфеля Марковица

А. Ю. Полетаев[1], Е. М. Спиридонова[1]

DOI: 10.18255/1818-1015-2020-1-62-71

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, 14, Ярославль, 150003, Россия.

УДК 311.2:004.021
Научная статья
Полный текст на русском языке

Получена 12 декабря 2019 г.
После доработки 14 февраля 2020 г.
Принята к публикации 28 февраля 2020 г.

Составление оптимального портфеля ценных бумаг является важным и частым случаем решения задачи оптимизации. Практическое применение существующих методов составления оптимального портфеля часто затруднено из-за большого числа доступных для инвестирования ценных бумаг (и, как следствие, большой размерности исходных данных). В данной работе предлагается метод снижения размерности исходных данных, основанный на иерархической кластеризации доступных для инвестирования ценных бумаг. Для кластеризации, широко используемой в компьютерных науках, уже разработано множество алгоритмов и методов. В качестве меры близости ценных бумаг для иерархической кластеризации используется коэффициент парной корреляции Пирсона. Далее исследуется влияние предложенного метода на качество получаемого оптимального решения на нескольких примерах составления оптимального портфеля ценных бумаг по модели Марковица. Также исследуется влияние параметров иерархической кластеризации (метрики межкластерного расстояния и порогового значения кластеризации) на изменение качества получаемого оптимального решения. Исследуется зависимость между целевой доходностью портфеля и возможностью снижения размерности с помощью предложенного метода. Для каждого рассмотренного примера приводятся графики и таблицы с основными полученными результатами применения метода — понижением размерности и падением доходности (снижением качества оптимального решения) у портфеля, построенного с применением предложенного метода по сравнению с портфелем, построенным без применения предложенного метода. Для проведения экспериментов используется язык программирования Python и его библиотеки: scipy для проведения кластеризации и cvxpy для решения задачи оптимизации (построения оптимального портфеля).

Ключевые слова: кластеризация; оптимизация; портфель Марковица

ИНФОРМАЦИЯ ОБ АВТОРАХ

Анатолий Юрьевич Полетаев
автор для корреспонденции
orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru
магистрант.

Елена Михайловна Спиридонова
orcid.org/0000-0002-1089-7072. E-mail: lena@uniyar.ac.ru
докт. экон. наук, доцент.

Для цитирования: A. Y. Poletaev and E. M. Spiridonova, “Hierarchical Clustering as a Dimension Reduction Technique for Markowitz Portfolio Optimization”, *Modeling and analysis of information systems*, vol. 27, no. 1, pp. 62-71, 2020.

Введение

Составление оптимального портфеля ценных бумаг является важным и частым случаем решения задачи оптимизации. Согласно портфельной теории, впервые сформулированной Гарри Марковицем в 1952 г., для составления оптимального портфеля из n ценных бумаг необходимо оценить лишь два показателя [1].

1. ожидаемую доходность $R = \sum_{i=1}^{n} R_i X_i$;
2. меру риска (изменчивости) $V = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{i,j} X_i X_j$.

Здесь R_i — ожидаемая доходность i-ой ценной бумаги; X_i — доля средств, инвестированных в неё ($\sum_{i=1}^{n} X_i = 1$); $\sigma_{i,j}$ — ковариация доходностей ценных бумаг i и j.

Портфель может быть оптимизирован по заданной ожидаемой доходности (для минимизации риска), по заданному риску (для максимизации доходности) и по RAPOC (risk-adjusted return) — тогда максимизируется функция $R - \gamma V$, где γ — некоторый коэффициент. Результатом оптимизации является вектор долей $X = [X_1, \dots, X_n]$.

В настоящее время разработано достаточно много математических методов оптимизации портфеля по Марковицу [2, 3], однако их общим недостатком является достаточно высокая вычислительная сложность. Учитывая, что объём биржевых данных, как правило, велик (например, в 2015 году только на Нью-Йоркской фондовой бирже торговались акции более 3000 компаний), а оптимизация портфеля на динамичном рынке может требоваться достаточно часто, необходимо искать пути ускорения оптимизации.

1. Предлагаемый метод понижения размерности

Предлагаемый подход заключается в предварительной кластеризации — разделении n доступных ценных бумаг на k групп (кластеров) ($k < n$). Кластеризация проводится иерархическим методом, в качестве матрицы расстояний используется матрица парных корреляций доходностей ценных бумаг.

Затем для каждого кластера рассчитывается доходность, как средняя доходностей входящих в него ценных бумаг, и строится ковариационная матрица доходностей кластеров. После этого можно будет решать задачу оптимизации портфеля с меньшим числом параметров, получив вектор долей для кластеров $W = [W_1, \dots, W_k]$. Рассчитать долю каждой ценной бумаги можно по формуле:

$$X_i = \frac{W_j}{S_j},$$

где

j — кластер, в который входит ценная бумага i;

S_j — число ценных бумаг в кластере j.

Из-за того, что кластеры представляют собой объединения ценных бумаг, оптимальный портфель, рассчитанный для кластеров, будет по своим характеристикам хуже, чем оптимальный портфель, рассчитанный для отдельных ценных бумаг. Величина снижения будет тем меньше, чем более схожие по поведению (с высокими коэффициентами парной корреляции) ценные бумаги оказались объединены в кластеры. Следовательно, можно сделать вывод о том, что для успешного применения предложенного метода (т.е. приводящего к достаточно сильному снижению размерности при допустимом снижении качества полученного оптимального решения) требуется, чтобы кластеры в исходных данных выделялись достаточно хорошо.

Возможно, результаты проведённой однажды кластеризации можно будет использовать в течение некоторого времени для решения нескольких задач оптимизации (до тех пор, пока значительно

не изменится матрица парных корреляций). Однако, этот вопрос однозначно требует дополнительного изучения.

Схожий с предложенным метод предлагается в работе [4], однако в упомянутом исследовании не производится расчёта ковариационной матрицы доходностей кластеров, а для оптимизации портфеля по модели Марковица в качестве матрицы σ используется полученная в ходе иерархической кластеризации матрица межкластерных корреляций. Такой подход, с одной стороны, ускоряет проведение расчётов, но с другой — авторами [4] признаётся риск того, что полученная матрица окажется отрицательно определённой (что сделает невозможным дальнейшие вычисления), однако данный риск игнорируется, поскольку он ни разу не реализовался в ходе экспериментов.

Подобная идея высказывается и в работе [5], выполненной в рамках проекта по исследованию оптимизации портфелей, основанной на кластеризации. Однако из-за слишком сильной ориентированности на практику (например, использование для оценки качества полученного оптимального решения метода Шарпа, специфичного для инвестиционных портфелей), результаты [5] сложно использовать для решения других задач оптимизации, кроме оптимизации инвестиционных портфелей.

2. Исследование влияния предлагаемого метода на качество получаемого оптимального решения

Для исследования влияния предлагаемого метода на качество получаемого оптимального решения был использован следующий метод:

Выбирались значения t_i — порогового значения для проведения кластеризации и R_{ic} — ожидаемой доходности. Затем с использованием предложенного метода при $t = t_i$ строился кластеризованный портфель с доходностью $R = R_{ic}$, его риск — V_i. Далее для тех же исходных данных, но без использования предложенного метода, строился портфель, оптимизированный по риску $V = V_i$, его доходность — $R_i u$.

Оптимизация портфелей проводилась по методике, описанной в [3], с использованием *Python* и библиотеки *CVXPY*, для кластеризации применялась библиотека *scipy.cluster*.

Для оценки влияния необходимы два показателя:

- $E = \frac{n}{k}$ — уровень снижения размерности в задаче оптимизации («экономия» размерности задачи)
- $L = R_i u - R_i c$ — снижение оптимизируемого показателя («потеря» оптимизируемого показателя)

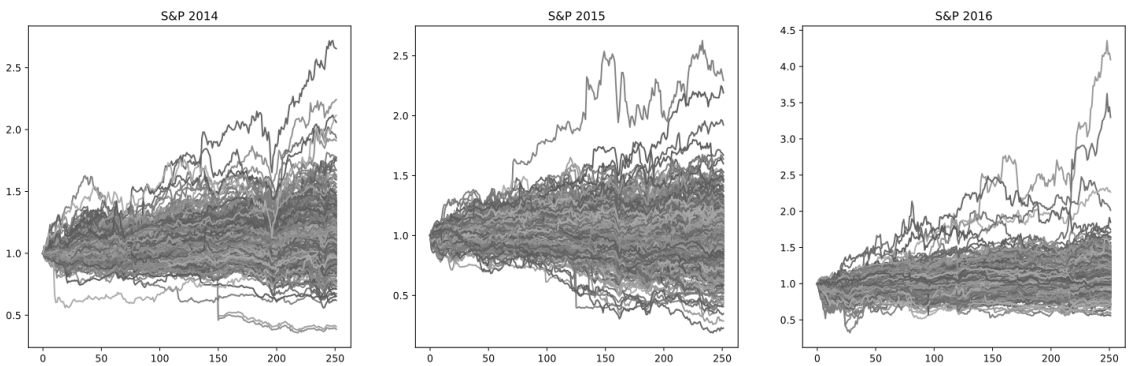

Fig. 1. Changes in prices of stocks of companies from the S&P rating in 2014-2016

Рис. 1. Изменение цен на акции компаний из рейтинга S&P в 2014-2016 годах

Эксперименты проводились на следующих данных:

1. Акции компаний из рейтинга Standard & Poor's 500 за 2014-2016 г.г. (данные об акциях 480, 486 и 494 компаний соответственно). Данные об изменениях цен на акции приведены на рисунке 1 (все данные нормированы).
2. Акции компаний, торгуемые на Нью-Йоркской фондовой бирже (NYSE) в 2004-2006 г.г. (данные об акциях 1285, 1354 и 1421 компаний соответственно). Данные об изменениях цен на акции приведены на рисунке 2 (все данные нормированы).

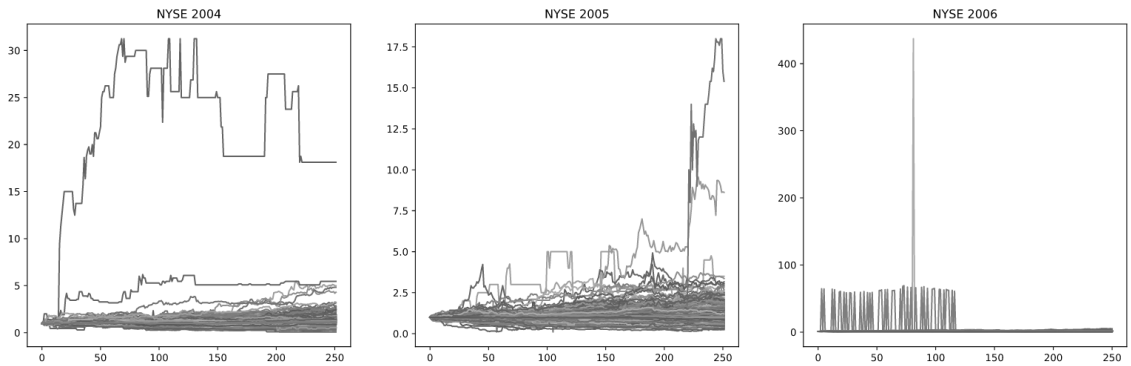

Fig. 2. Changes in prices of stocks of companies traded on the NYSE in 2004-2006

Рис. 2. Изменение цен на акции компаний, торгуемых на Нью-Йоркской фондовой бирже в 2004-2006 годах

3. Акции компаний, торгуемых на бирже NASDAQ в 2004-2006 г.г. (данные об акциях 1143, 1207 и 1280 компаний соответственно). Данные об изменениях цен на акции приведены на рисунке 3 (все данные нормированы).

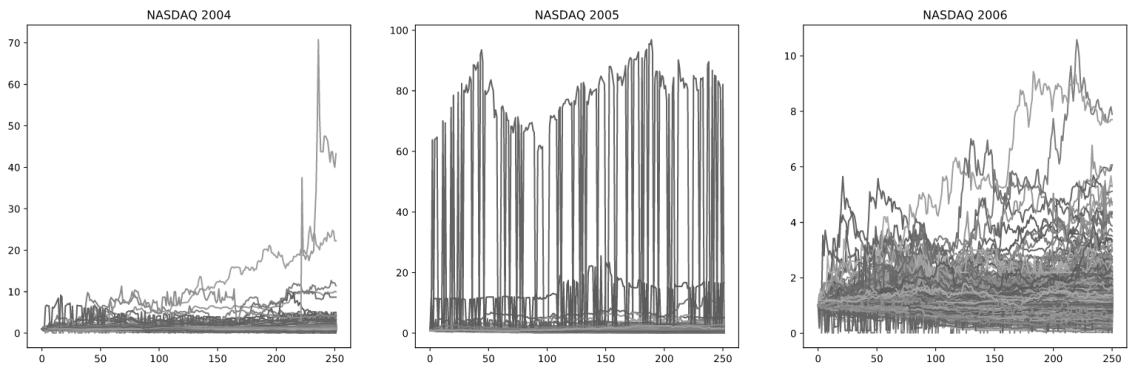

Fig. 3. Changes in prices of stocks of Companies traded on the NASDAQ in 2004-2006

Рис. 3. Изменение цен на акции компаний, торгуемых на бирже NASDAQ в 2004-2006 годах

Для проведения экспериментов использовались три основных метода иерархической агломеративной кластеризации: метод одиночной связи, метод полной связи и метод средней связи.

Число кластеров в зависимости от t для всех трёх наборов данных приведено на рисунках 4, 5, 6.

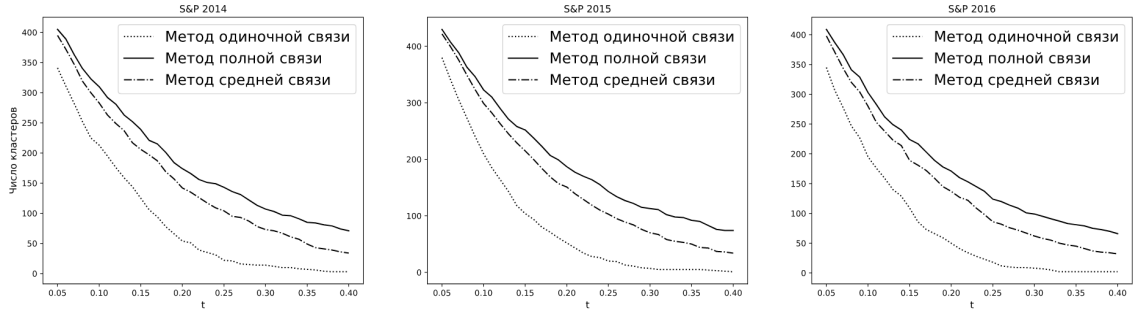

Fig. 4. Dependency between number of clusters and clustering threshold t for stocks of companies from the S&P rating

Рис. 4. Число кластеров в зависимости от порога кластеризации t для акций компаний из рейтинга S&P 500

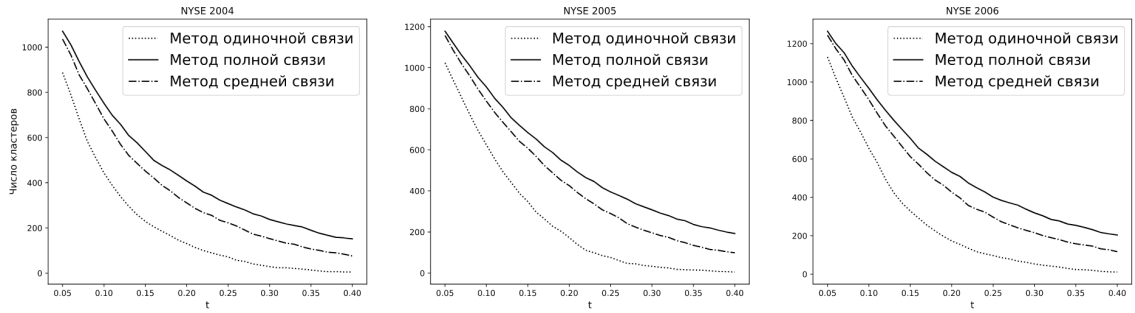

Fig. 5. Dependency between number of clusters and clustering threshold t for stocks of companies traded on the NYSE

Рис. 5. Число кластеров в зависимости от порога кластеризации t для акций компаний, торгуемых на Нью-Йоркской фондовой бирже

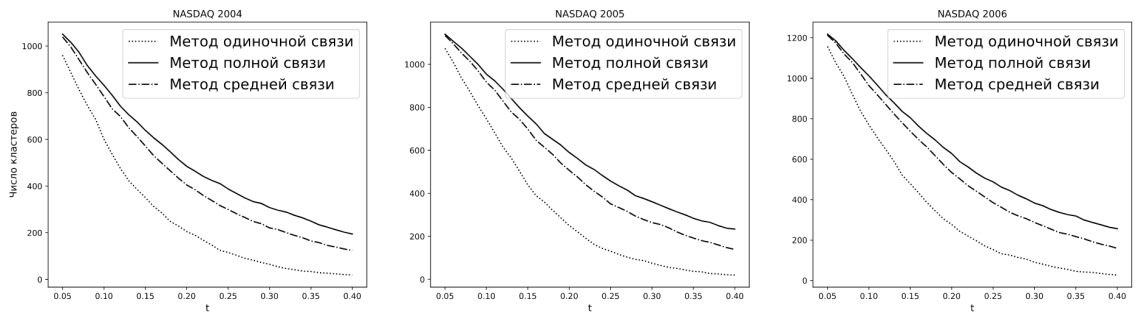

Fig. 6. Dependency between number of clusters and clustering threshold t for stocks of companies traded on the NASDAQ

Рис. 6. Число кластеров в зависимости от порога кластеризации t для акций компаний, торгуемых на бирже NASDAQ

2.1. Влияние предлагаемого метода на качество получаемого оптимального решения при кластеризации по методу одиночной связи

При кластеризации по методу одиночного соседа кластеры во всех трех наборах данных выделяются примерно одинаково, быстро и «гладко». В акциях компаний из рейтинга S&P кластеры выделяются при меньших пороговых значениях, чем в акциях компаний Нью-Йоркской фондовой биржи и акциях компаний биржи NASDAQ, что обусловлено, во-первых, меньшим объёмом данных, во-вторых, тем, что в рейтинг S&P попадают, в первую очередь, крупные компании, цены на акции которых достаточно стабильны и не показывают как значительного, «взрывного» роста, так и сильного снижения.

Средние результаты применения предлагаемого метода при кластеризации по методу одиночной связи приведены в таблице 1 (прочерк означает, что построить кластеризованный оптимальный портфель с заданным параметром доходности не удалось).

Table 1. Average results of the proposed method for clustering using the single linkage method

Таблица 1. Средние результаты применения предлагаемого метода при кластеризации по методу одиночной связи

t_i	R_{ic}	S&P 500 2014-2016		NYSE 2004-2006		NASDAQ 2004-2006	
		E	L	E	L	E	L
1	1,05	1,228	0,002	1,030	0,002	1,009	0,002
	1,1		0,007		0,000		0,000
	1,2		0,009		0,000		0,000
	1,3		0,009		0,000		0,000
2	1,05	2,990	0,018	1,575	0,002	1,403	0,007
	1,1		0,023		0,003		0,005
	1,2		0,036		0,007		0,007
	1,3		0,099		0,016		0,009
3	1,05	15,252	0,087	3,285	0,010	2,711	0,009
	1,1		0,105		0,021		0,015
	1,2		–		0,041		0,021
	1,3		–		0,060		0,027
4	1,05	87,694	0,202	8,399	0,022	7,564	0,038
	1,1		–		0,029		0,049
	1,2		–		0,051		0,074
	1,3		–		0,074		0,101

Как можно видеть, кроме очевидной зависимости E от t, существует ещё несколько зависимостей:

- L возрастает с ростом E при постоянном R_{ic}.
- L, в целом, возрастает с ростом R_{ic} при постоянном E.

Кроме того, можно отметить, что при меньшем объёме данных (акции компаний из рейтинга S&P 500) E растёт с ростом t быстрее, чем при большем объёме данных, и при некоторых значениях t и R_{ic} оптимальный портфель с заданными параметрами после кластеризации построить не удалось.

При значении $t = 1$ применение метода является практически бессмысленным, т.к. оно не приводит к существенному снижению размерности задачи оптимизации, а при $t = 2$ достаточно сильное снижение размерности происходит только для одного набора данных из трёх. В то же время, выбор $t = 4$ приводит к тому, что становится невозможно построить оптимальный портфель.

2.2. Влияние предлагаемого метода на качество получаемого оптимального решения при кластеризации по методу полной связи

Как можно видеть, объединение в кластеры при использовании метода полной связи происходит достаточно неравномерно и медленнее, чем при кластеризации по методу одиночной связи. В то же время, на всех трёх наборах данных кластеры выделяются достаточно хорошо, а для данных об акциях компаний Нью-Йоркской фондовой биржи и биржи NASDAQ — ещё и достаточно стабильно.

Средние результаты применения предлагаемого метода при кластеризации по методу полной связи приведены в таблице 2.

Table 2. Average results of the proposed method for clustering using the complete linkage method

Таблица 2. Средние результаты применения предлагаемого метода при кластеризации по методу полной связи

t_i	R_{ic}	S&P 500 2014-2016		NYSE 2004-2006		NASDAQ 2004-2006	
		E	L	E	L	E	L
1	1,05	1,125	0,002	1,025	0,001	1,009	0,002
	1,1		0,007		0,000		0,001
	1,2		0,009		0,000		0,000
	1,3		0,011		0,000		0,000
2	1,05	1,836	0,010	1,287	0,002	1,203	0,010
	1,1		0,014		0,003		0,010
	1,2		0,034		0,007		0,012
	1,3		0,050		0,014		0,020
3	1,05	3,175	0,026	1,836	0,005	1,713	0,023
	1,1		0,030		0,010		0,026
	1,2		0,043		0,019		0,033
	1,3		0,065		0,033		0,038
4	1,05	5,251	0,042	2,736	0,014	2,500	0,041
	1,1		0,046		0,021		0,046
	1,2		0,064		0,046		0,049
	1,3		0,103		0,099		0,058

Все зависимости, описанные для метода одиночной связи, имеют место и для метода полной связи, с единственным важным отличием — не было таких R_{ic} и t, при которых не получилось бы построить оптимальный портфель. Поскольку объединение в кластеры при использовании метода полной связи происходит медленее, чем при кластеризации по методу одиночной связи, в целом рост E и L при возрастании t и R_{ic} происходит медленнее, чем при использовании метода одиночной связи. Как и в случае с использованием метода одиночной связи, использование метода при $t = 1$ и $t = 2$ не имеет практического смысла, в то же время, наилучшие результаты получаются при выборе $t = 3$ или $t = 4$.

2.3. Влияние предлагаемого метода на качество получаемого оптимального решения при кластеризации по методу средней связи

При кластеризации по методу средней связи объединение в кластеры происходит немного более быстро и «гладко», чем при использовании метода полной связи, но всё же не так быстро, как при использовании метода одиночной связи. Относительно стабильные кластеры формируются только для акций компаний, торгуемых на бирже NASDAQ.

Средние результаты применения предлагаемого метода при кластеризации по методу средней связи приведены в таблице 3.

Table 3. Average results of the proposed method for clustering using the average linkage method

Таблица 3. Средние результаты применения предлагаемого метода при кластеризации по методу средней связи

t_i	R_{ic}	S&P 500 2014-2016		NYSE 2004-2006		NASDAQ 2004-2006	
		E	L	E	L	E	L
1	1,05	1,142	0,002	1,026	0,001	1,009	0,002
	1,1		0,007		0,000		0,000
	1,2		0,009		0,000		0,000
	1,3		0,011		0,000		0,000
2	1,05	2,054	0,011	1,343	0,002	1,233	0,014
	1,1		0,016		0,003		0,012
	1,2		0,030		0,008		0,010
	1,3		0,048		0,016		0,009
3	1,05	4,364	0,032	2,088	0,006	1,884	0,015
	1,1		0,037		0,010		0,019
	1,2		0,057		0,019		0,028
	1,3		0,096		0,033		0,040
4	1,05	9,354	0,065	3,511	0,015	3,173	0,025
	1,1		0,073		0,022		0,031
	1,2		0,099		0,049		0,044
	1,3		0,178		0,100		0,060

При кластеризации по методу средней связи результаты, во многом, обусловлены скоростью объединения акций в кластеры — средней между методами одиночной и полной связей. Как и в предыдущих случаях, использование метода при $t = 1$ лишено практического смысла, а наилучшие результаты получаются, в зависимости от размерности исходных данных, при выборе $t = 3$ или $t = 4$.

Заключение

По результатам описанных экспериментов можно сделать следующие выводы:

1. Предложенный метод позволяет существенно (в 5 раз и более) понижать размерность в задачах оптимизации при умеренном снижении качества полученного оптимального решения (до 5%) при «хорошем» выборе t и метода кластеризации.
2. Метод одиночной связи часто приводит к слишком быстрому понижению размерности, метод полной связи — к слишком медленному, «осторожному»; при использовании метода средней связи скорость будет средней. Выбор конкретного метода зависит, в первую очередь, от особенностей набора данных.

В дальнейшем следует изучить вопрос возможности построения регрессионной зависимости L от R, t и используемого метода кластеризации, чтобы формализовать процедуру выбора «хороших» параметров метода. Также потенциально перспективным является использование для проведения кластеризации ковариационной матрицы вместо матрицы парных корреляций, однако, этот вопрос требует отдельного изучения.

References

[1] H. Markowitz, “Portfolio Selection”, *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[2] V. Dubrovin and O. Os’kiv, “Modeli i metody optimizacii vybora investicionnogo portfelja”, *Radiojelektronika, informatika, upravlenie*, vol. 1, pp. 49–60, 2008.

[3] J. Chaitanya, *Markowitz Portfolio Optimization*, 2017. [Online]. Available: https://chaitjo.github.io/markowitz/.

[4] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna, “Cluster analysis for portfolio optimization”, *Journal of Economic Dynamics and Control*, vol. 32, no. 1, pp. 235–258, 2008.

[5] D. León and et. al., “Clustering algorithms for Risk-Adjusted Portfolio Construction”, in *ICCS*, vol. 108, 2017, pp. 1334–1343.