МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра теории вероятностей и математической статистики

ФОРМИРОВАНИЕ ИНВЕСТИЦИОННОЙ СТРАТЕГИИ НА ОСНОВЕ ПРЕДСКАЗАНИЙ БАЙЕСОВСКОЙ РЕГРЕССИИ

Курсовая работа

Рымкевич Виктории Сергеевны

студентки 4 курса,
специальность «актуарная математика»

Научный руководитель:

кандидат физико-математических наук,

доцент В.А. Морозов

Минск, 2016

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра теории вероятностей и математической статистики

ЗАДАНИЕ НА КУРСОВУЮ РАБОТУ

Студент

1. Тема «Формирование инвестиционной стратегии на основе предсказаний байесовской регрессии»

2. Срок представления курсовой работы 17.05.2016

3. Исходные данные для научного исследования:

Devavrat Shah. Bayesian regression and Bitcoin. / Shah Devavrat, Kang Zhang // Massachusetts Institute of Technology [Electronic resource]. – 2014. – Mode of access: http://arxiv.org/abs/1410.1231. – Date of access: 09.05.2015.

Медведев Г.А., Морозов В.А. Практикум на ЭВМ по анализу временных рядов: Учеб. пособие. – Мн .: Университетское, 2001. – 192 с

Модельные данные и реальные данные ряда экономических показателей.

4. Содержание курсовой работы

4.1. Изучить математические модели временных рядов.
4.2. Изучить методы байесовского оценивания.
4.3. Реализовать алгоритмы предсказания на основе байесовской регрессии.
4.4 Сформировать инвестиционную стратегию и оценить ее эффективность.

Руководитель курсовой работы _____ 12.09.2015г. /В.А. Морозов/
(подпись, дата)

Задание принял к исполнению _____ 12.09.2015г.
(подпись, дата)

Оглавление

Введение

Предсказание финансовых инструментов – необходимый элемент любой инвестиционной деятельности. Сама идея инвестиций – вложения денег сейчас с целью получения дохода в будущем – основывается на идее прогнозирования будущего. Соответственно, предсказание финансовых временных рядов лежит в основе деятельности всей индустрии инвестиций – всех бирж и внебиржевых систем торговли ценными бумагами.

Целью данной работы было разработать алгоритм динамически формируемого портфеля, на основе предсказания цен акций по байесовской регрессии, и применить его на практике. В основу алгоритма легла идея расширения стратегии, разработанной для спекулятивной торговли биткойнами двумя сотрудниками лаборатории искусственного интеллекта при Массачусетском технологическом институте (Massachusetts Institute of Technology – MIT) Девавратом Шах и Каном Чжан[1]. Основные внесенные изменения будут перечислены в разделе 4.

Эффективность стратегии оценивается путем сравнения полученных результатов с результатами следующих инвестиций:

- оптимальный портфель Марковица минимального риска;
- оптимальный портфель Марковица максимальной доходности;
- фондовый индекс S&P 500 (SPX);
- фондовый индекс Dow Jones Industrial Average (DJI);

В разделах 1 и 2 будет рассмотрена основная информация по портфельной теории Марковица и фондовым индексам соответственно. Раздел 3 посвящен описанию алгоритма прогнозирования изменения цен на основе байесовской регрессии. В разделе 4 приведена разработанная торговая стратегия, а в разделе 5 производится сравнение результатов.

1. Портфельная теория Марковица

Гарри Марковиц в 1952 году впервые предложил математическую модель формирования инвестиционного портфеля. В основе его модели лежат два ключевых показателя любого финансового инструмента: доходность и риск, которые были количественно измерены. Доходность по модели представляет собой математическое ожидание доходностей, а риск определяется как разброс доходностей возле математического ожидания и рассчитывается через стандартное отклонение.

До модели Марковица инвестирование происходило, как правило, в выборочные активы или финансовые инструменты, предложенная же им модель позволила снизить систематические (рыночные) риски за счет группировки активов с отрицательной корреляцией доходностей.

Следует заметить универсальность модели, так инвестиционный портфель может быть технически составлен для любых видов финансовых инструментов и активов: акций, облигаций, фьючерсов, индексов, недвижимости и т.д.

Выделяют две инвестиционные стратегии при формировании портфеля:

- стратегия максимизации доходности инвестиционного портфеля при ограниченном уровне риска;
- стратегия минимизации риска инвестиционного портфеля при минимально допустимом уровне доходности.

Расчет доходности. Общая доходность портфеля будут представлять собой взвешенную сумму доходностей каждого отдельного финансового инструмента (актива):

$$r_p = \sum_{i=1}^{n} w_i \cdot r_i;$$

где: n – количество финансовых инструментов инвестиционного портфеля;

r_p – доходность инвестиционного портфеля;

w_i – доля i – го финансового инструмента в портфеле;

r_i – ожидаемая доходность i – го финансового инструмента.

Оценка риска. В модели Марковица риск отдельно взятого финансового инструмента рассчитывается как стандартное отклонение доходностей. Для расчета общего риска портфеля необходимо отразить их

совокупное изменение и взаимное влияние (через ковариацию), для этого воспользуемся следующей формулой:

$$\sigma_p = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i \cdot w_j \cdot V_{ij}} = \sqrt{\sum_{i=1}^{n} w_i^2 \cdot \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i \cdot w_j \cdot k_{ij} \cdot \sigma_i \cdot \sigma_j};$$

где: n – количество финансовых инструментов инвестиционного портфеля;

σ_p – риск инвестиционного портфеля;

σ_i – стандартное отклонение доходностей $i-$го финансового инструмента;

k_{ij} – коэффициент корреляции между i-ым и j-ым финансовыми инструментами;

w_i – доля i-ого финансового инструмента в портфеле;

V_{ij} – ковариация доходностей i-ого и j-ого финансовых инструментов.

Рассматривая теоретически предельный случай, при котором в портфель можно включать бесконечное количество ценных бумаг, дисперсия (мера риска портфеля) асимптотически будет приближаться к среднему значению ковариации.

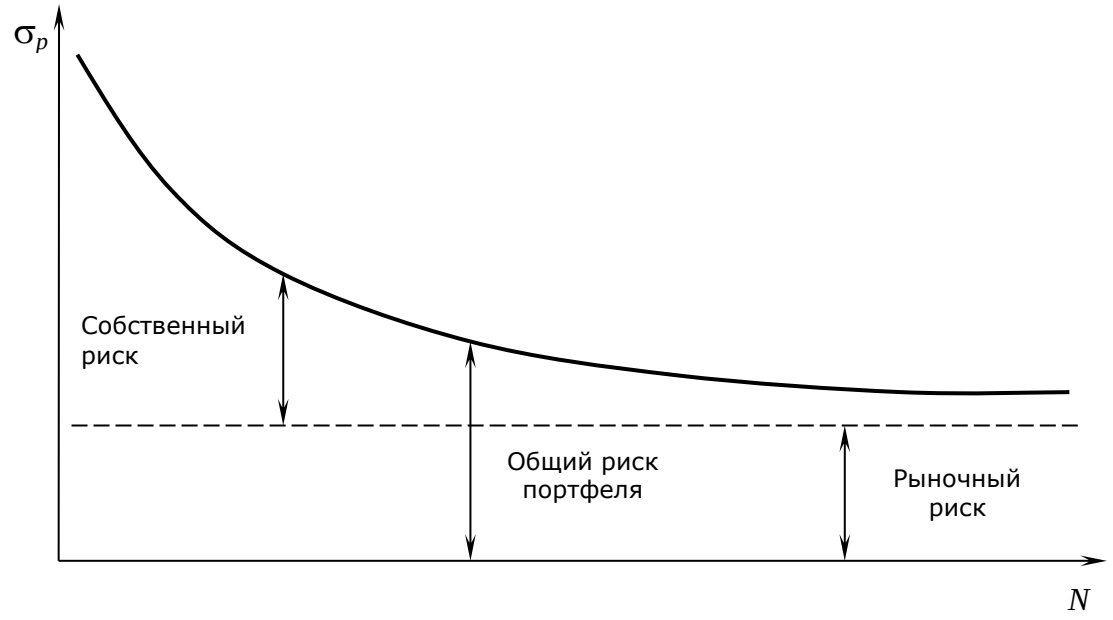

Рисунок 1.1 – Риск портфеля

Совокупный риск портфеля можно разложить на две составные части: рыночный риск, который нельзя исключить и которому подвержены все ценные бумаги практически в равной степени, и собственный риск, который можно избежать при помощи диверсификации. При этом сумма вложенных средств по всем объектам должна быть равна общему объему инвестиционных вложений, т.е. сумма относительных долей в общем объеме должна равняться единице.

Проблема заключается в численном определении относительных долей акций и облигаций в портфеле, которые наиболее выгодны для владельца. Марковиц ограничивает решение модели тем, что из всего множества «допустимых» портфелей, т.е. удовлетворяющих ограничениям, необходимо выделить те, которые рискованнее, чем другие. При помощи разработанного Марковицем метода критических линий можно выделить неперспективные портфели. Тем самым остаются только эффективные портфели.

Отобранные таким образом портфели объединяют в список, содержащий сведения о процентном составе портфеля из отдельных ценных бумаг, а также о доходе и риске портфелей.

Объяснение того факта, что инвестор должен рассмотреть только подмножество возможных портфелей, содержится в следующей теореме об эффективном множестве: «Инвестор выберет свой оптимальный портфель из множества портфелей, каждый из которых обеспечивает максимальную ожидаемую доходность для некоторого уровня риска и минимальный риск для некоторого значения ожидаемой доходности». Набор портфелей, удовлетворяющих этим двум условиям, называется эффективным множеством.

На рисунке 1.2 представлены недопустимые, допустимые и эффективные портфели, а также линия эффективного множества.

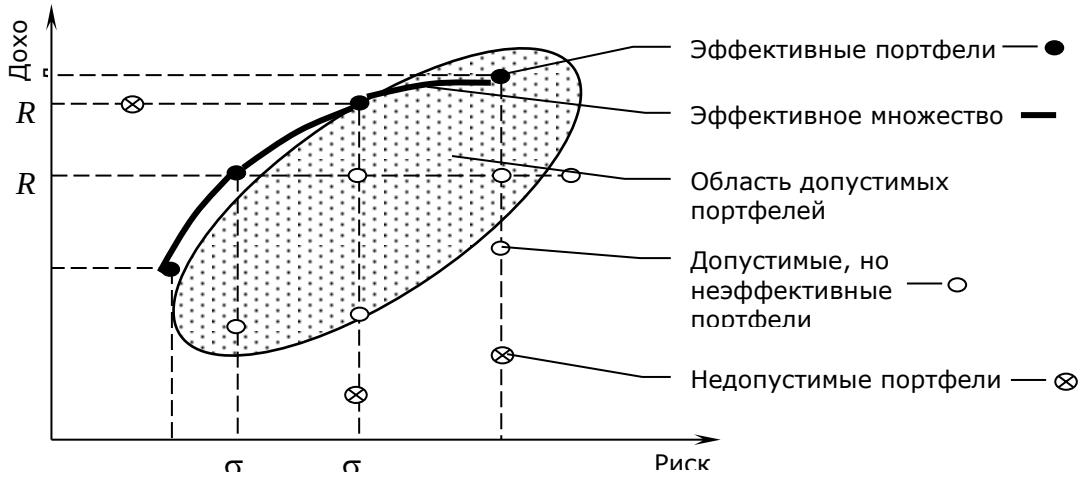

Рисунок 1.2 – Допустимое и эффективное множества

Эконометрический вид модели. Для того чтобы сформировать инвестиционный портфель необходимо решить оптимизационную задачу. Существует два вида задач: поиск долей акций в портфеле для достижения максимальной эффективности при заданном уровне риска (σ_p) и минимизация риска при заданном уровне доходности портфеля (r_p). Помимо этого, на уравнения накладываются дополнительные очевидные ограничения: сумма долей активов должна быть равна 1 и сами доли активов должны быть положительными. В таблице 1.1 показаны формулы и наложенные на них ограничения для поиска оптимальных долей финансовых инструментов:

Таблица 1.1 – Два вида оптимизационных задач

Портфель Марковица минимального риска	Портфель Марковица максимальной эффективности
$\begin{cases} \sqrt{\sum_{i=1}^{n} w_i^2 \cdot \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i \cdot w_j \cdot k_{ij} \cdot \sigma_i \cdot \sigma_j} \to min \\ \sum_{i=1}^{n} w_i \cdot r_i > r_p \\ \sum_{i=1}^{n} w_i = 1 \\ w_i \geq 0 \end{cases}$	$\begin{cases} \sum_{i=1}^{n} w_i \cdot r_i \to max \\ \sqrt{\sum_{i=1}^{n} w_i^2 \cdot \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i \cdot w_j \cdot k_{ij} \cdot \sigma_i \cdot \sigma_j} < \sigma_p \\ \sum_{i=1}^{n} w_i = 1 \\ w_i \geq 0 \end{cases}$

У модели существует ряд недостатков, на которые следует обратить внимание:

1. Модель строится на основании средней доходности по данным прошлых периодов, поэтому ее рационально использовать только при стабильном состоянии фондового рынка.

2. Не всегда выполняется предпосылка о нормальном распределении доходности, следовательно, математическое ожидание и среднеквадратичное отклонение не могут служить адекватными мерами доходности и риска.

3. Существуют ситуации, когда кратность стоимости активов приводит к вынужденному добавлению целочисленных ограничений. Это ведет к росту размерности задачи и трудоемкости ее решения.

2. Фондовые индексы

Фондовый индекс - показатель состояния и динамики рынка ценных бумаг. Через сопоставление текущего значения индекса с его предыдущими значениями можно оценить поведение рынка, его реакцию на те или иные изменения макроэкономической ситуации, различные корпоративные события (слияния, поглощения, дробления акций, отставки и назначения ведущих менеджеров), спекулятивные процессы.

В зависимости от того, какие ценные бумаги составляют выборку, используемую при расчете индекса, он может характеризовать рынок в целом, рынок определенного класса ценных бумаг (государственные обязательства, корпоративные облигации, акции и т. п.), отраслевой рынок (ценные бумаги компаний одной отрасли: телекоммуникации, транспорт, страхование, Интернет-сектор и т. п.). Сравнение динамики различных индексов может показать, какие сектора экономики развиваются самыми быстрыми темпами. Индекс может представлять национальный фондовый рынок в целом или определенную торговую площадку на этом рынке (например, индекс фондовой биржи). Фондовые индексы рассчитываются и публикуются различными организациями, чаще всего информационными или рейтинговыми агентствами и фондовыми биржами.

Методика определения фондовых индексов. Чтобы фондовый индекс адекватно отражал процессы, происходящие на рынке ценных бумаг, и как можно меньше зависел от субъективных факторов, таких, как манипулирование ценами отдельных финансовых инструментов, корпоративная политика компаний-эмитентов, включающая новые эмиссии, дробление или консолидацию акций, выпуск варрантов и т.п., необходимо применять правильные и обоснованные методики расчета фондовых индексов. Кроме того, понимание методики расчета индекса необходимо для правильной интерпретации его изменений.

При определении методики вычисления фондовых индексов необходимо рассмотреть следующие вопросы:

• формулы вычисления фондовых индексов;

• достоверность и полнота информации, используемой при расчете фондовых индексов;

• порядок корректировки расчетной формулы, необходимость которой вызвана теми или иными корпоративными событиями, изменением рыночных условий.

Методы расчета фондовых индексов. Существует четыре основных метода расчета фондовых индексов:

1. Метод вычисления невзвешенного среднего арифметического. Эта формула используется при расчете среднего промышленного индекса Доу-Джонса (Dow Jones Industrial Average).

2. Метод вычисления взвешенного среднего арифметического с использованием различных способов взвешивания:

- взвешивание по цене акций в выборке;

- взвешивание по стоимости выборки;

- взвешивание путем приравнивания весов акций компаний;

Данная методика используется для вычисления среднего индекса рейтингового агентства Standard & Poor's (S&P 500).

3. Метод вычисления невзвешенного среднего геометрического. По этой формуле рассчитывается старейший фондовый индекс Великобритании ФТ-30 (FT-30 Share Index, Financial Times Industrial Ordinary Index), который стал публиковаться с 1935 г.

4. Метод вычисления взвешенного среднего геометрического. Эта формула применяется для расчета композитного индекса Value Line Composite Average, используемого на фондом рынке США.

Требования к информации, используемой при вычислении фондовых индексов.

Любая формула будет бесполезна, если в нее будут вводиться недостоверные или неполные данные. Для обоснованного использования в расчетах информация должна отвечать следующим критериям:

• размер выборки. Желательно использовать при расчете индекса достаточно большое число компаний, что позволяет уменьшить вероятность влияния на конечный результат случайных отклонений стоимости ценных бумаг отдельных компаний относительно среднего рыночного значения.

• репрезентативность выборки. Перечень компаний, ценные бумаги которых входят в состав, например, отраслевого индекса, должен быть достаточно полным для того, чтобы индекс адекватно отражал состояние определенного сегмента экономики. Кроме того, чтобы изменения индекса правильно отражали изменения, происходящие на рынке, распределение эмитентов по размеру капитализации и отраслевой принадлежности должно соответствовать распределению на рынке в целом. Использование компьютеров позволило начать расчет индекса по всем акциям, торгуемым на том или ином рынке, не прибегая к некоторой выборке.

• вес. Желательно, чтобы стоимость ценных бумаг, входящих в индекс, имела свой вес, пропорциональный их влиянию на фондовый рынок в целом.

• объективность финансовой информации. Следует учитывать, что фондовый индекс рассчитывается на основе открыто сообщаемых сведений об изменении цен на финансовые инструменты. Большинство индексов рассчитывается в течение торгового дня, причем их обновленные значения появляются через короткие промежутки времени.

Корректировка индексов. Методика расчета индекса может время от времени меняться, что связано главным образом с различными корпоративными событиями, переживаемыми компаниями, ценные бумаги которых входят в состав индекса. Изменения могут касаться и перечня ценных бумаг, участвующих в расчете индекса.

Чем большую историю имеет фондовый индекс, тем большую ценность он представляет для прогнозирования будущей реакции рынка на те или иные события на основе его прошлого поведения. Но ситуация на рынке постоянно меняется - слияния и поглощение, банкротства старых компаний и появление новых, стремительно наращивающих свою капитализацию. Поэтому периодически появляется необходимость внести изменения в выборку, на основе которой рассчитывается индекс.

Если такие корректировки осуществлять редко, есть опасность, что индекс начнет отставать от развития рынка, если к корректировкам прибегать слишком часто - индекс начнет "терять" историю и, сохраняя прежнее название, отражать изменения уже другого сектора рынка.

2.1 Индексы Доу-Джонса

Наибольшей известностью в данном семействе индексов пользуется Dow Jones Industrial Average (средний промышленный индекс Доу-Джонса). Этот индекс был впервые опубликован в 1884 г. Чарльзом Доу, основателем компании, которая была издателем известной финансовой газеты "Wall Street Journal". Этот индекс сначала рассчитывался по акциям 11 железнодорожных компаний. В 1897 г. список был увеличен до 20 железнодорожных компаний. Первый промышленный индекс Доу-Джонса был рассчитан в 1896г. по акциям 12 компаний. В 1916 г. размер выборки был увеличен до 20 компаний, а в 1928г. - до 30. Последнее изменение в составе индекса было произведено 1 ноября 1999г., когда вместо компаний Union Carbide, Goodyear Tire & Rubber, Sears и Chevron в индекс были включены компании Home Depot, Intel, Microsoft и SBC Communications.

Индекс рассчитывается как среднее арифметическое цен акций 30 крупнейших компаний. В качестве делителя используется не число 30 (число компаний в выборке), а специальный делитель, учитывающий многочисленные сплиты (дробления акций), произведенные компаниями-эмитентами с 1928г. (с момента увеличения выборки до 30 компаний).

Используются и другие индексы Доу-Джонса: взвешенный индекс акций Доу-Джонса, рассчитанный по 700 акциям, котируемых на Нью-Йоркской фондовой бирже (публикуется с 1988 г.), индексы Доу-Джонса по транспортным и коммунальным компаниям (Dow Jones Transportation Average (20), Dow Jones Utilities Average(15)) и по 40 облигациям.

AMEX Composite - взвешенный по рыночной капитализации индекс всех акций, торгуемых на Американской фондовой бирже (American Stock Exchange).

NASDAQ 100 - индекс 100 крупнейших компаний нефинансового сектора на бирже NASDAQ.

NASDAQ Composite - взвешенный по капитализации индекс внебиржевого рынка, ежедневно публикуемый Национальной Ассоциацией фондовых дилеров и охватывающий около 3500 акций, торгуемых в рыночной системе Nasdaq (Nasdaq Market System).

NYSE Composite - взвешенный по рыночной капитализации индекс всех акций, торгуемых на Нью-Йоркской фондовой бирже (NYSE).

2.2 Семейство индексов Standard & Poor's

Standard & Poor's Composite 500 Index. В состав индекса входят 400 индустриальных, 20 транспортных, 40 коммунальных и 40 финансовых компаний. Взвешен по рыночной капитализации. Охватывает примерно 80% общей капитализации компаний, торгуемых на Нью-Йоркской фондовой бирже. Капитализация компаний в выборке составляет от 73 миллионов до 75 миллиардов долларов.

Standard & Poor's 400 Index (S&P Midcap) аналогичен S&P 500, но охватывает 400 промышленных компаний, капитализация которых варьируется от 85 миллионов до 6.8 миллиардов долларов. Standard & Poor's 100 аналогичен S&P 500, но охватывает только 100 акций, на которые существуют опционные контракты на Чикагской бирже опционов. "OEX" - название опциона на данный индекс, являющегося один из самых популярных и торгуемых опционов.

Взвешенный по цене индекс как противоположность индексу, взвешенному по капитализации. Некоторые считают, что данный индекс дает лучшее представление об эффективности инвестиций, так как отдельные акции не перешивают в нем, и большинство индивидуальных инвесторов не строят свой портфель с взвешиванием по рыночной капитализации. (пока они не покупают индексные фонды).

3. Прогнозирование

3.1 Байесовская регрессия. Классический подход

Рассмотрим следующую задачу. Пусть дан ряд из n исторических маркированных точек данных (x_i, y_i), для $1 \leq i \leq n$, где $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, при некотором фиксированном $d \geq 1$. Необходимо, используя эти исторические данные, предсказать неизвестный маркер $y \in \mathbb{R}$ для данного $x \in \mathbb{R}^d$.

Классический подход решения данной задачи предполагает для генерации маркера y использование модели следующего типа:

$$y = f(x) + \epsilon, \tag{3.1}$$

где ϵ – независимая случайная величина, представляющая шум. Обычно предполагается, что она является стандартной Гауссовской величиной с математическим ожиданием 0 и дисперсией 1.

Регрессионный метод сводится к оцениванию функции f по n наблюдениям $(x_1, y_1), \dots, (x_n, y_n)$ и использованию ее для прогнозирования будущего.

Например, если f – линейная функция, т.е. $f(x) = x^T\theta^*$, то для оценивания θ^* можно применить классический метод наименьших квадратов:

$$\hat{\theta}_{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^T\theta)^2 \tag{3.2}$$

В классической постановке, d предполагается фиксированным и $n \gg d$, что оправдывает такую высокую эффективность оценки. В различных современных приложениях, более реалистично $n \approx d$ или даже $n \ll d$, и, таким образом, это оставляет весьма неопределенную проблему оценивания θ^*. В этом случае можно сделать предположение о «разрежённости» θ^*, т.е.:

$$||\theta^*||_0 \ll d, \text{где } ||\theta^*||_0 = |\{i: \theta_i^* \neq 0\}| \tag{3.3}$$

Правильным решением будет использование регулярной оценки наименьших квадратов (так же известной как оценка Лассо [4]) для соответствующего выбора $\lambda > 0$:

$$\hat{\theta}_{LASSO} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^T\theta)^2 + \lambda\|\theta\|_1 \tag{3.4}$$

Стоит отметить, что вышеупомянутая конструкция, с различными функциональными формами, была чрезвычайно успешен на практике. На данный момент, это очень активная область исследований.

Ключ к успеху для вышеуказанного подхода заключается в возможности выбирать разумное параметрического пространство, среди которого пытаются оценивать параметры, используя наблюдения. В различных современных приложениях сделать такой выбор кажется сложным. Основная причина этого, заключается в том, что данные очень высокой размерности (например, временные ряды) делают параметрическое пространство либо слишком сложным, либо бессмысленным. Сейчас во многих таких случаях кажется, что есть лишь несколько заметных вариантов, в которых основное событие проявляет себя. Например, фраза или набор слов становятся популярными в социальной сети Twitter по немногим различным причинам: публичное мероприятие, событие, меняющее жизнь знаменитости, природная катастрофа и т.п. Точно так же, есть только несколько различных типов людей с точки их выбора фильмов: те, кто любит комедии и мелодрамы, те, кто любит боевики и фильмы ужасов, и т.д. Таковы были новые идеи, оформленные в работах [2] и [3] как модели скрытого источника. Модель скрытого источника формально описана ниже в контексте рассматриваемой структуры.

3.2 Модель скрытого источника в рассматриваемом контексте

Рассмотрим следующую задачу. Пусть существуют:

1) K различных скрытых источников $s_1, \dots, s_K \in \mathbb{R}^d$;
2) скрытое распределение над $\{1, \dots, K\}$ с соответствующими вероятностями $\{\mu_1, \dots, \mu_K\}$;
3) K распределений над $\mathbb{R}$, обозначенных $P_1, \dots, P_K$.

Необходимо получить данные, удовлетворяющие данной модели.

Каждая маркированная точка данных (x, y) генерируется следующим образом:

1) задается индекс $T \in \{1, \dots K\}$ так, что $P(T = k) = \mu_k$ для $1 \leq k \leq K$;
2) $\boldsymbol{x} = s_T + \epsilon$, где ϵ — d-мерная независимая случайная величина, обозначающая шум, которым в нашем случае предполагается Гауссовским с вектором математического ожидания $\mathbf{0} = (0, \dots, 0) \in \mathbb{R}^d$ и единичной матрицей ковариации.
3) $\boldsymbol{y}$ находится из $\mathbb{R}$ по распределению P_T.

3.3 Байесовская регрессия для модели скрытого источника

Учитывая вышеописанную модель, для предсказания маркера $\boldsymbol{y}$ с учетом соответствующего наблюдения $\boldsymbol{x}$, мы можем использовать условное распределение[1] $\boldsymbol{y}$ при заданном $\boldsymbol{x}$, которое задается следующим образом:

$$\begin{aligned} & P(y|x) \\ &= \sum_{k=1}^{T} P(y|x, T=k) P(T=k|x) \\ &\propto \sum_{k=1}^{T} P(y|x, T=k) P(x|T=k)\, P(T=k) \\ &= \sum_{k=1}^{T} P_k(y) P\big(\epsilon = (x - s_k)\big) \mu_k \\ &= \sum_{k=1}^{T} P_k(y) exp\left(-\frac{1}{2}\|x - s_k\|_2^2\right) \mu_k \end{aligned} \tag{3.5}$$

Таким образом, на основании модели скрытого источника, задача предсказания сводится к решению простой задачи Байесовской регрессии. Однако, остаются неизвестными «скрытые» параметры модели. В частности, неизвестны K, источники $s_1, \dots, s_K$, вероятности $\{\mu_1, \dots, \mu_K\}$ и распределения $P_1, \dots, P_K$.

Для преодоления данной проблемы, авторы предлагают следующий простой алгоритм: использовать эмпирические данные как основу для оценки условного распределения $P(y|x)$, приведенного в (3.5). В частности, при заданных n точек наблюдений (x_i, y_i), $1 \le i \le n$ эмпирическая условная вероятность равна:

$$P_{\text{эмп.}}(y|x) = \frac{\sum_{i=1}^{n} \mathbb{1}(y = y_i) exp\left(-\frac{1}{4}\|x - x_i\|_2^2\right)}{\sum_{i=1}^{n} exp\left(-\frac{1}{4}\|x - x_i\|_2^2\right)} \tag{3.6}$$

Предложенная эмпирическая оценка (3.6) имеет следующие применения. В контексте двоичной классификации, т.е. когда $\boldsymbol{y}$ принимает значение из множества $\{0, 1\}$, для установления значения $\boldsymbol{y}$ требуется посчитать следующий коэффициент:

[1] Здесь предполагается, что случайные величины имеют четко определенные плотности в соответствующем пространстве. И когда это уместно, условные вероятности эффективно представляют условную плотность вероятности.

$$\frac{P_{\text{эмп.}}(y=1|x)}{P_{\text{эмп.}}(y=0|x)} = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i=1) exp\left(-\frac{1}{4}\|x-x_i\|_2^2\right)}{\sum_{i=1}^{n} \mathbb{1}(y_i=0) exp\left(-\frac{1}{4}\|x-x_i\|_2^2\right)} \quad (3.7)$$

Если коэффициент больше 1, то $y=1$, а иначе $y=0$.

В общем случае, для оценки условного математического ожидания $\boldsymbol{y}$ при условии наблюдения $\boldsymbol{x}$, из (3.6) вытекает следующая формула:

$$E_{\text{эмп.}}[y|x] = \frac{\sum_{i=1}^{n} y_i \exp\left(-\frac{1}{4}\|x-x_i\|_2^2\right)}{\sum_{i=1}^{n} \exp\left(-\frac{1}{4}\|x-x_i\|_2^2\right)} \quad (3.8)$$

Оценку (3.8) можно рассматривать как линейную. Пусть вектор $X(x) \in \mathbb{R}^d$ такой, что:

$$X(x)_i = \frac{\exp\left(-\frac{1}{4}\|x-x_i\|_2^2\right)}{\sum_{i=1}^{n} \exp\left(-\frac{1}{4}\|x-x_i\|_2^2\right)} \quad (3.9)$$

И пусть вектор $y \in \mathbb{R}^n$ – вектор с -ой компонентой y_i. Тогда $\hat{y} \equiv E_{\text{набл.}}[y|x]$, исходя из формул (3.8) и (3.9), равен:

$$\hat{y} = X(x)y \quad (3.10)$$

Эта оценка (3.10) используется в дальнейшем для предсказания будущей вариации цены.

3.4 Предшествующие исследования

Начать стоит с того, что Байесовский вывод является основополагающим методом и использование эмпирических данных как основу было хорошо известным подходом, который был потенциально обнаружен и повторно открыт в различных контекстах в течение десятилетий, если не столетий.

Использование модели скрытого источника с целью определения точной выборки для Байесовской регрессии впервые было исследовано в [2]. В работе [2], авторы показали эффективность такого подхода для прогнозирования тенденций в социальной сети Twitter. С целью конкретного применения, авторам пришлось использовать модель шума, которая отличалась от Гауссовской, что привело к небольшим изменениям в (3.6) – вместо использования квадратичной функции, была взята квадратичная функция, примененная к логарифму лежащих в основе векторов.

В различных современных приложениях, таких как онлайн рекомендаций, наблюдения (x_i в обозначениях выше) только частично

наблюдаются. Это требует дальнейшей модификации вывода (3.6), чтобы сделать его эффективным. Такая модификация была предложена в [3] вместе с соответствующими теоретическими гарантиями для полученной выборки.

Стоит заметить, что в обоих работах [2] и [3], Байесовская регрессия для модели скрытого источника была использована в первую очередь для бинарной классификации. Вместо этого, в данной работе [1], авторы использовали ее для оценки вещественной величины.

4. Разработка торговой стратегии

Как уже упоминалось ранее, для прогнозирования будущих изменений цен была использована байесовская регрессия на основе модели скрытого источника. На основе величины предполагаемого изменения цены, принималось решение о покупке либо продаже определенного количества акций.

В алгоритм из статьи [1] были внесены следующие основные изменения:

1. рассматривается больший период выборки, и используются дневные, а не 10секундные данные;
2. весь временной ряд делится на два, а не три периода, и все оценки производятся на первом;
3. количество переменных регрессии и размер предшествующих интервалов не фиксируется, а выбирается за счет вида данных;
4. вместо отношения объемов спроса к предложению использовался относительный объем сделок по акции;
5. пороговое значение изменения цены не фиксируется, а также измеряется как параметр на первом периоде;
6. количество покупаемых либо продаваемых акций не фиксируется, а определяется как целая часть отношения прогнозируемого изменения к пороговому (единственным ограничением является наличие у нас необходимых ресурсов);

Ниже приведены подробности.

4.1 Актуальность применения модели скрытого источника

Количественные торговые стратегии были тщательно изучены и применены в финансовой отрасли, хотя многие из них и держат в секрете. Один из наиболее распространенных подходов, сообщенный в литературе, — это технический анализ, который предполагает, что ценовые движения следуют набору шаблонов и можно использовать прошлые движения цен, чтобы в некоторой степени предсказать будущие [5]. Исследования нашли, что некоторые эмпирически развитые геометрические шаблоны, такие как «голова-и-плечи», «треугольник» и «дважды-сверху», «дважды-снизу» (см. рисунок 4.1), могут быть использована для прогнозирования будущего изменения цены [6], [7].

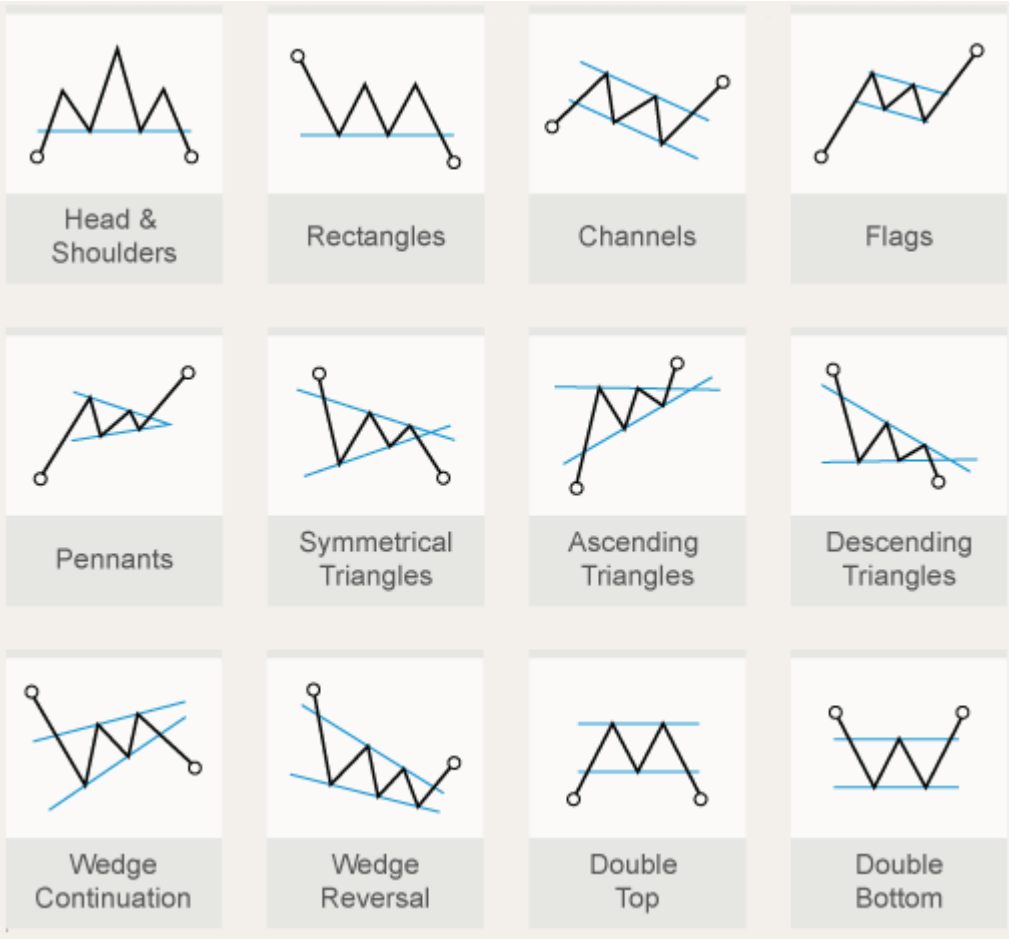

Рисунок 4.1 – Основные геометрические шаблоны

Модель скрытого источника пытается смоделировать существование таких основных шаблонов, ведущих к изменению цены. Попытки разработать модели с помощь человеческого опыта или попытки определить модели в явном виде могут быть сложными и в какой-то степени субъективными. Вместо этого, использование подхода Байесовской регрессии, как описано выше, позволяет нам использовать существование моделей, с целью лучшего прогноза, при этом явно не находя их.

4.2 Предсказание изменения цены

Напомню, что имеющиеся данные разделяются на два периода. Первый период используется для поиска шаблонов и нахождения параметров, а второй – для теоретической торговли и оценки эффективности построенного алгоритма. Далее изложено более подробное описание.

Для конечной оценки изменения цены $\Delta \boldsymbol{p}$ в заданной точке использовалась следующая формула:

$$\Delta p = w_0 + \sum_{j=1}^{n} w_j \Delta p^j + w_{n+1} r \tag{4.1}$$

где r – коэффициент, показывающий относительную величину объема сделок за день, либо наоборот; Δp^j – среднее изменение цены в некотором предшествующем интервале; n – количество переменных регрессии; параметры $w_1, \dots, w_n$ выражают зависимость последующего изменения цены от текущих данных, а параметр w_0 представляет шум.

Коэффициент r рассчитывается по следующей формуле:

$$r = \frac{V}{V_{average}}, \tag{4.2}$$

где V представляет общий объем сделок по акции за день, а $V_{average}$ – средний объем сделок, оцениваемый на первом периоде данных.

Для поиска Δp^j используются исторические данные из задаваемых длин интервалов, обозначаемые x^1, x^2,…, x^n. Вектор x^j используется с историческими шаблонами S^j для предсказания среднего изменения цены Δp^j при помощи байесовской регрессии (3.10) для $1 \le j \le n$.

Поиск исторических шаблонов $S^j, 1 \le j \le n$, производился на первом периоде. Для этого были взяты всевозможные временные ряды соответствующей задаваемой длины. Для каждого ряда x_i (в обозначениях, используемых в (3.10)) соответствующее значение y_i полагается равным среднему изменению цен в задаваемом количестве последних изменений в x_i. Однако, такое количество шаблонов слишком велико. Поэтому из всевозможных пар (x_i, y_i) было выбрано при помощи алгоритма k-средних по несколько кластеров для каждого множества $S^j, 1 \le j \le n$, в нормализированной форме (с математическим ожиданием 0 и дисперсией 1).

Параметры $w_0, \dots, w_{n+1}$ определяются опять же на первом периоде с помощью метода наименьших квадратов (3.2), т.к. количество наблюдений предполагается много большим размерности $d = n + 2$.

Также, для более быстрого вычисления, вместо нормы пространства l_2 в (3.10) было использовано *сходство*. Сходство между двумя векторами $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^M$ определяется следующим образом:

$$s(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_{z=1}^{M} (a_z - mean(\boldsymbol{a}))(b_z - mean(\boldsymbol{b}))}{M\, std(\boldsymbol{a}) std(\boldsymbol{b})} \tag{4.3}$$

где $mean(\boldsymbol{a}) = (\sum_{z=1}^{M} a_z)/M$ и $std(\boldsymbol{a}) = (\sum_{z=1}^{M} (a_z - mean(a))^2)/M$ (аналогично для $\boldsymbol{b}$).

Таким образом, в (3.10) вместо $\exp\left(-\frac{1}{4}\|x - x_i\|_2^2\right)$ используем $\exp(c \cdot s(x, x_i))$. Постоянная $\boldsymbol{c}$ определяется вместе с параметрами $w_0, \ldots, w_4$.

4.3 Торговая стратегия

Разработанная стратегия заключается в следующем:

1) Задается начальный капитал и выбирается конечный набор активов, которые будут участвовать в торговле. (т.е. на начальный момент ни один актив не приобретается).

2) Активы располагаются в порядке приоритета операций с их участием, опираясь на личные предпочтения.

3) На каждый момент времени прогнозируем среднее движение цены ($\Delta\boldsymbol{p}$) в последующем интервале, используя байесовскую регрессию (точные детали описаны в предыдущем подразделе).

4) По очереди принимается решение о покупке/продаже некоторого количества каждого вида акций:

   - если $\Delta\boldsymbol{p} > \boldsymbol{t}$ и текущий капитал $\geq$ текущей цены акции, то **покупаем** $[\Delta\boldsymbol{p}/\boldsymbol{t}]$ акций (либо меньше, насколько позволяет текущий капитал);

   - если $\Delta\boldsymbol{p} < -\boldsymbol{t}$ и текущее количество акций ≥ 1, то **продаем** $[-\Delta\boldsymbol{p}/\boldsymbol{t}]$ акций (либо меньше, смотря сколько акций данного типа у нас есть в наличии);

   - в противном случае ничего не делаем.

В конце торгового периода, продаем все имеющиеся акции и оцениваем эффективность алгоритма.

5. Результаты практической реализации

5.1 Исходные данные

Для практического рассмотрения были взяты дневные данные о ценах и объемах торговли некоторых акций, торгуемых на Нью-Йоркской фондовой бирже (New York Stock Exchange, NYSE), с сайта *nyse.com* за 2013 - 2015 года. В таблице 5.1 приведены сведения о выбранных акциях и выпустивших их компаниях.

Таблица 5.1 – Сведения о рассматриваемых акциях

№	Тикер	Название	Отрасль	Капитализация
1	DDD	3D Systems Corporation	Программное обеспечение	1.05×10^9
2	IBM	International Business Machines Corp	Компьютерные системы	1.3367×10^{11}
3	TWX	Time Warner Inc	Средства массовой информации	5.173×10^{10}
4	CCE	Coca – Cola Enterprises Inc	Пищевая промышленность (напитки)	1.135×10^{10}
5	JNJ	Johnson & Johnson	Фармацевтика	2.8812×10^{11}
6	XOM	Exxon Mobil Corporation	Нефтяная промышленность	3.3066×10^{11}

Заметим, что выбирались компании с большой, а не малой или средней, капитализаций. Также большинство из них являются лидерами в своих отраслях. На рисунке 5.1 представлены графики изменения цен акций за указанный период.

Весь период выборки составил три года, и был разделен на две части. Оценка параметров модели каждого актива производилась отдельно на данных за первые два года (2013 - 2014). Затем, с использованием полученных значений коэффициентов, производилась имитация торговли на данных последнего года (2015).

Как видно на рисунке 5.1, в большинстве случаев цены акций на последнем периоде падают, что осложняет возможность получения большого дохода. Однако, как будет показано в подразделе 5.3, разработанная стратегия хотя бы позволяет сократить убытки. Эффективность стратегии производилась путем сравнения в конце периода торговли результата нашей стратегии с результатами владения портфелями, построенными согласно портфельной теории Марковица.

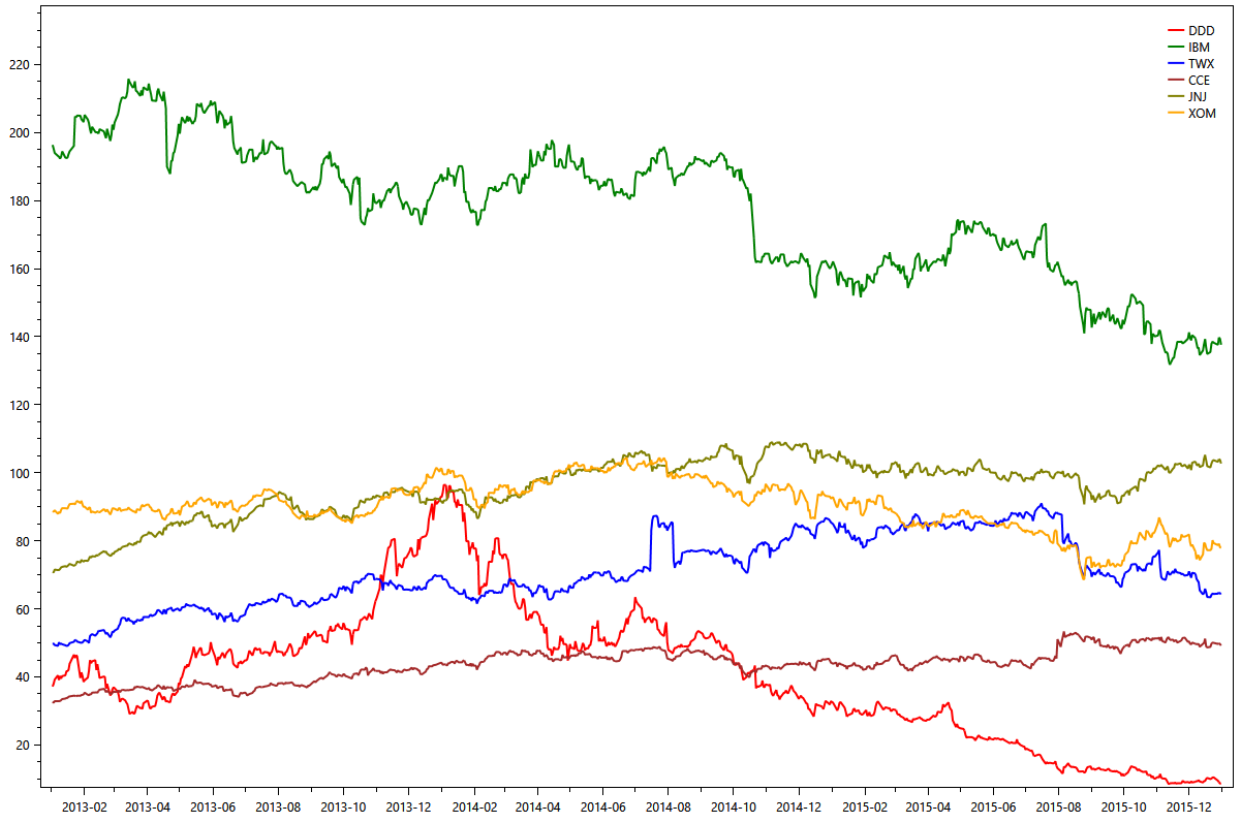

Рисунок 5.1 – Графики цен акций

5.2 Оценка параметров и прогнозирование цены

Было решено использовать три регрессионные переменные (т.к. предыдущие исследования говорят о том, что для большинства моделей этого достаточно), использующие данные о предыдущих интервалах длины 20, 40 и 80 дней соответственно.

На оценочном периоде в подмножествах каждой длины были найдены по 10 наиболее эффективных шаблонов классическим методом k-средних.

Далее методом наименьших квадратов были получены, и использованы в дальнейшем для предсказания изменения цены в (3.1), следующие значения параметров регрессии:

Таблица 5.2 – Оцененные параметры моделей акций

	TWX	JNJ	CCE	XOM	DDD	IBM
Среднеквадратичная ошибка аппроксимации	1.11333	0.83148	0.47688	0.88627	1.78506	1.89823
w_0	0.06793	0.03934	-0.07795	-0.15044	0.06460	-4.73667
w_1	0.00381	0.03468	0.00327	-0.13481	-0.28486	2.34746

w_2	-0.38539	0.00017	0.00925	0.13164	-1.74241	-1.79345
w_3	0.30332	-0.24034	-0.08113	-1.27391	-0.91080	-14.57292
w_4	0.10693	0.00298	0.10674	0.10847	-0.04477	-0.27842
Средний объем сделок	5369534.6	7981670.6	2218128.5	12082829.2	4403237.2	4492401.7
Пороговое значение	0.18679	0.06620	0.03265	0.0324	0.07417	0.13393

На рисунке 5.2 для наглядной демонстрации результатов аппроксимации изображены графики реального и предсказанного изменения цены акций DDD за первый период, т.е. 2013-2014 года. Из рисунка видно, что прогнозирование было проведено достаточно точно, т.к. в каждый момент времени использовались последние наблюдаемые данные.

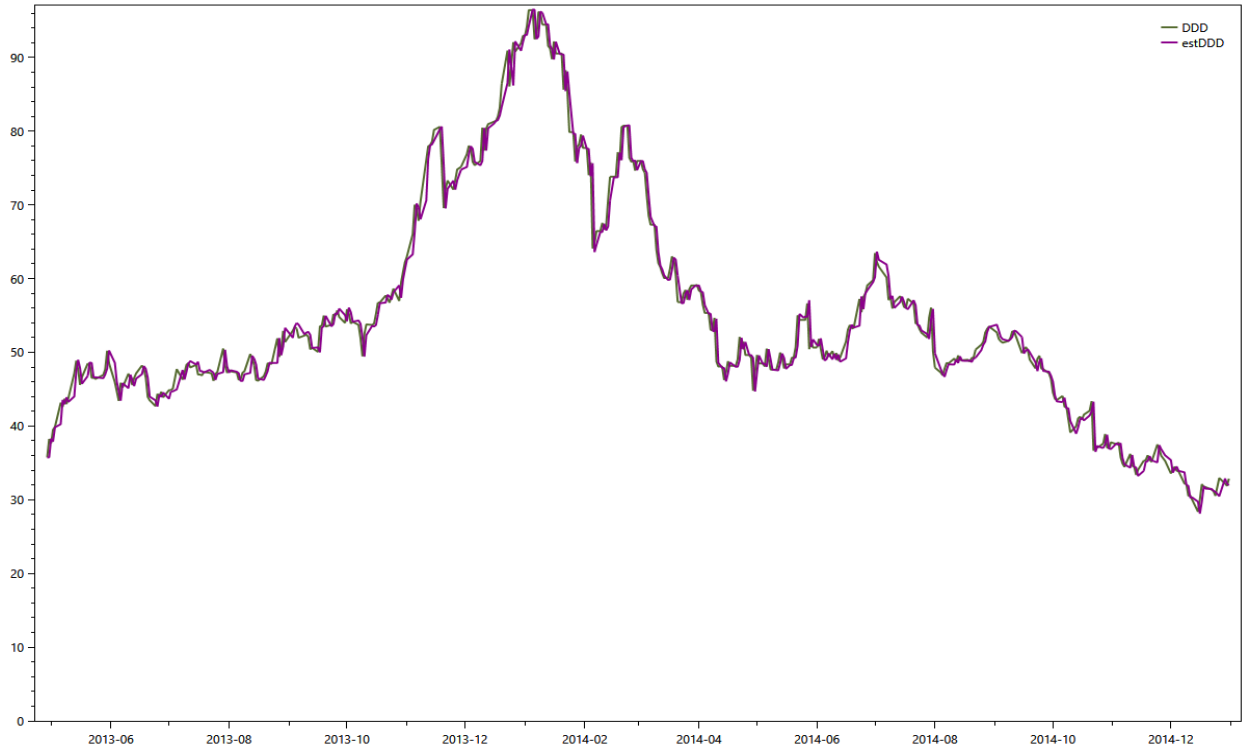

Рисунок 5.2 – Графики реальной и спрогнозированной динамики цены акции DDD

5.3 Имитация торговли

В качестве начального капитала была взята сумма 5000$. Приоритет акции определялся по величине средних ожидаемых доходов на оценочном периоде:

1. TWX, $\mu = 0.064032$;
2. JNJ, $\mu = 0.0461797$;
3. CCE, $\mu = 0.0355928$;

4. XOM, $\mu = 0.0099438$;
5. DDD, $\mu = -0.01275602$;
6. IBM, $\mu = -0.01641357$.

С учетом данных за оценочный период, были получены следующие веса оптимальных портфелей Марковица:

Таблица 4.3 – Веса акций в оптимальных портфелях Марковица

	Минимальный риск	Максимальная доходность
TWX	0.073236	0.563571
JNJ	0.369994	0.436426
CCE	0.106351	$2.37744*10^{-6}$
XOM	0.257233	$3.27853*10^{-7}$
DDD	0.0161861	$1.54406*10^{-7}$
IBM	0.177	$1.49623*10^{-7}$

5.4 Эффективность стратегии

В конце периода итоговый капитал, накопленный в ходе торговли по разработанному алгоритму, составил 5031.14$, что немногим превышает первоначальный капитал. Однако владение оптимальными портфелями Марковица минимального риска и максимальной доходности привело в итоге к величине капитала 4546.87$ и 4290.22$ соответственно, что меньше изначальных инвестиций. То же самое, однако в меньшей степени, наблюдается и у индексов: 4885.61$ для DJI и 4965.35$ для SPX.

Таким образом, доходность составила:

- торговая стратегия:
  $0{,}6228\%$;
- портфель минимального риска:
  $-9{,}0626\%$;
- портфель максимальной доходности:
  $-14{,}1956\%$;
- индекс Dow Jones:
  $-2{,}2877\%$;
- индекс S&P 500:

−0,6928%.

На рисунке 5.3 приведена динамика стоимости портфеля стратегии и рассматриваемых портфелей Марковица, а на рисунке 5.4 – портфеля стратегии и фондовых индексов (все с учетом наличия свободного капитала).

Рисунок 5.3 – Графики динамик стоимости портфелей

Рисунок 5.4 – Графики динамик стоимости индексов

Таким образом, на рассматриваемом периоде данных разработанная стратегия не принесла больших доходов, однако помогла в итоге избежать убытков. С учетом состояния рынка, это можно считать неплохим результатом.

Заключение

В данной работе была расширена и применена на практике исходная стратегия, описанная в статье[1]. По моему мнению, основным ее преимуществом стала замена фиксированного количества покупаемых/продаваемых активов на более гибкую величину, которая определяется на основании текущего и спрогнозированного состояний рынка. Практические исследования показали, что данное изменение позволяет ограничить риски потерь, а также увеличить доходы.

То, что торговля ведется с дневным периодом, дает гораздо больше времени на оценку имеющихся данных рынка. В связи с этим, в параметрах моделей можно использовать гораздо больше переменных регрессии, предшествующие интервалы большей длины и большее количество шаблонов. Также есть возможность чаще обновлять эти параметры. Все это теоретически должно хорошо сказываться на результатах прогнозирования и, как следствие, результатах торговли. Однако с практической точки зрения это не всегда так.

Список использованной литературы

1. Devavrat Shah. Bayesian regression and Bitcoin. / Shah Devavrat, Kang Zhang // Massachusetts Institute of Technology [Electronic resource]. – 2014. – Mode of access: http://arxiv.org/abs/1410.1231. – Date of access: 09.05.2015.
2. Chen, G. H. A latent source model for nonparametric time series classification. / G. H. Chen, S. Nikolov, D. Shah // Neural Information Processing Systems [Electronic resource]. – 2013. – Mode of access: http://arxiv.org/abs/1302.3639. – Date of access: 09.05.2015.
3. Bresler, G. A latent source model for online collaborative filtering. / G. Bresler, G. H. Chen, D. Shah // Neural Information Processing Systems [Electronic resource]. – 2014. – Mode of access: http://arxiv.org/abs/1411.6591. – Date of access: 09.05.2015.
4. Tibshirani, R. Regression shrinkage and selection via the lasso / R. Tibshirani // Journal of the Royal Statistical Society. Series B (Methodological). – 1996. – P. 267-288.
5. Lo, A.W. Stock market prices do not follow random walks: Evidence from a simple specification test / A.W. Lo, A.C. MacKinlay // Review of Financial Studies – 1988. – Vol. 1. – P. 41-66.
6. Lo, A.W. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation / A.W. Lo, H. Mamaysky, J. Wang // Journal of Finance. – 2000. – Vol.4.
7. Caginalp, G. The predictive power of price patterns / G. Caginalp, H. Laurent // Applied Mathematical Finance. – 1988. – Vol.5. – P. 181-206.
8. Sharpe, W.F. The sharpe ratio / W.F. Sharpe // Streetwise–the Best of the Journal of Portfolio Management. – 1998. – P. 169-185.