МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «КубГУ»)

Факультет компьютерных технологий и прикладной математики
Кафедра анализа данных и искусственного интеллекта

КУРСОВАЯ РАБОТА

МЕТОД АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТОВ И КЛАСИФИКАЦИЯ. STATISTICA И PYTHON

Работу выполнила _____ М.Г. Шарыпова
(подпись)

Направление подготовки 09.03.03 Прикладная информатика

Направленность (профиль) Прикладная информатика в экономике

Научный руководитель
д-р. техн. наук, проф. _____ А.А. Халафян
(подпись)

Нормоконтролер
канд. физ.-мат. наук, доц. _____ Г.В. Калайдина
(подпись)

Краснодар
2021

РЕФЕРАТ

Курсовая работа 26 страниц, 21 рисунок, 10 источников.

STATISTICA, КОМПОНЕНТ, PYTHON, МЕТОД, ГЛАВНЫХ, АНАЛИЗ, ПЕРЕМЕННЫЕ, КЛАССИФИКАЦИЯ, ПРИЗНАК, ФАКТОР

Объектом исследования являются STATISTICA, Python, а также один из методов для работы с ними.

Цель курсовой работы – рассмотрение метода анализа главных компонентов, определение взаимосвязей между переменными, их классификация, проведение анализа данным методом в statistica и python и сравнение их.

Итог проделанной работы – сравнение метода в statistica и python и нахождение оптимального варианта.

В результате был проведён статистический анализ данных, функциями которого являются:

– Поиск взаимосвязи между переменными в данных;

– Интерпретация и визуализация данных;

– Упрощение дальнейшего анализа посредством уменьшения количества переменных;

– Визуализация генетической дистанции и взаимосвязь между популяциями.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую очередь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на нем признаков. Так, например, для определения специфики фигуры человека при покупке одежды достаточно назвать значения двух признаков (размер-рост), являющихся производными от измерений ряда параметров фигуры. При этом, конечно, теряется какая-то доля информации (портной измеряет до одиннадцати параметров на клиенте), как бы огрубляются (при агрегировании) получающиеся при этом классы. Однако, как показали исследования, к вполне удовлетворительной классификации людей с точки зрения специфики их фигуры приводит система, использующая три признака, каждый из которых является некоторой комбинацией от большого числа непосредственно замеряемых на объекте параметров.

Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам[1].

Таким образом тема «Методы главных компонент» является актуальной.

Целью данной курсовой является рассмотрение метода главных компонент. В соответствии с поставленной целью необходимо выполнить следующие задачи:

- рассмотрение статистического подхода в методе главных компонент;
- примеры использования главных компонент в экономике;
- экономико-математическое моделирование факторов[2].

1 Основные методы редукции данных

1.1 Метод главных компонент и классификация. Определение. Задачи метода

Метод главных компонент и классификация – один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Изобретен Карлом Пирсоном в 1901 г. Применяется во многих областях, таких как распознавание образов, компьютерное зрение, сжатие данных и т. п. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных. Иногда метод главных компонент называют преобразованием Кархунена-Лоэва или преобразованием Хотеллинга. Другие способы уменьшения размерности данных – это метод независимых компонент, многомерное шкалирование, а также многочисленные нелинейные обобщения: метод главных кривых и многообразий, метод упругих карт, поиск наилучшей проекции, нейросетевые методы «узкого горла», самоорганизующиеся карты Кохонена и др.

Как известно, социально-экономическое явление можно характеризовать целым рядом признаков. При большом наборе таких признаков в корреляционно-регрессионном анализе влияние связей становится затруднен, поэтому возникает необходимость сжатия, т. е. описание изучаемого явления (объекта) более укрупненным показателям, так называемыми "главными компонентами". Исходным степени здесь корреляционная матрица, на основании которой с использованием метода главных компонент может быть продлен анализ значений наблюдаемых признаков.

Правильно отобранные в корреляционную модель признаки, как правило, связаны между собой. Наличие таких связей между ними позволяет на основе одного фактора иметь информацию о другом. Существование

тесной связи между признаками дает основание для исключения одной из них. Например, если в модель урожайности включены две переменные x и x_2, характеризующих денежные затраты на гектар, первая - все виды, вторая - затраты на удобрения. Здесь практически будет лишним при включении в модель признаки x исследовать также и признак x_2, поскольку она тесно связана с первой. Идея учета одного признака на основании второго лежит в основе метода главных компонент.

Следует отметить, что речь не идет только о двух признаки. В таком случае метод главных компонент малоэффективен. Его используют, как правило, при десятках взаимосвязанных признаков. При этом ставится цель "набрать" определенную часть общей вариации результативного признака минимальным количеством переменных. Последние подбирают до тех пор, пока сумма их дисперсий НЕ достигает заданной доли в дисперсии исследуемого явления (например, 60%, 80%, 90% и т. д.).

Метод главных компонент решает следующие задачи:

– возмещение скрытых, объективно существующих закономерностей в изменении явлений[3];

– характеристика изучаемого, числом признаков, значительно меньше взятых, на начальном этапе. Число главных компонент, выделенных в процессе исследования, будет содержать (в компактной форме) больше информации, чем изначально измерены признаки;

– выявление признаков, наиболее тесно связанных с главной компонентой. Иначе говоря, изучение связи при которой с изменением одной переменной изменяется закон распределения второй, между ними;

– прогнозирования уровней изучаемых явлений на основании уравнения регрессии, которое получено по информации главных компонент.

Преимущества такого метода прогнозирования в отличие от классического регрессионного анализа можно объяснить тем, что при последнем в модель пытаются включить максимально возможное количество факторов, в экономических явлениях часто характеризуются существенной

коррелируемости. Прогноз по таким переменным, как правило, бывает не точным. Поэтому возникает задача о замене исходных взаимосвязанных переменных совокупности некоррелированных параметров. Эта задача решается математическим аппаратом - методом главных компонент, который представляет собой характеристики, построенные на основе первично измеренных признаков.

Реализация практических возможностей указанных выше задач, которые решаются методом главных компонент в области экономики, может быть представлена различным направлениям.

Назовем их:

– анализ причинно-следственных взаимосвязей показателей и установления их стохастического связи с главными компонентами.

– выделение обобщающих экономических показателей;

– ранжирования результатов наблюдений по главным компонентам

– классификация объектов наблюдения;

– список исходной информации;

– построение уравнений регрессии по обобщающим экономическим показателям.

Как негативную сторону метода главных компонент следует назвать сложность математического аппарата, обусловленного абсолютностью знаний теории вероятностей, математической статистики, линейной алгебры, а также математического обеспечения ЭВМ. Формальное использование стандартных программ без понимания математической сути вычислительных процедур может привести к необоснованным выводам. Следует также помнить о профессиональные знания сути изучаемых экономических явлений. Только при таких условиях метод главных компонент может стать мощным математическим средством познания существующих реалий в области социально-экономических явлений.

1.2 Статистический подход в методе главных компонент. Примеры использования главных компонент в экономике

Компонентный анализ относится к многомерным методам снижения размерности. Он содержит один метод - метод главных компонент. В этом методе линейные комбинации случайных величин определяются характеристическими векторами ковариационной матрицы. Главные компоненты представляют собой ортогональную систему координат, в которой дисперсии компонент характеризуют их статистические свойства.

В зависимости от конкретных задач, решаемых в экономике, используется один из методов факторного анализа, или метод главных компонент.

Метод главных компонент считается статистическим методом. Однако есть другой подход, приводящий к методу главных компонент, но не являющийся статистическим. Этот подход связан с получением наилучшей проекции точек наблюдения в пространстве меньшей размерности. Для решения подобной задачи необходимо знать матрицу вторых моментов.

В статистическом подходе задача будет заключаться в выделении линейных комбинаций случайных величин, имеющих максимально возможную дисперсию. Он опирается на ковариационную или корреляционную матрицу этих случайных величин. У этих двух разных подходов есть общий аспект: использование матрицы вторых моментов как исходной для начала анализа[5].

Из сказанного следует, что для овладения методом главных компонент необходимо пользоваться методами теории вероятностей и математической статистики на основе моделей линейной алгебры. Рассмотрим основные положения этих математических дисциплин, на которые опирается метод главных компонент.

Учитывая, что объекты исследования в экономике (фирма, завод, министерство, отрасль народного хозяйства, экономика страны)

характеризуются большим, но конечным количеством признаков (характеристик), влияние которых подвергается воздействию большого количества случайных причин, в качестве моделей в статистическом плане возьмем многомерные распределения, а в алгебраическом - многомерное пространство признаков.

Если рассматривать с экономический точки зрения то метод главных компонент применяется в оценке стоимости бизнеса, так же этом метод применяется при анализе экономической безопасности региона, для анализа признаков, оказывающих наибольшее влияние на результаты деятельности банков.

Применение метода осуществляется так же в анализе рыночной конъюнктуры, модели рыночной конъюнктуры.

Говоря о методе многомерного статистического анализа при помощи главных компонент, а также оценки эффективности экономических организаций, экономических систем и систем управления рассматривают задачи обработки многомерных наблюдений в экономике и проблемы совершенствования метода главных компонент и расширения области его применения. Изучаются основные принципы исследования операций, используемые в теории эффективности; дается оценка эффективности на основе критериев – игровых, информационных, теории массового обслуживания[4].

2 Практическая реализация

2.1 Метод главных компонент и классификация в STATISTICA

Метод анализ главных компонент и классификация служит для достижения двух целей: уменьшение общего числа переменных и классификация переменных и наблюдений, при помощи строящегося факторного пространства[6].

Проведем анализ экономики в различных странах (Рисунок 1).

	1 Нас. 1990	2 Нас. 2000	3 Нас. 2010	4 Пл., км2	5 Форма пр-я	6 Пр.жизни (М)	7 Пр.жизни (Ж)	8 Рост населения	9 Пл.населени	10 ВВП ($ млн.)
Австралия	16937	19038	22152	768685(	M	77.8	83.6	Да	2.8818046	153240
Австрия	7645	8002	8375	768685(	P	76.3	82.3	Да	1.0895230	39964
Бельгия	9948	10239	1084(	32545	M	75.8	82.2	Да	333.07727	48370
Болгария	8767	8191	7564	110910	P	70	76.4	Нет	68.19944	5103(
Великобритания	57157	58785	62027	244101	M	76.2	81.3	Да	254.10383	247178
Германия	79113	82163	81802	357022	P	76	82.1	Да	229.12313	342813
Греция	10121	10904	11305	131940	P	76.9	82.1	Да	85.682886	24909
Дания	5135	533(	5535	43094	M	75.7	80.4	Да	128.44015	31424
Ирландия	3507	3778	4468	70273	P	75.3	80.7	Да	63.580607	21033
Испания	3882(	4005(	4598(	504782	M	76.5	83.3	Да	91.106655	132296
Италия	56694	56924	6034(	30123(	P	77	83.1	Да	200.31205	201467
Канада	27464	3052(	3391(	998467(	M	77	83.9	Да	3.3968073	182142
Латвия	2668	2382	2248	64589	P	66.4	77.1	Нет	34.804688	2837
Литва	3694	3512	3329	65200	P	69.5	79.7	Нет	51.058282	4224(
Нидерланды	14893	15864	16575	4152(	M	76.8	81.3	Да	399.14752	77222
Польша	38038	38654	38167	312685	P	73.9	79.4	Да	122.06213	48979
Россия	147665	14689(	142834	1707540(	P	64	75.6	Нет	8.3648992	201477
США	248147	28072(	308061	937261(	P	75.2	81	Да	32.868219	1624460(
Украина	51557	49115	45783	603700	P	62.2	74	Нет	75.837336	17630
Финляндия	4974	5171	5351	338145	P	75.2	82.3	Да	15.824572	25002
Франция	56577	60545	64659	547030	P	77.7	84.3	Да	118.20009	261287
Чехия	10362	10278	10507	78866	P	73.1	79.9	Да	133.22597	19565
Швейцария	6674	7164	7786	41290	P	77.8	83.6	Да	188.56866	63219
Швеция	8527	8861	9341	449964	M	78.4	83	Да	20.759438	52574
Япония	123205	12668(	12751(	377835	M	78.7	85.6	Да	337.47535	595971

Рисунок 1 – Таблица данных

Исходная таблица содержит выборку из 29 переменных.

Столбцы таблицы:

– количество населения за 1990, 2000, 2010 года в тысячах человек;

– площадь в км^2;

– форма правления, где Р-республика, М-монархия;

– средняя продолжительность жизни мужчин и женщин;

– показатель, выросло ли население за выбранный временной промежуток;

– плотность населения (чел./км2);

– размер ВВП ($ млн.).

2.2 Проведение анализа:

Выберем переменные:

для анализа – Нас 1990, Нас 2000, Нас 2010, пр. жизни(М), пр. жизни(Ж), ВВП.;

вспомогательные переменные – Пл. км, Пл. нас;

переменные с основными наблюдениями – Рост населения;

группирующая переменная – Форма пр-я.

Получим Рисунок 2:

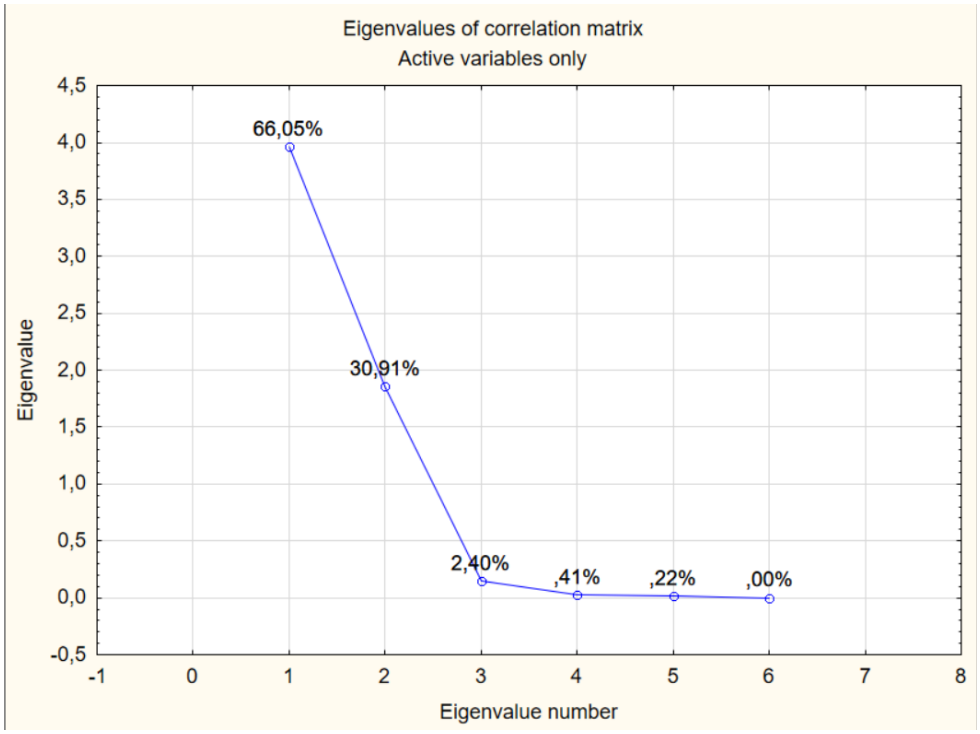

Рисунок 2 – График 1

Из графика видно, что число выделяемых факторов может быть 2 или 3. Эта таблица собственных значений(Рисунок 3).

Value number	Eigenvalues of correlation matrix, and related statisti Active variables only			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,962844	66,04740	3,962844	66,0474
2	1,854772	30,91286	5,817615	96,9603
3	0,143965	2,39941	5,961580	99,3597
4	0,024839	0,41398	5,986419	99,7736
5	0,013419	0,22365	5,999838	99,9973
6	0,000162	0,00270	6,000000	100,0000

Рисунок 3 – Таблица собственных значений

Во втором столбце (Eigenvalue) приведены дисперсии выделенных факторов – собственные числа, в третьем – процент от общей дисперсии. Как видно, первый фактор объясняет 66,04% общей дисперсии, второй – 30,9% и т. д. По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются только 2 фактора.

Variable	Factor coordinates of Active and Supplemer *Supplementary varial	
	Factor 1	Factor 2
Нас. 1990	0,995321	-0,012993
Нас. 2000	0,997817	-0,028690
Нас. 2010	0,992189	-0,052054
ВВП ($ млн)	0,991031	-0,019067
Пр. жизни (М)	0,018737	0,964689
Пр. жизни (Ж)	0,098025	0,959209
*Пл., км2	0,403016	0,098138
*Пл. населения	0,016887	0,077691

Рисунок 4 – Факторные координаты переменных

В данной таблице(Рисунок 4) представлены факторные координаты переменных (факторные нагрузки). Большее абсолютное значение факторной нагрузки переменной с каким-либо фактором говорит о том, что переменная сильнее связана с этим фактором. Вспомогательные переменные обозначены «*». Первая факторная ось наиболее сильно коррелирует с переменными Нас. 1990, 2000, 2010, и ВВП. (сильные положительные корреляции), Пл., км2 (умеренные положительные корреляции). Вторая факторная ось наиболее сильно коррелирует с переменной Пр. жизни (М), Пр. жизни (Ж). (сильные положительные корреляции).

График факторных координат переменных и наблюдений(Рисунок 5):

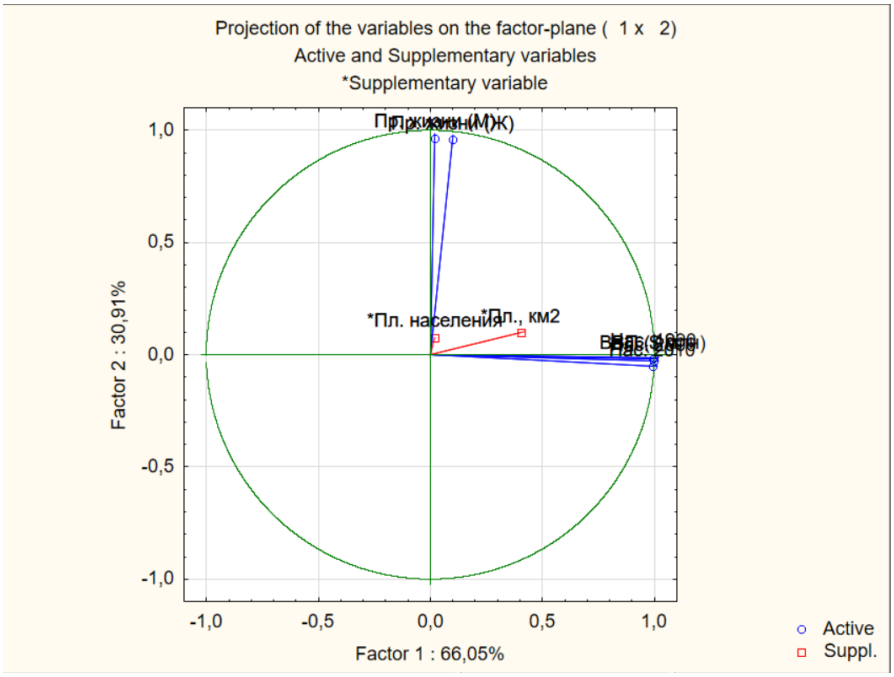

Рисунок 5 – График выделенных факторов

Чем ближе переменная к единичной окружности, тем лучше она воспроизведена в найденной системе координат (лучше воспроизводится текущим набором выделенных факторов).

Вклад переменных в дисперсию факторной оси(Рисунок 6):

Variable	Variable contribution	
	Factor 1	Factor 2
Нас. 1990	0,249988	0,000091
Нас. 2000	0,251243	0,000444
Нас. 2010	0,248417	0,001461
ВВП ($ млн)	0,247838	0,000196
Пр. жизни (М)	0,000089	0,501747
Пр. жизни (Ж)	0,002425	0,496062

Рисунок 6 – Вклад переменных

Variable	Communalities, base Active and Suppleme *Supplementary varia	
	From 1 factor	From 2 factors
Нас. 1990	0,990665	0,990833
Нас. 2000	0,995638	0,996461
Нас. 2010	0,984440	0,987149
ВВП ($ млн)	0,982142	0,982505
Пр. жизни (М)	0,000351	0,930976
Пр. жизни (Ж)	0,009609	0,929690
*Пл., км2	0,162422	0,172053
*Пл. населения	0,000285	0,006321

Рисунок 7 – Таблица общностей переменных

Это таблица общностей переменных(Рисунок 7). Общность – это доля объясненной дисперсии, которая характеризует степень общности переменной с другими переменными по заданному числу факторов[7]. Из таблицы общностей переменных видно, что самая высокая степень общности с другими переменными для первого фактора у переменной Нас. 2000, у второго фактора - у Нас. 2000.

Case	Cosine squares, based on correlations (Spreadsheet1 Active cases variable: Рост населения Labelling varia Code for active cases: Да ; Suppl. cases highlighted **Factor 1**	Factor 2	Рост населения	Форма пр-я
Австралия	0,174865	0,802167	Да	M
Австрия	0,987039	0,000619	Да	P
Бельгия	0,872158	0,058824	Да	M
Болгария	0,052912	0,937640	Нет	P
Великобритания	0,236682	0,498942	Да	M
Германия	0,881251	0,073858	Да	P
Греция	0,878751	0,043506	Да	P
Дания	0,505440	0,414051	Да	M
Ираландия	0,532668	0,456219	Да	P
Испания	0,044078	0,583045	Да	M
Италия	0,181648	0,739069	Да	P
Канада	0,070112	0,790730	Да	M
Латвия	0,036880	0,847472	Нет	P
Литва	0,067359	0,791010	Нет	P
Нидерланды	0,646877	0,026471	Да	M
Польша	0,034328	0,950821	Да	P
Россия	0,043365	0,843310	Нет	P
США	0,960875	0,037344	Да	P
Украина	0,002048	0,908639	Нет	P
Финляндия	0,422255	0,174909	Да	P
Франция	0,097241	0,878405	Да	P
Чехия	0,156113	0,803510	Да	P
Швейцария	0,363710	0,626021	Да	P
Швеция	0,307399	0,564384	Да	M
Япония	0,497856	0,492093	Да	M

Рисунок 8 – Косинусные квадраты, основанные на корреляции

В данной таблице(Рисунок 8) представлена информация о принадлежности наблюдения к основным или вспомогательным наблюдениям. Например, страны Болгария, Латвия, Литва, Россия и Украина относятся к не основным наблюдениям.

График наблюдений в факторном пространстве(Рисунок 9):

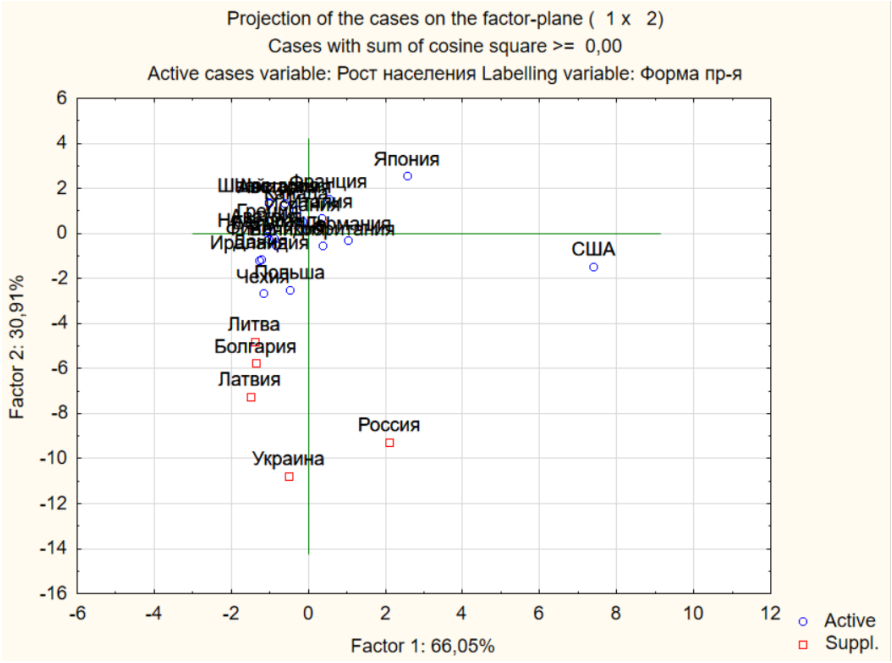

Рисунок 9 – График наблюдений в факторном пространстве

Из всех наблюдений выше можно сказать, что ВВП страны сильно зависит от средней продолжительности мужчин и женщин, а также от количества населения за 1990, 2000 и 2010 года. Так, чем выше количество населения и средняя продолжительность жизни, тем выше ВВП в стране. Наблюдения синего цвета - основные (страны, у которых ВВП значительно возрос за год), красного - вспомогательные (страны, у которых ВВП не возрос значительно за год).

2.3 Метод главных компонент и классификация в Python

Для начала, чтобы провести необходимый анализ, импортируем необходимые нам библиотеки[8] и базу данных. Набор данных, который мы будем использовать в этой статье, – это знаменитый набор данных Iris. Набор данных состоит из 150 записей растений ириса с четырьмя признаками: “длина чашелистика”, “ширина чашелистика”, “длина лепестка” и “ширина лепестка”. Все функции числовые. Записи были классифицированы на один из трех классов: “Iris-setosa”, “Iris-versicolor” или “Iris-virginica”. Выполним

следующий скрипт для загрузки набора данных с помощью pandas(Рисунок 10).

```
Ввод [1]: import numpy as np
          import pandas as pd
          url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
          names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'Class']
          dataset = pd.read_csv(url, names=names)
```

Рисунок 10 – Импорт библиотек и загрузка набора данных

Далее посмотрим, как выглядит наш набор данных(Рисунок 11):

```
Ввод [2]: dataset.head()
```

Out[2]:

	sepal-length	sepal-width	petal-length	petal-width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Рисунок 11 – Предварительный просмотр набора данных

Выполняем предварительную обработку[10]. Первый шаг предварительной обработки состоит в том, чтобы разделить набор данных на набор объектов и соответствующие метки. Данная команда хранит наборы функций в переменной x и ряд соответствующих меток в переменной y. Следующим шагом предварительной обработки является разделение данных на обучающие и тестовые наборы(Рисунок 12).

```
Ввод [3]: X = dataset.drop('Class', 1)
          y = dataset['Class']

          C:\Users\shary\AppData\Local\Temp/ipykernel_41824/2098761270.py:1: FutureWarning: In a future version of pandas all arguments o
          f DataFrame.drop except for the argument 'labels' will be keyword-only
            X = dataset.drop('Class', 1)

Ввод [*]: from sklearn.model_selection import train_test_split

          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

Ввод [ ]:
```

Рисунок 12 – Разделение набора данных

Метод главных компонент лучше всего работает с нормализованным набором функций. Мы выполним стандартную скалярную нормализацию, чтобы нормализовать наш набор функций(Рисунок 13).

Рисунок 13 – Выполнение препроцессорной части

Применяем функции метода анализа главных компонент в обучающем и тестовом наборе для анализа. Для этой цели используется класс PCA. PCA зависит только от набора функций, а не от данных метки. Поэтому PCA можно рассматривать как неконтролируемую технику машинного обучения. Выполнение PCA с помощью Scikit-Learn это двухэтапный процесс (Рисунок 14):

– инициализируем класс PCA, передав конструктору количество компонентов;

– вызовем методы fit, а затем transform, передав набор функций этим методам. Метод transform возвращает заданное количество основных компонентов.

Рисунок 14 – Применение метода анализа главных компонент и классификации

В приведенном выше коде мы создаем объект PCA с именем pca. Мы не указывали количество компонентов в конструкторе. Следовательно, все четыре функции в наборе функций будут возвращены как для обучающего, так

и для тестового наборов. Класс PCA содержит explained_variance_ratio_, который возвращает дисперсию, вызванную каждым из основных компонентов. Выполните следующую строку кода, чтобы найти “объясненный коэффициент дисперсии”. (Рисунок 15):

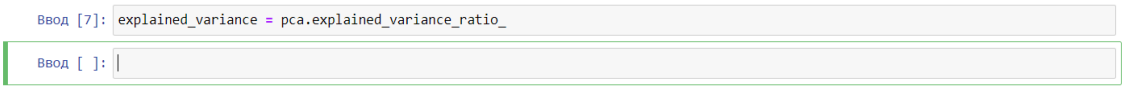

Рисунок 15 – Функция возвращения дисперсии

Переменная explained_variance теперь является массивом типа float, который содержит коэффициенты дисперсии для каждого основного компонента. Значения переменной explained_variance выглядят следующим образом(Таблица 1):

Таблица 1 – Коэффициенты дисперсии

Коэффициенты
0.722265
0.239748
0.033381
0.004606

Можно видеть, что первый главный компонент отвечает за 72,22% дисперсии. Аналогично, второй основной компонент вызывает 23,9% – ную дисперсию в наборе данных. В совокупности мы можем сказать, что (72.22 + 23.9) 96.21% процентов классификационной информации, содержащейся в наборе признаков, захватывается первыми двумя основными компонентами. Для начала попробуем использовать 1 главный компонент для обучения нашего алгоритма (Рисунок 16):

```
Ввод [8]: from sklearn.decomposition import PCA
          pca = PCA(n_components=1)
          X_train = pca.fit_transform(X_train)
          X_test = pca.transform(X_test)

Ввод [ ]:
```

Рисунок 16 – Использование 1 главного компонента

Далее мы будем использовать классификацию случайных лесов для составления прогнозов(Рисунок 17):

```
Ввод [*]: from sklearn.ensemble import RandomForestClassifier
          classifier = RandomForestClassifier(max_depth=2, random_state=0)
          classifier.fit(X_train, y_train)
          y_pred = classifier.predict(X_test)

Ввод [ ]:
```

Рисунок 17 – Классификация случайных лесов для составления прогнозов

Далее проводим оценку эффективности(Рисунок 18):

```
Ввод [10]: from sklearn.metrics import confusion_matrix
           from sklearn.metrics import accuracy_score

           cm = confusion_matrix(y_test, y_pred)
           print(cm)
           print('Accuracy' + accuracy_score(y_test, y_pred))
```

Рисунок 18 – Оценка Эффективности

Вывод приведенного выше выглядит следующим образом(Рисунок 19):

```
[[11  0  0]
 [ 0 12  1]
 [ 0  1  5]]
0.933333333333
```

Рисунок 19 – Матрица 1

Из выходных данных видно, что только с одной функцией алгоритм случайного леса способен правильно предсказать 28 из 30 экземпляров, что приводит к точности 93,33%. Теперь попробуем оценить эффективность

классификации алгоритма случайного леса с 2 основными компонентами. (Рисунок 20):

```
Ввод [12]: from sklearn.decomposition import PCA
           pca = PCA(n_components=2)
           X_train = pca.fit_transform(X_train)
           X_test = pca.transform(X_test)
```

Рисунок 20 – Оценка эффективности 2

Здесь количество компонентов для PCA было установлено равным 2. Результаты классификации с 2 компонентами следующие(Рисунок 21):

```
[[11  0  0]
 [ 0 10  3]
 [ 0  2  4]]
0.833333333333
```

Рисунок 21 – Матрица 2

При двух основных компонентах точность классификации снижается до 83,33% по сравнению с 93,33% для 1 компонента.[9] С тремя основными компонентами результат выглядит следующим образом(Рисунок 22):

```
[[11  0  0]
 [ 0 12  1]
 [ 0  1  5]]
0.933333333333
```

Рисунок 22 - Матрица 3

С тремя основными компонентами точность классификации снова возрастает до 93.33% Далее попробуем найти результаты с полным набором функций. Для этого просто удалим компьютер отдельно от сценария, который мы написали выше. Результаты с полным набором функций, без применения PCA выглядят следующим образом(Рисунок 23):

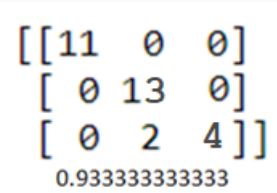

Рисунок 23 -Матрица 4

Точность, полученная с полным набором функций для алгоритма случайного леса, также составляет 93,33%. В результате вышеприведенных экспериментов мы достигли оптимального уровня точности при значительном сокращении количества объектов в наборе данных. Мы видели, что точность, достигнутая только с 1 основным компонентом, равна точности, достигнутой с набором функций воли, то есть 93,33%. Уместно также отметить, что точность классификатора не обязательно повышается с увеличением числа основных компонентов. Из полученных результатов видно, что точность, достигнутая с одним основным компонентом (93,33%), была больше, чем точность, достигнутая с двумя основными компонентами (83,33%). Количество основных компонентов, сохраняемых в наборе функций, зависит от нескольких условий, таких как емкость хранилища, время обучения, производительность и т. Д. В некоторых наборах данных все функции в равной степени участвуют в общей дисперсии, поэтому все основные компоненты имеют решающее значение для предсказаний, и ни один из них не может быть проигнорирован. Общее эмпирическое правило состоит в том, чтобы взять количество главных компонентов, которые вносят значительный вклад в дисперсию, и игнорировать те, которые имеют уменьшающуюся дисперсию

отдачи. Хороший способ-построить дисперсию по основным компонентам и игнорировать основные компоненты с уменьшающимися значениями, как показано на следующем графике(Рисунок 24):

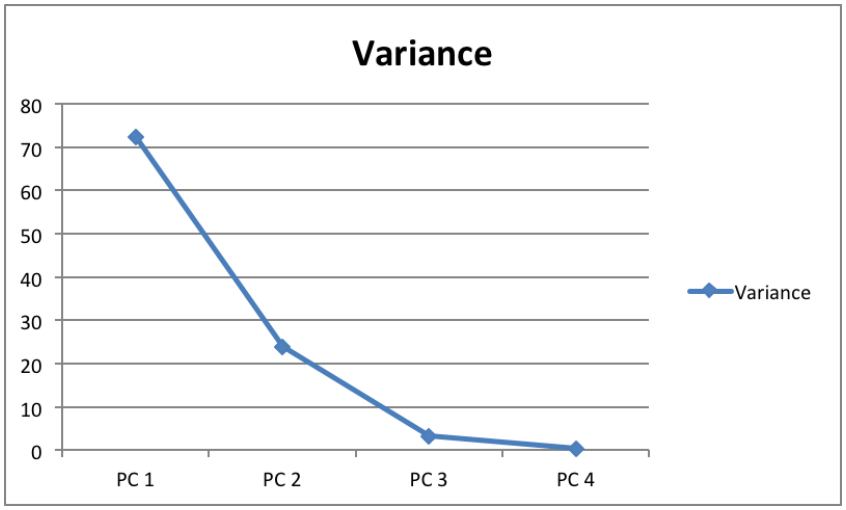

Рисунок 24 - График переменных

Например, на приведенной выше диаграмме мы видим, что после третьего главного компонента изменение дисперсии почти уменьшается. Поэтому можно выбрать первые три компонента.

3 Сравнение результатов и общие выводы

Мы провели анализ в STATISTICA и Python, по результатам можно выделить наиболее удобный способ проведения анализа. Можно сказать, что проведение анализа в STATISTICA наиболее выгодно, так как в ней очень простая и понятная навигация. Так как основной код для программы уже написан, все, что необходимо пользователю – ввести нужные данные и проанализировать результаты. Программа очень производительная и проведение анализа не требует много времени. Так же плюсом можно отметить, что программа прекрасно работает даже на слабых компьютерах.

В то время как в Python пользователю требуются необходимые знания для проведения анализа. Необходимо знать и понимать написание кода и работы с программой. Пользователь не может сразу приступить к анализу, для этого необходимо установить и подключить нужные библиотеки. Для загрузки базы данных и проведения полного анализа необходимо вводить код вручную. Если учитывать подключение библиотек, написание кода и общую компиляцию по проведению анализа, программа занимает довольно много времени, она требует очень высокой производительности и наличие мощного компьютера с хорошими параметрами.

ЗАКЛЮЧЕНИЕ

Подводя итог всему вышесказанному, можно сказать о том, что наличие множества исходных признаков, характеризующих процесс функционирования объектов, заставляет отбирать из них наиболее существенные и изучать меньший набор показателей. Чаще исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации. Такое решение может быть обеспечено методами снижения размерности, куда относят факторный и компонентный анализ. Эти методы позволяют учитывать эффект существенной многомерности данных, дают возможность лаконичного или более простого объяснения многомерных структур. Они вскрывают объективно существующие, непосредственно наблюдаемые закономерности при помощи полученных факторов или главных компонент. Они дают возможность достаточно просто и точно описать наблюдаемые исходные данные, структуру и характер взаимосвязей между ними. Сжатие информации получается за счет того, что число факторов или главных компонент – новых единиц измерения – используется значительно меньше, чем было исходных признаков[4].

На основании изученной темы и проделанной работы по написанию данной работы можно сделать вывод, что поставленные цель и задачи нашли здесь свое отражение.

Учитывая, что объекты исследования в экономике характеризуются большим, но конечным количеством признаков (характеристик), влияние которых подвергается воздействию большого количества случайных причин, в качестве моделей в статистическом плане берутся многомерные распределения.

В данной курсовой работе была построена математическая модель и программная реализация метода главных компонент в STATISTICA и Python. Следует отметить, что первая программа была более оптимизирована и предназначена для более широкого круга лиц, т. е. её могут использовать как

опытные пользователи, так и новички-любители. К достоинствам обоих использованных программ можно отнести то, что они могут работать с массивами исходных данных достаточно большой размерности.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. HOW PRINCIPAL COMPONENT ANALYSIS, PCA WORKS – URL: https://dataaspirant.com/principal-component-analysis-pca/ (10.12.2021).
2. Principal Component Analysis (PCA) – Part 1 – Fundamentals and Applications – URL: https://medium.com/analytics-vidhya/principal-component-analysis-pca-part-1-fundamentals-and-applications-8a9fd9de7596 (дата обращения 24.12.2021)
3. Факторный анализ. Метод главных компонент – URL: https://www.myuniversity.ru/Экономико-математическое_моделирование/Факторный_анализ_Метод_главных_компонент/487608_3531057_страница1.html (дата обращения 28.12.2021).
4. Метод главных компонент – URL: https://www.yaneuch.ru/cat_24/metod-glavnyh-komponent/270617.2265010.page2.html (дата обращения 28.12.2021)
5. Лекция: Метод главных компонент – URL: http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html (дата обращения 29.12.2021).
6. Халафян А. А. STATISTICA 6. Статистический анализ данных/ А. А. Халафян – М.: Бином, 2007.– 507 с.
7. Метод Главных Компонент (PCA) – URL: https://www.chemometrics.ru/old/Tutorials/pca.htm (дата обращения 8.01.2022).
8. Principal Component Analysis in Python - A Step-by-Step Guide (PCA)|Nick McCullum – URL: https://nickmccullum.com/python-machine-learning/principal-component-analysis-python/ (дата обращения 15.01.2022).
9. АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТОВ С ПОМОЩЬЮ PYTHON| ПОРТАЛ ИНФОРМАТИКИ ДЛЯ ГИКОВ – URL: http://espressocode.top/principal-component-analysis-with-python/ (дата обращения 18.01.2022).

10. Principal Component Analysis (PCA) in Python|DataCamp – URL: https://www.datacamp.com/community/tutorials/principal-component-analysis-in-python (дата обращения 20.01.2022).