

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Механико-математический факультет

УТВЕРЖДАЮ:

Декан



Л. В. Гензе

«30» 06 2022 г.

Рабочая программа дисциплины

Многомерные статистические методы обработки данных

по направлению подготовки

01.04.01 Математика

Направленность (профиль) подготовки :

Фундаментальная математика

Форма обучения

Очная

Квалификация

Магистр

Год приема

2022

Код дисциплины в учебном плане: Б1.В.3.ДВ.02.03

СОГЛАСОВАНО:

Руководитель ОП



П.А.Крылов

Председатель УМК



Е.А.Тарасов

Томск – 2022

1 Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен формулировать и решать актуальные и значимые проблемы математики.

ПК-1 Способен самостоятельно решать исследовательские задачи в рамках реализации научного (научно-технического, инновационного) проекта.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК 1.1 Формулирует поставленную задачу, пользуется языком предметной области, обоснованно выбирает метод решения задачи.

ИПК 1.1 Проводит исследования, направленные на решение отдельных исследовательских задач

2. Задачи освоения дисциплины

- Освоить теоретические методы многомерной статистики при анализе данных
- Научиться применять математические пакеты для решения практических задач профессиональной деятельности, связанных с обработкой и сортировкой больших данных

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплина (модули)».

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Первый семестр, зачет с оценкой

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 4 з.е., 144 часов, из которых:

-лекции: 16 ч.

-практические занятия: 16 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Многомерное нормальное распределение и его основные свойства

Двумерное нормальное распределение. Коэффициент корреляции. Корреляционная матрица. Оценивание параметров многомерного нормального распределения. ММП. Доверительные области для параметров многомерного нормального распределения. Проверка гипотез о параметрах распределения, гипотезы о независимости компонент вектора.

Тема 2. Метод главных компонент

Задача максимизации дисперсии главных компонент. Оценивание параметров главных компонент. Сокращение количества главных компонент (Критерий каменистой

осыпи, критерий Кайзера, критерий доли выделенной дисперсии, правило сломанной трости).

Тема 3. Факторный анализ.

Каноническая модель факторного анализа. Оценивание параметров модели факторного анализа методом главных компонент.

Тема 4. Метод канонических корреляций.

Выделение наиболее зависимых признаков на основе максимизации их корреляций. Оценивание параметров канонических корреляций. Проверка значимости канонических корреляций, отсеивание незначимых признаков.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, выполнения домашних практических заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет с оценкой в первом семестре проводится в письменной форме по билетам. Билет содержит один теоретический вопрос. Продолжительность зачета 1,5 часа.

Примерный перечень теоретических вопросов

1. Многомерное нормальное распределение, основные свойства.
2. Эллипсоид рассеяния многомерной нормальной случайной величины.
3. Вектор выборочных средних и выборочная матрица как оценки вектора математических ожиданий и матрицы ковариаций многомерного нормального вектора.
4. Оценки ММП для вектора математических ожиданий и матрицы ковариаций многомерного нормального вектора.
5. Доверительная область для вектора средних многомерной нормальной случайной величины.
6. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о равенстве вектора средних заданному вектору.
7. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о равенстве векторов средних двух совокупностей.
8. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о равенстве векторов средних нескольких совокупностей (критерий отношения правдоподобия).
9. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о равенстве матриц ковариаций нескольких совокупностей.
10. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о однородности нескольких совокупностей.
11. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о независимости множеств случайных величин.
12. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о сферичности распределения.
13. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о равенстве матрицы ковариаций заданной матрице.
14. Проверка статистических гипотез о параметрах многомерных нормальных совокупностей: гипотеза о равенстве матрицы ковариаций заданной матрице и вектора средних заданному вектору.
15. Метод главных компонент. Определение параметров линейных комбинаций исходя из условий максимума дисперсии и ортогональности.

16. Метод главных компонент. Оценка числа главных компонент.
17. Факторный анализ. Метод главных компонент, как метод факторного анализа. Оценка векторов факторных нагрузок исходя из условий максимума дисперсии и ортогональности.
18. Классификация обобщенных факторов в факторном анализе, вращение факторов.
19. Метод канонических корреляций. Определение параметров линейных комбинаций исходя из условий максимума корреляции и ортогональности.
20. Метод канонических корреляций, проверка значимости канонических корреляций, отсеив незначимых признаков.
21. Метод канонических корреляций. Определение выделенной каноническими переменными дисперсии и оценка избыточности множеств исходных признаков.

Примеры задач:

Задача №1.

1. По выборке из двумерной нормальной совокупности $\{X_2, X_3\}$ (таблица 1) построить доверительную область уровня $\beta = 0,92$ для вектора средних, если матрица ковариаций: а) известна (таблица 2); б) неизвестна.
2. Используя генератор стандартных нормальных чисел, получить выборку из $n = 100$ значений многомерной нормальной случайной величины $\{X_4, X_5, X_6\}$ с матрицей ковариаций A (таблица 2) и вектором средних μ (таблица 3). Используя полученную выборку, на уровне значимости $0,09$, проверить гипотезу (в предположении, что данные распределены нормально) о равенстве вектора средних и матрицы ковариаций генеральной совокупности вектору μ и матрице A соответственно.
3. По двум независимым выборкам объемов $n_1 = 30$ и $n_2 = 30$ из многомерных нормальных совокупностей $\{X_1, X_2, X_3\}$ и $\{X_4, X_5, X_6\}$ (таблица 4), на уровне значимости $0,05$ проверить гипотезу H_0 о равенстве матриц ковариаций $A(X_1, X_2, X_3) = A(X_4, X_5, X_6)$.
4. По выборке объема $n = 30$ из многомерной нормальной совокупности $\{X_4, X_5, X_6\}$ (таблица 1) проверить гипотезу о независимости компонент данной совокупности (указать достигнутый уровень значимости).

Задача № 2.

1) Для приведенных в варианте задания данных (цены закрытия акций на ММВБ за период с 01.01.2015 по 01.09.2015 с периодичностью 1 день: файл «Данные» лист «Котировки») выяснить целесообразность применения метода главных компонент (проверить гипотезу о зависимости рассматриваемых факторов).

2) Найти собственные значения и собственные векторы выборочной матрицы ковариаций, упорядоченные по убыванию собственных значений. Оценить долю общей дисперсии объясняемой каждой главной компонентой и совокупностью главных компонент (двух, трех и т.д.)

3) Определить количество главных компонент, ориентируясь на: долю выделенной дисперсии, критерий Кайзера, правило сломанной трости, критерий Каттеллера (каменистой осыпи),

4) Записать выражения для главных компонент (выписать в явном виде линейные комбинации, определяющие компоненты), а также получить оценки векторов значений главных компонент по всем наблюдениям.

Задача № 3.

- 1) Для относительных приращений, приведенных в варианте задания данных (цены закрытия акций на ММВБ за период с 01.01.2015 по 01.09.2015 с периодичностью 1 день: файл «Данные» лист «Котировки»), выяснить целесообразность применения факторного анализа (проверить гипотезу о зависимости рассматриваемых факторов – относительных приращений котировок ценных бумаг).
- 2) Провести факторный анализ относительных приращений методом главных компонент. Определить количество главных компонент, ориентируясь на долю выделенной дисперсии, а также руководствуясь критериями Кайзера и Кэттелла (каменистой осыпи). Провести факторный анализ относительных приращений методом главных компонент. Определить количество главных компонент, используя критерии Кайзера, Кэттелла (каменистой осыпи), а также ориентируясь на долю выделенной дисперсии.
- 3) Найти оценки главных компонент, упорядоченные по убыванию дисперсий (получить оценки векторов факторных нагрузок, оценки векторов значений обобщенных факторов по всем наблюдениям, записать выражения для обобщенных факторов и разложение для исходных признаков через обобщенные и характерные факторы). Оценить долю общей дисперсии, объясняемой каждым обобщенным фактором и всей совокупностью обобщенных факторов. Оценить дисперсии характерных факторов.

Задача №4.

В таблице (файл “Данные” лист “Показатели”) приведены значения показателей производственной деятельности для 53 предприятий машиностроения. Использовались следующие показатели: Y_1 - производительность труда; Y_2 – индекс снижения

себестоимости продукции; Y_3 – рентабельность; X_1 – трудоемкость единицы продукции; X_2 – удельный вес рабочих в составе ППП; X_3 – удельный вес покупных изделий; X_4 – коэффициент сменности оборудования; X_5 – премии и вознаграждения на одного работника; X_6 – удельный вес потерь от брака; X_7 – фондоотдача; X_8 – среднегодовая численность ППП; X_9 – среднегодовая стоимость ОПФ; X_{10} – среднегодовой фонд заработной платы ППП; X_{11} – фондовооруженность труда; X_{12} – оборачиваемость нормируемых оборотных средств; X_{13} – оборачиваемость ненормируемых оборотных средств; X_{14} – непроизводственные расходы.

Требуется для выбранных в соответствии с вариантом факторов (смотри таблицу 1) провести канонический анализ для двух групп факторов Y и X :

1. Найти оценки канонических переменных (найти канонические веса, использующиеся для вычисления значений канонических переменных) и канонических корреляций.
2. Произвести оценку значимости полученных канонических корреляций и, соответственно, отсеять незначимые пары канонических переменных. Записать выражения для значимых канонических переменных через исходные признаки.
3. Найти корреляции между каноническими переменными и переменными из каждого множества Y и X (координаты векторов канонических нагрузок).
4. Вычислить извлеченную дисперсию каждой канонической переменной и совокупностью канонических переменных (для каждого множества).

Результаты зачета с оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Итоговая оценка по курсу выставляется как средняя оценка между оценкой по практике и по теории. Для допуска к зачету должны быть сданы все практические задания.

Каждое практическое задание оценивается по пятибалльной шкале. Оценка по практике – средняя оценка по всем практическим заданиям.

Ответ на теоретический вопрос оценивается следующим образом. Студент получает оценку:

- 1) «отлично», если дан полный ответ на вопрос с выводом и также студент ответил на все дополнительные вопросы;
- 2) «хорошо», если студент не полностью изложил доказательство по теоретическому вопросу, но при этом ответил на все дополнительные вопросы;
- 3) «удовлетворительно», если студент ответил на вопрос из билета с существенными пропусками в доказательстве и ответил хотя бы на половину дополнительных вопросов.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - <https://moodle.tsu.ru/course/view.php?id=25418>

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине представлены в электронном курсе в «Moodle» в виде списка вопросов и практических заданий по дисциплине.

в) План практических занятий по дисциплине:

- 1) Занятия № 1-4 – выполнение задачи №1;
- 2) Занятия № 5-8 – выполнение задачи №2;
- 3) Занятия № 9-12 – выполнение задачи №3;
- 4) Занятия № 13-16 – выполнение задачи №4.

г) Методические указания по организации самостоятельной работы студентов.

Для изучения теории по курсу используются источники из перечней основной и дополнительной литературы. Также для выполнения практических заданий необходимо изучить один из математических пакетов «Matlab», «Statistica», «R» и другие.

Для изучения пакета «Matlab» можно использовать видеокурс - https://www.youtube.com/playlist?list=PLC-Umin_a5bMNXvaPz-mdg4nPjTuO0EnM.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

- Прикладной статистический анализ/ Е.И. Куликов. – М.: ГЛТ, 2018. – 464 с.
- Тюрин Ю.Н. Анализ данных на компьютере: Учебное пособие / Ю.Н. Тюрин, А.А. Макаров; Науч. ред. В.Э. Фигурнов. — М.: ИД ФОРУМ, 2017. — 368 с.

б) дополнительная литература:

- Дубров А. М. Многомерные статистические методы : учебник для экономистов и менеджеров / А. М. Дубров [и др.]. — М: Финансы и статистика, 1998. — 350 с.
- Факторный анализ : пер. с нем. / К. Иберла. — М: Статистика, 1980. — 398 с.
- Кендалл, М. Многомерный статистический анализ и временные ряды : пер. с англ. / М. Кендалл (Кендэл), А. Стьюарт. — М: Наука, 1976. — 736 с.
- Методы эконометрики/ С.А. Айвазян. — М: Магистр, 2014, 506 с.

в) ресурсы сети Интернет:

- <http://www.mathnet.ru.ru/> - общероссийский математический портал
- <http://www.lib.mexmat.ru> – электронная библиотека механико-математического факультета Московского государственного университета

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

- Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);
- публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

б) информационные справочные системы:

- Электронный каталог Научной библиотеки ТГУ – <http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system>
- Электронная библиотека (репозиторий) ТГУ – <http://vital.lib.tsu.ru/vital/access/manager/Index>
- ЭБС Лань – <http://e.lanbook.com/>
- ЭБС Консультант студента – <http://www.studentlibrary.ru/>
- Образовательная платформа Юрайт – <https://urait.ru/>
- ЭБС ZNANIUM.com – <https://znanium.com/>

в) профессиональные базы данных (при наличии):

– Академия анализа данных – http://statsoft.ru/#tab-case_studies-link

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешанном формате («Актру»).

15. Информация о разработчиках

Губин Владимир Николаевич, кандидат физико-математических наук, ММФ ТГУ, доцент.