**Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"**

Факультет социальных наук
Департамент социологии

Рабочая программа дисциплины
Data Mining в социальных науках: методы факторизации и кластеризации

для образовательной программы «Социология»
направления подготовки 39.03.01
бакалавр

Разработчик программы:
А.Н. Ротмистров, к.с.н, доцент
alexey.n.rotmistrov@gmail.com

Одобрена на заседании департамента социологии «___»_____ 2018 г.
Руководитель департамента А.Ю. Чепуренко _____

Рекомендована Академическим советом образовательной программы бакалавриата «Социология» «___»_____ 2018 г., № протокола_____

Утверждена «___»_____ 2018 г.

Академический руководитель образовательной программы бакалавриата «Социология»
О.Е. Кузина _____

Москва, 2018

Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы

1. Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчётности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направления 39.03.01 «Социология» подготовки бакалавра (4 курс), изучающих дисциплину «Data Mining в социальных науках: методы факторизации и кластеризации».

Программа разработана в соответствии с:

- Образовательным стандартом ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО АВТОНОМНОГО ОБРАЗОВАТЕЛЬНОГО УЧРЕЖДЕНИЯ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ». http://www.hse.ru/data/2015/04/09/1095226565/Бакалавриат_ОС_Социология.pdf.
- Образовательной программой 39.03.01 «Социология» подготовки бакалавра
- Учебным планом на весь период обучения подготовки бакалавра 2015-2019 (4 года, очная форма) по направлению 39.03.01 СОЦИОЛОГИЯ, утверждённым в 2015 г.
- Рабочим учебным планом университета по направлению 39.03.01 «Социология» подготовки бакалавра, утверждённым в 2017 г.

2. Цели и задачи дисциплины

Учебная дисциплина нацелена на формирование у студентов разносторонних компетенций в области факторизации и кластеризации переменных и объектов в «трудных» случаях: случаи категориальных переменных (номинальных, порядковых и бинарных) и больших данных (Big Data, далее – BD). При этом один из треков дисциплины предполагает в качестве основного инструмента сбора и анализа данных программную среду Python.

Достижение обозначенной цели предлагается осуществить через решение следующих задач:

а. Сопоставление современных парадигм анализа данных: «конфирматорной», Theory-Driven, генетически связанной со статистикой и эконометрикой, и «эксплораторной», Data-Driven, генетически связанной с компьютерными науками и с идеями Дж. О. Юла, Э. Симпсона, Дж. Тьюки и И. Пятецкого-Шапиро. Обоснование логической связи «эксплораторной» парадигмы с явлением BD;

б. Рассмотрение структуры Data Mining (далее – DM) в социальных науках: как первичного нереактивного измерения, как идеологии работы с собранными данными, как техники углублённого «понимания» респондентов в контексте массовых опросов, как интеллектуального анализа данных;

в. Изучение и сравнение видов «оцифровки» (оптимального шкалирования): оцифровка, основанная на линейности связи, на сходстве распределений, на совпадениях категорий;

г. Рассмотрение возможностей инферентных (предполагающих проверку нулевых гипотез) выводов из эксплораторных методов;

д. Знакомство с модулями Python, предназначенными для решения релевантных задач web-scraping (нереактивного измерения) и «обучения без учителя» (факторизации и кластеризации).

Учебная дисциплина предназначена для студентов 4 курса направления подготовки «Социология».

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения материала данной дисциплины студенты:

– **научатся** ориентироваться в многомерном пространстве методов анализа латентных переменных (формирующихся в результате применения алгоритмов «обучения без учителя»), выбирать среди них наиболее подходящий для каждой конкретной задачи, корректно применять его, грамотно интерпретировать результаты и формулировать выводы, полезные для их проектной, консалтинговой или аналогичной деятельности;

– **овладеют** методами нереактивного измерения (сбора данных), в т.ч. предполагающего дальнейшую аналитическую траекторию в русле факторизации и/или кластеризации;

– **расширят свои знания и навыки** применения статистических и программных сред Python, MS Excel и SPSS.

Тем самым у них будут развиты следующие **компетенции**:

Компетенция	Код по ФГОС/НИУ	Дескрипторы - основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
способность к восприятию, обобщению, анализу информации, постановке цели и выбору путей её достижения	ОК-1, продвинутый уровень	студент способен выбрать из нескольких близких методов наиболее подходящий каждой конкретной задаче и обосновать свой выбор, также способен выбрать из нескольких адекватных решений оптимальное и обосновать свой выбор	предоставление студентам схемы методов DM; практика по выбору методов, адекватных поставленной задаче и постановки задач под конкретные методы на практических занятиях, в рамках домашних и контрольной работ
владение основными методами, способами и средствами получения, хранения, переработки информации, навыки работы с компьютером как средством управления информацией	ОК-13, продвинутый уровень	студент способен быстро и правильно выбирать и применять нужные для решения поставленной задачи пакеты в программных средах MS Excel, SPSS, Python	регулярная практика пользования перечисленными программными и статистическими средами
способность самостоятельно формулировать цели, ставить конкретные задачи научных	ПК-2, продвинутый уровень	студент способен выстроить цепочку: содержательная задача -	практика по выбору методов, адекватных поставленной

исследований в различных областях социологии и решать их с помощью современных исследовательских методов с использованием новейшего отечественного и зарубежного опыта и с применением современной аппаратуры, оборудования, информационных технологий		> метод анализа -> нужные для решения поставленной задачи методов DM	задаче, и постановке задач под конкретные методы на практических занятиях, в рамках домашних и контрольной работ
умение обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций	ПК-8, продвинутый уровень	студент способен выстроить цепочку: результаты применения методов DM -> решение содержательной задачи -> ясная развёрнутая интерпретация -> практические рекомендации	практика решения поставленных преподавателем и самостоятельно сформулированных задач; ясная развёрнутая интерпретация результатов и предложение практических рекомендаций на практических занятиях, в рамках домашних и контрольной работ
способность использовать базовые теоретические знания, практические навыки и умения для участия в научных и научно-прикладных исследованиях, аналитической и консалтинговой деятельности	ПК-10, продвинутый уровень	студент способен самостоятельно соотносить имеющуюся теоретическую или практическую проблему с располагаемым арсеналом методов	изложение краткой истории и теории DM

4. Место дисциплины в структуре ООП

Дисциплина относится к вариативной части профессионального цикла (Б.ПЦ.В.ДВ.).

Изучение курса предполагает предварительное освоение материала учебных дисциплин:

- алгебра и анализ
- прикладное программное обеспечение
- теория вероятностей и математическая статистика
- анализ данных в социологии
- методология и методы социологического исследования
- анализ и визуализация данных в R
- анализ и визуализация данных в Python

Из материала этих дисциплин ключевыми для освоения методов анализа латентных признаков являются:

- знания о том,…

… что такое случайная величина и каковы основные законы её распределения;

… что такое статистическая гипотеза;

… каковы типы шкал социологического измерения и их основные свойства;

… что такое парная и множественная связь и какими методами она измерима;

… что такое язык программирования Python и какие структуры данных он поддерживает, каковы основные опции статистических приложений MS Excel и SPSS;

… что такое нереактивные и большие данные;

- умение…

… оценивать законы распределения случайных величин;

… проверять статистические гипотезы;

… измерять парную и множественную связи;

… строить простые логические и математические модели для решения содержательных задач в области анализа данных;

… писать простейшие коды на языке Python, в т.ч. обращаясь и к MS Excel, применять основные аналитические опции статистического приложения SPSS.

Дисциплины, для которых данная дисциплина является предшествующей: отсутствуют.

5. Тематический план учебной дисциплины

№ п.	Название темы	Всего часов по дисциплине	Аудиторные часы		Самосто-ятельная работа
			Лекции	Сем. и практ. занятия	
1	Основы современного DM в социальных науках: эксплораторность, BD, программирование	7,6	0,5	1,5	5,6
2	Основы web-scraping: программные среды и специфика социальных BD. Релевантные модули и фреймворки на языке Python: requests, BeautifulSoup, scrapy	30,4	2,0	6,0	22,4
3	Роль методов факторизации и кластеризации в системе DM: сжать объём информации для формирования и корректировки аналитической траектории. Новый взгляд на применение методов главных компонент и кластеризации К-средних. Релевантные модули Python: sklearn, statsmodels, prince	7,6	0,5	1,5	5,6
4	Роль оцифровки в системе DM: факторизация и кластеризация категориальных переменных и «понимающая» парадигма в количественном воплощении	15,2	1,0	3,0	11,2
5	Оцифровка, основанная на линейности связи: категориальный метод главных компонент, и метод кластеризации Cosine. Релевантные модули Python: sklearn, mca, prince	30,4	2,0	6,0	22,4
6	Оцифровка, основанная на сходстве распределений: методы анализа соответствий и 2-ступенчатого кластерного анализа. Релевантные	30,4	2,0	6,0	22,4

	модули Python: mca, prince				
7	Оцифровка, основанная на совпадениях категорий: методы Jaccard, Yule's, Phi, Lambda и пр. Релевантные модули Python: sklearn	30,4	2,0	6,0	22,4
	Итого:	152,0	10,0	30,0	112,0

6. Формы контроля знаний студентов

Тип контроля	Форма контроля	Неделя					Параметры **
		2	4	6	8	10	
Текущий	Оценка активности на занятиях	еженедельно					Ответы на вопросы
	Домашняя работа	*	*	*	*	*	Применение методов в программных средах MS Excel, SPSS, Python для решение содержательных задач, интерпретация результатов – дома
	Контроль-ная работа				*		Применение методов в программных средах MS Excel, SPSS, Python для решение содержательных задач, интерпретация результатов – в классе
Итоговый	Экзамен					*	Ответы на вопросы в тестовой форме

6.1 Критерии оценки знаний, навыков

Оценки по всем формам текущего контроля, кроме оценки активности на занятиях выставляются по 10-ти балльной шкале. Суммарная оценка за активность на занятиях **может превышать 10 и тем самым отчасти компенсировать недоработки студентов по другим составляющим итоговой оценки.**

Контрольная работа. Студент должен правильно подобрать метод под каждое задание (ОК-1, продвинутый уровень), применить его в подходящем статпакете (ОК-13, продвинутый уровень, ПК-8, продвинутый уровень) и проинтерпретировать результаты в соответствующем поле листа с заданиями.

Домашняя работа. Студент должен выполнить требования, сформулированные в п.9.1. ОК-1, продвинутый уровень; ОК-13, продвинутый уровень; ПК-2, продвинутый уровень; ПК-8, продвинутый уровень; ПК-2, продвинутый уровень

Экзаменационный тест по обязательной литературе. Студент должен выбрать правильный(/е) ответ(ы) каждый вопрос (ОК-1, продвинутый уровень) и отметить его.

6.2 Порядок формирования итоговой оценки по дисциплине

- активность на занятиях; оценка выставляется преподавателем на каждом занятии за обоснованные ответы на его вопросы (не более 3 баллов за занятие); **суммарное значение может превышать 10 и тем самым отчасти компенсировать недоработки студентов по другим составляющим итоговой оценки:** *0,1 итоговой оценки*
- контрольная работа в формате решения практических задач: *0,3 итоговой оценки*
- домашние работы в формате решения практических задач: *0,3 итоговой оценки*

- экзамен в формате теста со множественным выбором; **студент может не приходить на экзамен, если его устраивает итоговая**: *0,3 итоговой оценки*

7. Содержание дисциплины

Раздел 1. Введение

Основы современного DM в социальных науках: эксплораторность, BD, программирование.

Парадигмы анализа данных («конфирматорная» и «эксплораторная»): история конкуренции и современное состояние. Идеи Дж. О. Юла, Э. Симпсона, Дж. Тьюки и И. Пятецкого-Шапиро.

Влияние на анализ данных появления нереактивных данных и BD.

Перенос подходов DM из компьютерных наук в социальные. Определение DM в социальных науках:

- DM как первичное нереактивное измерение – прежде всего в online-среде (web-scraping);
- DM как идеология работы с собранными данными, лежащая в основе дизайна исследования data-driven в противовес дизайну theory-driven;
- DM как техники углублённого «понимания» респондентов в контексте массовых опросов. Под «пониманием» респондентов в массовых опросах подразумевается адаптация исследовательских шкал под имеющиеся ответы респондентов для более насыщенной интерпретации этих ответов. Оцифровка – одна из процедур такой адаптации;
- DM как интеллектуальный анализ данных, т.е. анализ, предполагающий разработку уникального методно-аналитического комплекса под каждую сложную аналитическую задачу. Уникальность состоит в учёте множества логических развилок на пути решения сложной задачи и обоснованном выборе более подходящей ветви в рамках каждой развилки; в возврате к предыдущей развилке в случае неудовлетворительно результата ранее выбранной развилки; в итеративности таких выборов и возвратов. Под сложной задачей мы понимаем задачу, включающую много содержательных слоёв, или уровней, каждый из которых требует содержательной интерпретации.

Основные источники

1. Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery: an overview. In: Fayyad, U.M. et al (Eds.) Advances in Knowledge Discovery and Data Mining, pp. 1-34. Menlo Park, CA, USA: AAAI/MIT
2. Morgan, J. N. and J. A. Sonquist (1963). Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association 58, 415–434.
3. Piatetsky-Shapiro, G. (1991) Knowledge discovery in real databases: A report on the IJCAI-89 workshop Al Magazine, 11 (5), pp. 68-70
4. Ratner, B. (2012) Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, CRC Press
5. Tukey, J.W. (1962) The future of data analysis, Annals of Mathematical Statistics, 33, 1-67.
6. Tukey, J.W. (1977) Exploratory Data Analysis, Addison-Wesley, Reading, MA

Дополнительные источники

1. Fabris, C. C., & Freitas, A. A. (2006). Discovering Surprising Instances of Simpson's Paradox in Hierarchical Multidimensional Data. International Journal of Data Warehousing and Mining (IJDWM), 2(1), 27-49

2. Goltz, Heather Honoré & Matthew Lee Smith (2010). Yule-Simpson's Paradox in Research. Practical Assessment, Research & Evaluation, 15(15)

3. Kievit, R., Frankenhuis, W., Waldorp, L. and Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. Frontiers in Psychology, 4

4. Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, 13(2), 238-241

5. Мальцева А. В., Шилкина Н. Е., Махныткина О. В. Data mining в социологии: опыт и перспективы проведения исследования // Социологические исследования. 2016. №3. С. 35-44

6. Толстова Ю. Н. Анализ социологических данных. М.: Научный мир, 2000

7. Толстова Ю. Н. Математико-статистические модели в социологии. М.: Издательский дом ГУ-ВШЭ, 2008

Раздел 2. Web-scraping и Big Data. Релевантные модули и фреймворки на языке Python: requests, BeautifulSoup, scrapy

Тема 2. Основы web-scraping: программные среды и специфика социальных BD. Знакомство с возможностями программной среды Python для web-scraping и видами данных в онлайн-среде; пробное применение методов web-scraping для сбора данных. Специфика применения методов DM к социальным BD по сравнению с ситуацией в компьютерных науках:

- в социальных науках при обычном (реактивном) сборе данных в большей мере используются категориальные шкалы, тогда как BD обычно организованы в формате метрических и бинарных переменных. Однако в социальных BD тоже есть место для категориальных шкал: категориальными шкалами обычно выражается экспертная оценка, с которой как с эталоном машина соотносит собираемые данные;
- в компьютерных науках модели должны иметь значения метрик качества (например, коэффициент детерминации или показатель Loss в задачах регрессии, доля правильных предсказаний в задачах классификации и т.п.) не ниже 0,9 из 1. Такая объяснительная способность достигается даже в ущерб непереобученности модели. Это не удивительно, т.к., во-первых, BD в компьютерных науках – это обычно и есть генеральная совокупность, и, во-вторых, практические задачи в компьютерных науках требуют крайне высокой точности: скажем, распознать образ на 50% – это не распознать его вовсе. Обратной стороной смещения баланса между объяснительной способностью и непереобученностью модели является риск неработоспособности или невоспроизводимости модели за пределами BD, на которых она построена;
- в компьютерных и социальных науках разный язык касательно области анализа данных.

Основные источники

1. Майер-Шенбергер, В., Кукьер В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Манн, Иванов и Фербер. М. 2014

2. Golder, S. A., and Macy, M. W. Digital footprints: Opportunities and challenges for online social research. Annual Review of Sociology, 2014, 40(1), 129-152

3. Kosinski M., Stillwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior. PNAS, 2013, 110 (15), 5802-5805

Дополнительные источники

1. Hasler B. S., Tuchman P., & Friedman D. Virtual research assistants: Replacing human interviewers by automated avatars in virtual worlds. Computers in Human Behavior, 2013, 29, 1608-1616

2. Lazer D., Kennedy R., King G., and Vespignani A. (2014). The parable of Google flu: Traps in big data analysis, Science, 343. P. 1203-1205

3. Lazer, D., & Radford, J. Data ex Machina: Introduction to big data. Annual Review of Sociology, 2017, 43(1), 7.1-7.21

4. Ruths D, and Pfeffer J. (2015). Social media for large studies of behavior. Science, 346 (6213). P. 1063-1064

5. Домингос, П. Верховный алгоритм: как машинное обучение изменит наш мир / пер. с англ. В. Горохова – М.: Манн, Иванов и Фербер, 2016.

Раздел 3. Оцифровка, факторизация и кластеризация. Релевантные модули Python: sklearn, statsmodels, mca, prince

Тема 3. Роль методов факторизации и кластеризации в системе DM: сжать объём информации для формирования и корректировки аналитической траектории. Методы кластеризации (объединение многих объектов в немногие однородные группы, описываемые значениями интегральной переменной) – разновидность методов факторизации (объединение многих переменных в немногие однородные группы, описываемые соответствующим числом интегральных переменных), для которой характерна единственная интегральная (латентная) переменная категориального типа. Методы факторизации сформировались на основе идей шкалирования – эксплораторных по своей сути. До сих пор большинство методов факторизации и кластеризации не предполагают инферентные (основанные на проверке статистических гипотез) выводы и в этом смысле выполняют сугубо разведочную роль. Эту роль нельзя переоценить, поскольку в русле DM исследователь принимает решения о формировании и корректировке аналитической траектории преимущественно на основании информации обозримого объёма, а методы факторизации и кластеризации как раз и направлены на сжатие объёма информации. С другой стороны, следует постепенно вводить в состав многих методов факторизации и кластеризации алгоритмы инферентной оценки получаемых моделей для более надёжной их оценки, что не противоречит идеям эксплораторной парадигмы Дж. Тьюки. В качестве примера предлагается рассмотреть введение регрессионного моделирования и критерия Lack-of-Fit для оценки качества моделей, построенных методом главных компонент и методом кластеризации К-средних. Эти методы выбраны как наиболее популярные в социальных науках методы факторизации и кластеризации соответственно.

Тема 4. Роль оцифровки в системе DM: факторизация и кластеризация категориальных переменных и «понимающая» парадигма в количественном воплощении. Оцифровка – одна из процедур адаптации исследовательских шкал под имеющиеся ответы респондентов для более насыщенной интерпретации этих ответов. Одновременно оцифровка – основа методов факторизации и кластеризации категориальных переменных. Изучение и сравнение видов оцифровки: оцифровка, основанная на линейности связи (пример – категориальный метод главных компонент), оцифровка, основанная на сходстве распределений (пример – метод анализа соответствий и метод 2-ступенчатого кластерного анализа), оцифровка, основанная на совпадениях категорий (пример – методы Jaccard, Yule’s, Phi, Lambda).

4.1. Оцифровка, основанная на линейности связи: категориальный метод главных компонент (далее – CatPCA), и метод кластеризации Cosine.

CatPCA – метод факторизации набора наблюдаемых переменных (далее – переменных-индикаторов) и их значений, хотя бы одна из которых является порядковой, а остальные могут быть порядковыми или номинальными. Содержательно факторизация может быть симметрична относительно индикаторов или происходить в контексте одного

из индикаторов (факторизующего). Эти содержательные подходы влияют только на интерпретацию результатов.

Модельные предположения метода: а) изучаемые индикаторы по своей природе континуальны, даже если измерены категориально. Гипотетическая причина наличия «категориальной ипостаси» индикаторов – «грубость» социологического инструментария. Одна из задач CatPCA – восстановить эту континуальность. б) Поскольку прямолинейные связи среди всех видов связи для человеческого восприятия самые простые, то CatPCA как будто «выпрямляет» связь любого вида.

CatPCA НЕ подвержен следующим ограничениям метода главных компонент (далее – МГК):

- тип шкалы переменных-индикаторов не ниже интервального;
- форма связи между переменными-индикаторами – прямолинейная;
- форма связи между переменными-индикаторами и компонентами – прямолинейная.

Для CatPCA предпочтительны переменные-индикаторы рангового типа. Латентные переменные в CatPCA интервального типа.

CatPCA = оцифровка исходных индикаторов + МГК.

Суть оцифровки: категории переменной-индикатора выстраиваются таким образом, чтобы максимизировать её связь с фактором/ами. Оцифровка в CatPCA может производиться посредством работы нейросети.

Выбор между МГК и CatPCA

- Ориентируясь только на тип переменных-индикаторов;
- Ориентируясь только на тип переменных-индикаторов и на их совместное «поведение» посредством, например, критерия Lack-of-Fit.

Критерии качества моделей CatPCA: Альфа Кронбаха, Variance Accounted for, Component Loadings, линейная регрессия, регрессионные остатки и критерий Lack-of-Fit.

Cosine – метод расчёта расстояний между объектами для последующей их **кластеризации**. Число кластеров определяется по скачку критерия.

Прочие распространённые методы кластеризации:

- большинство методов кластеризации основаны на координатных расстояниях (Евклидовом и его обобщениях);
- другая часть методов кластеризации основана на сходстве профилей объектов по кластеризующим переменным;
- координатные расстояния предназначены для поиска концентрированных/центроидных кластеров (варианты центроидов: среднее арифметическое, медиана, мода и т.п.).

Cosine уникален:

- измеряет расстояние как угол между векторами, задаваемыми координатами объектов по каждой кластеризующей переменной;
- угловое расстояние предназначено для поиска линейно-функциональных вытянутых кластеров.

Для Cosine требуются кластеризующие переменные интервального типа. Латентная переменная в Cosine номинального типа.

Суть оцифровки: объекты объединяются в линейно-функциональные группы, благодаря чему объединяются профили, соответствующие этим объектам, т.е. объединяются именно сочетания категорий кластеризующих переменных при том, что взятые по отдельности эти категории могут быть очень далеки друг от друга (если смотреть на них с т.з. координатных расстояний). Оцифровка в Cosine может производиться посредством работы нейросети, но эта идея ещё не реализована.

Выбор в пользу Cosine

- по типу кластеризующих переменных (интервальный);
- по графику рассеяния (если кластеризующих переменных две и объекты формируют вытянутые группы);
- по теоретической рамке;
- нужно разрабатывать дополнительные критерии.

Критерии качества моделей Cosine: линейная регрессия, регрессионные остатки и критерий Lack-of-Fit.

4.2. Оцифровка, основанная на сходстве распределений: методы анализа соответствий (далее – CorrAn) и 2-ступенчатого кластерного анализа (далее – 2SCA).

CorrAn предназначен для факторизации категорий переменных-индикаторов и, как следствие, самих индикаторов. Содержательно факторизация может быть симметрична относительно индикаторов или происходить в контексте одного из индикаторов (факторизующего индикатора). Эти содержательные подходы влияют на выбор алгоритма нормализации инерции и на интерпретацию результатов.

Модельное предположение метода: изучаемые индикаторы по своей природе континуальны, даже если измерены категориально. Т.е. предполагается, что обобщённый респондент, выбирая категории некой категориальной переменной, руководствуется имеющимся в его сознании небольшим числом латентных континуальных переменных. Это может быть объяснено через наличие в сознании человека множества переплетённых ассоциаций, благодаря которым категории почти любой категориальной переменной переплетены с категориями множества континуальных переменных. Посредством этих ассоциативных связей категории изучаемой категориальной переменной оказывается возможным встроить в некоторый континуум.

CorrAn НЕ подвержен никаким ограничениям МГК, кроме одного: между переменными-индикаторами должна быть связь (любого типа). Кроме того, у CorrAn есть специфическое ограничение: индикаторов всего два.

CorrAn актуален для любых категориальных переменных-индикаторов (при этом отношение порядка между категориями – если такое имеется – не учитывается методом). Латентные переменные в CorrAn интервального типа.

Критерии качества моделей CorrAn: общая объяснённая инерция, Total Contribution of Dimensions to Inertia of Point, Contribution of Point to Inertia of Dimensions, линейная регрессия, регрессионные остатки и критерий Lack-of-Fit. Важная уникальная характеристика – «расплывчатость» координат категории.

Метод множественного анализа соответствий (далее – **MultiCorrAn**) предназначен для факторизации набора переменных-индикаторов и их значений; он существенно отличается от 2-мерного анализа соответствий. Содержательно факторизация может быть симметрична относительно индикаторов или происходить в контексте одного из индикаторов (факторизующего индикатора). Эти содержательные подходы влияют только на интерпретацию результатов.

Модельные предположения метода объединяют в себе предположения из CorrAn и CatPCA: а) изучаемые индикаторы по своей природе континуальны, даже если измерены категориально. Т.е. предполагается, что обобщённый респондент, выбирая категории некой категориальной переменной, руководствуется имеющимся в его сознании небольшим числом латентных континуальных переменных, что объясняется через наличие в сознании человека множества переплетённых ассоциаций, благодаря которым категории почти любой категориальной переменной переплетены с категориями множества континуальных переменных. Посредством этих ассоциативных связей

категории изучаемой категориальной переменной оказывается возможным встроить в некоторый континуум. б) Поскольку прямолинейные связи среди всех видов связи для человеческого восприятия самые простые, то MultiCorrAn как будто «выпрямляет» связь любого вида.

Также MultiCorrAn отличается от CorrAn бОльшим набором критериев качества моделей: Loss, общая объяснённая инерция, Total Contribution of Dimensions to Inertia of Point, Альфа Кронбаха (>0), Contribution of Point to Inertia of Dimensions, линейная регрессия, регрессионные остатки и критерий Lack-of-Fit.

2SCA – метод **кластеризации** объектов. Число кластеров определяется по скачку одного из критериев.

Большинство методов кластеризации «работают» с интервальными кластеризующими переменными и основаны на координатных расстояниях (Евклидовом и его обобщениях).

2SCA интересен тем, что:

- может «работать» и с категориальными кластеризующими переменными;
- может использовать расстояния, основанные на сходстве распределений.

Для 2SCA желательно, чтобы:

- кластеризующие признаки не были связаны между собой;
- форма распределения интервальных кластеризующих признаков была нормальной, а категориальных – мультиномиальной.

Другие ограничения:

- кластеры, построенные на сходстве распределений, бывает трудно интерпретировать;
- неустойчивость некоторых моделей к порядку респондентов (особенно при использовании мер расстояний, основанных на координатах).

Латентная переменная в 2SCA номинального типа.

Суть оцифровки: объекты объединяются сначала согласно профилям, составленным из значений кластеризующих переменных (первая ступень), затем пошагово объединяются всё менее похожие друг на друга (по координатам или по формам распределения) профили (вторая ступень). Т.е. снова речь идёт в конечном счёте о трансформации (оцифровке) категорий согласно их профилям.

Выбор в пользу 2SCA:

- по графику рассеяния (если кластеризующих переменных две и объекты формируют вытянутые группы);
- по теоретической рамке;
- нужно разрабатывать дополнительные критерии.

Критерии качества моделей 2SCA: BIC, AIC, линейная регрессия, регрессионные остатки и критерий Lack-of-Fit.

4.3. Оцифровка, основанная на совпадениях категорий: методы Jaccard, Yule’s, Phi, Lambda и пр.

Методы кластеризации. Число кластеров определяется по скачку одного из критериев.

Большинство методов кластеризации «работают» с интервальными кластеризующими переменными и основаны на координатных расстояниях (Евклидовом и его обобщениях)

Методы кластеризации на бинарных переменных (далее – МКБП) интересны тем, что:

- могут «работать» с бинарными кластеризующими переменными;

- могут использовать расстояния, основанные на сходстве распределений.

Ограничения:

- кластеры, построенные на сходстве распределений, бывает трудно интерпретировать;
- неустойчивость некоторых моделей к порядку респондентов (особенно при использовании мер расстояний, основанных на координатах).

Латентная переменная в МКБП номинального типа.

Суть оцифровки (принцип такой же, как в 2SCA): объекты объединяются сначала согласно профилям, составленным из значений кластеризующих переменных (первая ступень), затем пошагово объединяются всё менее похожие друг на друга (по координатам или по формам распределения) профили (вторая ступень). Т.е. снова речь идёт в конечном счёте о трансформации (оцифровке) категорий, согласно их профилям.

Меры расстояний для МКБП могут быть систематизированы по 5 основаниям, отталкиваясь от конфигурации 4-клеточной таблицы частот совпадений и несовпадений нулей и единиц для любых двух объектов:

1. несходство/сходство;
2. совпадения/несовпадения;
3. диагональ/ячейка;
4. учёт асимметричности;
5. наличие границ.

Выбор в пользу МКБП: по типу кластеризующих переменных (бинарный). Нужно разрабатывать дополнительные критерии. Критерии качества моделей МКБП: линейная регрессия, регрессионные остатки и критерий Lack-of-Fit.

Основные источники

1. Калинина В.Н. Соловьев В.И. Введение в многомерный статистический анализ. М., 2003. URL: http://www.allmath.ru/highermath/probability/probability1/probability.htm
2. Клигер С. А., Косолапов М. С., Толстова Ю. Н. Шкалирование при сборе и анализе социологической информации. М., 1978.
3. Bacher, J., Wenzig, K., Vogler, M. SPSS Two Step Cluster – A First Evaluation. Universitat Erlangen-Nurnberg. URL: http://www.statisticalinnovations.com/products/twostep.pdf
4. Benzécri, J. P. 1992. Correspondence analysis handbook. New York: Marcel Dekker.
5. Di Franco G. Multiple correspondence analysis: one only or several techniques? // Qual. Quant. 2016. Vol. 50. No 3. Pp. 1299–1315.
6. Di Franco G., Marradi A. Factor analysis and principal component analysis. FrancoAngeli, 2013.
7. Diamantaras, Konstantinos I., and Sun Yuan Kung. Principal component neural networks: theory and applications. John Wiley & Sons, Inc., 1996
8. Greenacre, Michael. Correspondence analysis in practice. CRC press, 2017
9. Heiser, W.J., and J.J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: Correspondence Analysis in the Social Sciences: Recent Developments and Applications, M. Greenacre, and J. Blasius, eds. New York: Academic Press
10. Le Roux, Brigitte, and Henry Rouanet. Multiple correspondence analysis. Vol. 163. Sage, 2010
11. Meulman, J.J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical dat A.Journal of Classification, 13, 249-266

12. Meulman, J.J., A.J. Van der Kooij, and A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: Classification, Automation and New Media, W. Gaul, and G. Ritter, eds. Berlin: Springer-Verlag

13. Meulman, J.J., A.J. Van der Kooij, and W.J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: Handbook of Quantitative Methodology for the Social Sciences, D. Kaplan, eds. Thousand Oaks, Calif.: Sage Publications, Inc.

14. Norusis, M. SPSS 19.0 Statistical Procedures Companion. Ch. 19. URL: http://www.norusis.com/pdf/SPC_v19.pdf

15. Takane, Yoshio. Constrained principal component analysis and related techniques. CRC Press, 2013

16. Tenenhaus, Michel, and Forrest W. Young. "An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data." Psychometrika 50, no. 1 (1985): 91-119. Van Rijckevorsel, J. 1987. The application of fuzzy coding and horseshoes in multiple correspondence analysis. Leiden: DSWO Press

17. Wang, Haixun, Wei Wang, Jiong Yang, and Philip S. Yu. "Clustering by pattern similarity in large data sets." In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pp. 394-405. ACM, 2002

18. Young, F.W., Y. Takane, and J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. Psychometrika, 43, 279-281

Дополнительные источники

1. Applications of latent trait and latent class models in social sciences / Ed. by J. Rost and R. Langeheine. Münster, New York, München, Berlin: Waxmann, 1997. URL: http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

2. Gilula, Z., and S.J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. Journal of the American Statistical Association, 83, 760-771

3. Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: The Prediction of Personal Adjustment, P. Horst, eds. New York: Social Science Research Council

4. Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: International Encyclopedia of Statistics, W. H. Kruskal, and J. M. Tanur, eds. New York: The Free Press

5. Meulman, J.J., and P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. Psychometrika, 58, 7-35

6. Meulman, J.J., and W.J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: Visual Display of Categorical Data, M. Greenacre, and J. Blasius, eds. New York: Academic Press

8. Образовательные технологии

Программные среды MS Excel, Python и SPSS.

9. Задания для оценки качества освоения курса

Применяются многовариантные задания (в формате Excel) для оценки качества освоения методов и тестовые задания для оценки качества освоения литературы.

10. Учебно-методическое и информационное обеспечение дисциплины

Базовые источники:

1. Greenacre, Michael. Correspondence analysis in practice. CRC press, 2017
2. Kosinski M., Stillwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior. PNAS, 2013, 110 (15), 5802-5805
3. Meulman, J.J., A.J. Van der Kooij, and W.J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: Handbook of Quantitative Methodology for the Social Sciences, D. Kaplan, eds. Thousand Oaks, Calif.: Sage Publications, Inc.
4. Ratner, B. (2012) Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, CRC Press
5. Официальная документация для используемых модулей Python:

- http://www.statsmodels.org/ (для модуля statsmodels),
- http://scikit-learn.org/stable/ (для модуля sklearn),
- https://github.com/MaxHalford/prince (для модуля prince),
- https://github.com/esafak/mca (для модуля mca),
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (для модуля BeautifulSoup),
- http://docs.python-requests.org/en/master/ (для модуля requests),
- https://docs.scrapy.org/en/latest/ (для фреймворка scrapy).

Основные и **дополнительные** источники приведены отдельно для каждого раздела.

Большинство публикаций можно найти в:

- библиотеке НИУ ВШЭ,
- электронном каталоге НИУ ВШЭ,
- базе научных публикаций Google Scholar

Базы данных, информационно-справочные и поисковые системы:

- единый архив экономических и социологических данных: http://sophist.hse.ru/
- портал Организации экономического сотрудничества и развития (OESD): http://stats.oecd.org/
- портал European Social Survey (ESS): http://www.ess-ru.ru/
- портал International Social Survey Programme (ISSP): http://www.issp.org/
- портал Russian Longitudinal Monitoring Survey: http://www.cpc.unc.edu/projects/rlms-hse
- портал World Values Survey: http://www.worldvaluessurvey.org/
- федеральный информационный портал «Экономика. Социология. Менеджмент»: http://ecsocman.hse.ru/
- и другие

11. Материально-техническое обеспечение дисциплины

Персональный компьютер на каждого студента, 1 проектор, 1 экран, 1 доска и маркеры.