ИНСТИТУТ СИСТЕМАТИКИ И ЭКОЛОГИИ ЖИВОТНЫХ СОРАН
ЛАБОРАТОРИЯ ЗООЛОГИЧЕСКОГО МОНИТОРИНГА

ИННОВАЦИОННЫЙ ЦЕНТР ЗАЩИТЫ РАСТЕНИЙ (ВИЗР РАСХН)

В.М. Ефимов, В.Ю. Ковалева

МНОГОМЕРНЫЙ АНАЛИЗ БИОЛОГИЧЕСКИХ ДАННЫХ

Учебное пособие

2-е исправленное и дополненное издание

Санкт-Петербург - 2008

В.М. Ефимов, В.Ю. Ковалева.
МНОГОМЕРНЫЙ АНАЛИЗ БИОЛОГИЧЕСКИХ ДАННЫХ
УЧЕБНОЕ ПОСОБИЕ

Рецензент:
Ю.С.Равкин, заслуженный деятель науки РФ, д.б.н., проф., зав. лабораторией зоологического мониторинга Института систематики и экологии животных СО РАН

В пособии рассмотрены многомерные методы исследования массовых биологических процессов и явлений: метод главных компонент, факторный анализ, дискриминантный анализ, регрессионные методы, многомерное шкалирование, нейронные сети. Основной упор делается на линейные и нелинейные методы анализа взаимного расположения объектов в многомерном пространстве и интерпретацию получаемых результатов с биологической точки зрения.

Пособие предназначено для научных работников и студентов биологических специальностей.

СОДЕРЖАНИЕ

ЛЕКЦИЯ 1. Введение

Исследования в области биологии неизбежно связаны с биологическими объектами. В качестве объектов можно рассматривать особи, популяции, сообщества, их состояния, динамику, поведение и другие характеристики. Каждый объект обладает набором свойств. В некотором смысле каждый объект является совокупностью своих свойств. Однако нас интересуют не все свойства, а только те, по которым объекты различаются между собой, формируя биологическое разнообразие. Если свойств много, то возникает необходимость в специальных методах изучения, позволяющих охватить сразу всю их совокупность.

Одним из выдающихся достижений научного естествознания прошлого тысячелетия является разработка и строгая формализация понятия метрического пространства и его размерности (Александров, 1987). Являясь абстрактным обобщением окружающего нас и доступного нашим органам чувств трехмерного физического пространства, оно позволяет представлять точками многомерного пространства объекты любой природы. Сходство между объектами отображается расстоянием в многомерном пространстве. Тем самым создается возможность получать глубокие содержательные результаты, исследуя геометрические и динамические свойства взаимного расположения точек и переводя их на язык соответствующей области знаний (Пуанкаре, 1983). Кроме того, подобным образом можно прослеживать параллели и искать структурное единство в очень далеких друг от друга научных областях, так как они могут быть описаны одним и тем же математическим аппаратом. Физики с большим успехом реализуют подобную программу, начиная с середины XIX века, создав, среди всего прочего, теорию относительности и квантовую механику (Фейнман и др., 1978; Дирак, 1990). О будущей геометризации биологии писал В.И.Вернадский (1975).

Основными понятиями многомерного анализа являются понятия пространства, его размерности и взаимного расположения объектов в этом пространстве, а также расстояния или сходства между его объектами. В многомерном евклидовом пространстве расстояние между двумя любыми объектами полностью определяется теоремой Пифагора: квадратный корень из суммы квадратов разностей между значениями координат:

$$d = \sqrt{\sum (x_i - y_i)^2}$$

В биологических исследованиях невозможно обойтись без понятия выборки. Если предполагается, что выборка извлечена из генеральной совокупности объектов, распределенных в этом же пространстве в соответствии с некоторым вероятностным законом, то мы имеем дело с многомерным статистическим анализом.

Понятие пространства и его размерности пронизывает практически всю математику от древности до наших дней. Уже в аксиомах Евклида (III век до н.э.) в качестве первичных сущностей приняты точка, линия, плоскость, пространство, отражающие основные геометрические свойства окружающего нас трехмерного мира. Все остальные свойства строго логически выводятся из аксиом. В прикладных науках, астрономии и географии, (но не в самой геометрии) не позже III–II веков до н.э. сформировались различные системы координат (Эратосфен, Гиппарх, Птолемей).

Система Евклида была настолько совершенна, что вплоть до XIX века

служила образцом интеллектуальных возможностей чистого разума. Одновременно она же была эталоном соответствия математики и реального мира – все ее утверждения немедленно могли быть подтверждены практикой. Собственно, никаких сомнений и не могло возникнуть, изначально предполагалось, что ее идеальные конструкции и лежат в основе реального мира, а возможные несоответствия вызваны исключительно неточностью измерений.

В XIX веке произошло невероятное событие – была открыта другая геометрия. Лобачевский, взяв за основу аксиомы Евклида и заменив постулат о параллельных на противоположный, построил геометрию, логически равноправную с геометрией Евклида, но, в отличие от нее, не имеющую никакого явного соответствия с реальным миром. После некоторой неразберихи стало ясно, что обе геометрии являются чисто математическими объектами, а вопрос о том, какая геометрия на самом деле лежит в основе реального мира, должны решать не математики, а физики. И хотя для геометрии Лобачевского позже и было найдено некоторое соответствие с реальностью, это уже не имело никакого значения и математики с энтузиазмом принялись конструировать все новые и новые геометрии. Появились пространства с произвольным и даже бесконечным числом измерений (Грассман, Кэли, Риман), с евклидовым, различными неевклидовыми расстояниями между объектами и даже совсем без метрики, аффинная и проективная геометрии, и т.д. Этому способствовало и то, что и в евклидовой геометрии к середине XVII века наконец появились координатные оси (Декарт). Для следующего шага, кажущегося сейчас очень простым, понадобился гений Ньютона, который ввел отрицательные координаты. Физики быстро добились огромных успехов, применяя геометрический подход и рассматривая многие свои задачи в подходящих пространствах большого, иногда бесконечного числа измерений и выбирая соответствующую метрику и удобную систему координат. Примерами могут служить теория относительности и квантовая механика.

Геометрический подход перспективен и для биологии. Он позволяет с единых позиций исследовать различные формы многомерной изменчивости биологических объектов, включая генетические, морфологические, функциональные и этологические характеристики особей, структуру, численность, пространственное распределение и динамику популяций и их параметров, а также влияние внешних и внутренних факторов. Таким единообразным способом могут быть решены научные проблемы самых различных областей биологии, которые не поддаются изучению традиционными биологическими средствами.

В связи с этим существует и очень актуальна следующая проблема: последовательный и корректный перевод биологических задач на язык геометрических расстояний и пространств для исследования математическими методами и интерпретация полученных результатов с целью выработки новых теоретических концепций биологии.

Подобная проблема стоит не только перед биологией. Как научные направления сформировались, например, психометрика – в психологии, хемометрика – в химии (Шараф и др., 1987; Родионова, Померанцев, 2006; Родионова, 2007), эконометрика – в экономике или клиометрия – в истории (Миронов, 1991). Однако математический аппарат в этих науках (кроме психометрики) на начальном этапе был полностью заимствован из биометрии.

которая исторически возникла вследствие усилий небольшой группы английских генетиков начала XX века, Ф.Гальтона, К.Пирсона, Р.Фишера, и американского экономиста Г.Хотеллинга. Сейчас эта область науки носит название многомерного статистического анализа (Кендалл, Стьюарт, 1976; Айвазян, 1985). К классическим методам многомерного статистического анализа относятся множественная регрессия, метод главных компонент, дискриминантный и канонический анализы. Психометрика развивалась параллельно и почти независимо от биометрии. К ее основным достижениям в области разработки математического аппарата относятся факторный анализ и многомерное шкалирование.

В биологии потребность в многомерных методах первыми, по-видимому, ощутили систематики (Гейнке, Смирнов; цит по Любищев, 1982) и геоботаники, работающие со списками и обилиями видов в растительных сообществах, которые в начале XX века предложили различные эмпирические и по этой причине, как правило, неевклидовы, индексы сходства. Однако уместно отметить, что "расстояние" между объектами, измеряемое этими индексами, обычно рассматривалось без пространства, в котором можно было бы отобразить их взаимное расположение, и до сих пор наиболее употребительным для этих целей остается применение методов кластерного анализа, например, плеяд П.В.Терентьева или малоинформативных дендрограмм. Реже дело доходило до ординации объектов, т.е. расположения их в линейном порядке, наиболее совпадающем с матрицей взаимного сходства. И только самые компьютеризированные биологи используют сегодня методы многомерного шкалирования неевклидовых расстояний для визуального представления взаимного расположения изучаемых ими объектов в многомерном евклидовом пространстве (напр., Васильев и др., 2003) или самоорганизующиеся карты признаков Кохонена (Kohonen, 1982).

С евклидовым расстоянием дело обстояло намного проще, поскольку при его использовании изначально предполагается, что объекты помещены в некоторое многомерное координатное признаковое пространство. В начале XX века К.Пирсон предложил множественную регрессию и метод главных компонент, который сильно опередил свое время и остался незамеченным. В 1930-е годы Р.Фишер разработал, в первую очередь, для систематиков, дискриминантный анализ, сутью которого является определение направлений, в отображении на которые в максимальной степени различаются группы объектов. В те же годы Хотеллинг переоткрыл метод главных компонент – выявление направлений, в проекции на которые в максимальной степени различаются объекты внутри одной группы – а также канонический анализ. После этого стало ясно, что одними и теми же методами можно обрабатывать данные любой природы. А когда в 50-е годы к ним присоединился факторный анализ, который вообще пришел из психологии, и обнаружилось, что это те же главные компоненты, только с вращениями, то со спецификой обрабатываемых данных было покончено окончательно, она полностью ушла в интерпретацию.

Исторически многомерный анализ биологических данных начался с работ Френсиса Гальтона (1822–1911), который попытался рассмотреть зависимость между средним ростом родителей и средним ростом их потомков. Таким образом, каждая семья характеризовалась значениями двух переменных. Предположив линейный характер зависимости и построив ее график по методу наименьших

квадратов, он обнаружил, что потомки в среднем ближе к популяционной средней, чем родители. Гальтон назвал это явление "регрессией" и с тех пор так называется любая функциональная зависимость одной переменной от одной или нескольких других, подобранная статистическими методами. {Ф.Гальтон – двоюродный брат Ч.Дарвина. Открыл антициклоны, основал дактилоскопию, евгенику, психометрику, генетику количественных признаков и биометрию (1889).}

Следующим был Карл Пирсон (1857–1936), который теоретически обосновал и разработал хорошо всем известный коэффициент линейной корреляции (коэффициент Браве–Пирсона) и много других коэффициентов, а также ввел понятие "множественной регрессии", т.е. функциональной зависимости одной переменной от нескольких других. Важнейшим частным случаем является множественная линейная регрессия. Он же вместе с Уэлдоном и Гальтоном (консультант-редактор) основал журнал "Биометрика" для статистического изучения биологических проблем (1901).

Однако наиболее известным статистиком XX века, безусловно, является Рональд Фишер (1890–1962), который заложил основы дисперсионного анализа. Кроме того, он первым начал систематически рассматривать объекты и выборки в многомерном пространстве и анализировать их разнообразие и взаимное расположение. Ему принадлежит заслуга разработки многомерного обобщения дисперсионного анализа – дискриминантного анализа – как способа нахождения одномерного направления, в проекции на которое наиболее различаются выборки (лекция 5). И хотя сам дискриминантный анализ, как сейчас становится ясным, не совсем адекватен биологической реальности и им нужно пользоваться, принимая некоторые меры предосторожности, для своего времени это был мощный шаг вперед. Следует отметить, что Гальтон и Фишер были биологами по основному образованию (Кембридж) и роду деятельности (генетики). {Термин "корреляция", безусловно, имеет биологическое происхождение, восходя к Кювье и отражая представления о целостности организмов и взаимозависимости его частей. В XX веке эти представления активно развивались И.И.Шмальгаузеном (1982).} Пирсон получил сначала юридическое образование, потом стал математиком, затем увлекся теорией эволюции и генетикой и даже заведовал кафедрой евгеники.

В связи с ними нельзя не упомянуть имя Гарольда Хотеллинга (1895–1973), который предложил метод главных компонент (не зная работы К.Пирсона) и канонический корреляционный анализ (Hotelling, 1933, 1936). Последний метод в общем виде не нашел применения из-за трудностей в интерпретации (следует иметь в виду, что иногда каноническим называется дискриминантный анализ Фишера, который является частным случаем канонического анализа Хотеллинга). Метод главных компонент сейчас применяется наиболее широко из всех многомерных методов и в следующих лекциях мы увидим, что это совсем не случайно. Хотеллинг был выдающимся американским экономистом, однако свою основополагающую работу по многомерному анализу опубликовал в психологическом образовательном журнале.

Все они заложили основы математической статистики и многомерного анализа, попутно разрабатывая недостающие разделы теории вероятностей, которой в оформленном виде тогда еще не существовало. Аксиоматическая теория вероятностей была создана позднее А.Н.Колмогоровым (1936). Далее

обнаружилось, что биометрические методы применяются или их можно применять и в других науках и, следовательно, они не являются специфически биологическими. Произошло слияние и поглощение биометрии математической статистикой и теорией вероятности, которые разрабатываются профессиональными математиками.

Тем временем психологи шли своим путем. Начало научному тестированию в психометрике (сама психометрика развивалась к тому времени уже несколько десятков лет) положил все тот же Ф.Гальтон, который пришел к необходимости измерять, кроме прочих, и психические характеристики человеческой личности: «Психометрия, необходимо твердо сказать, означает искусство охватывать измерением и числом операции ума (mind)», «Пока феномены какой-нибудь отрасли знания не будут подчинены измерению и числу, они не могут приобрести статус и достоинство науки» (Galton, 1879; цит. по Бурлачук, 2002). Ясно понимая, что человека нужно рассматривать по всей совокупности свойств как единое целое, он предложил схему обследования, в которую входили морфологические (рост, вес), физиологические (сила удара, скорость реакции) и психологические (ответы на тесты) признаки и обследовал более 9 тыс. человек. Примерно в это же время Дж.Кеттел, ученик Ф.Гальтона, предложил набор тестов, направленных именно на измерение психических свойств человека, т.е., тех, которые, с точки зрения обыденного сознания, меньше всего поддаются измерению (Cattell, 1890). Достаточно быстро выяснилось, что для измерения любого свойства необходима шкала, в которой можно выражать результаты измерений с тем, чтобы можно было сравнивать различных людей или одних и тех же в разные моменты времени или в разных условиях, а также исследовать влияние различных факторов, например, наследственности и среды. В естественных и технических науках измерение означает сравнение с эталоном. Однако в психологии, в отличие от естественных и технических наук, очень трудно предложить какие-либо универсальные эталоны, вроде метра или килограмма. Поэтому каждый психолог разрабатывал свой собственный набор характеристик личности, а также набор тестов для их выявления. В конце концов психологи, и в этом состоит их несомненная заслуга, сформировали расширенное понятие измерения: измерение есть приписывание чисел свойствам объектов по определенным правилам (Stevens, 1946; Стивенс, 1960).

Наряду с бесчисленным множеством разработанных и вновь разрабатываемых тестов (уже в двадцатых годах прошлого века их насчитывалось больше тысячи) велась кропотливая работа по разработке теории тестирования и математической обработке получаемых результатов. Еще Ф.Гальтон заметил, что результаты тестов должны коррелировать друг с другом (и использовал соответствующий коэффициент, который сейчас называется коэффициентом линейной корреляции Браве-Пирсона), а Ч.Спирмен (Spearman, 1904a, 1904б, 1927) положил это в основу своей теории G-фактора – генерального фактора, который должен обнаруживаться во всех тестах и который можно интерпретировать как проявление некоей умственной энергии. Он же предложил ранговый коэффициент корреляции, носящий теперь его имя. Фактически речь шла об одномерной шкале измерения интеллектуальных способностей. Но для того, чтобы отобразить какие-либо психологические особенности личности на числовой шкале, можно было воспользоваться двумя путями. Первый – измерять как можно больше разнообразных характеристик, отражающих эту особенность, и брать их линейную

или нелинейную комбинацию – факторный анализ (Thurstone, 1935, 1938) или метод главных компонент (Pearson, 1901; Hotelling, 1933). По историческим причинам психометрики применяли факторный анализ, а биометрики – метод главных компонент. Второй – предположить, что человек (эксперт) интуитивно ощущает расстояние на этой шкале и относительно двух объектов всегда может сказать, у какого из объектов эта особенность более выражена, чем у другого, или относительно двух пар объектов всегда может сказать, какая пара находится друг к другу ближе, чем другая. Отсюда с помощью математических операций можно определить упорядочивание на самой шкале или даже координаты объектов – шкалирование, неметрическое или метрическое.

В оба направления важный вклад внес Л.Терстоун, который использовал метод парных сравнений Кона для шкалирования одномерных различий между объектами (Thurstone, 1927), а также свой вариант факторного анализа (Thurstone, 1935, 1938). В отличие от подхода Ч.Спирмена, где интерпретация была определена заранее, факторный анализ Терстоуна допускал несколько групповых факторов и мог применяться к данным любой природы, а не только психологическим. Следует специально отметить, что у психологов речь шла не столько о математической модели, в которой естественно рассматривать несколько факторов, а один – считать просто частным случаем, сколько о том, какой именно вариант реализуется в действительности. Фактически Л.Терстоун предложил технологию, в которой сначала на основе метода парных сравнений строились одномерные шкалы, а затем из них конструировались групповые факторы с помощью факторного анализа. Таким образом, каждый объект получал набор координат и мог быть представлен точкой в многомерном пространстве. Факторный анализ Терстоуна требовал дополнительной интерпретации, что не нравилось многим психологам и вызвало их критику, но универсальность постановки привела к тому, что через некоторое время он вышел за пределы психологии и де-факто стал стандартом для других наук. Довольно скоро стало ясно, что, по сути, это те же главные компоненты, только с вращениями.

Если считать, что эксперт может оценить различия между парами объектов настолько, что можно их упорядочить, то можно поставить задачу определения координат объектов в многомерном пространстве с заданной метрикой (удобнее всего, евклидовой) таким образом, чтобы ранги различий как можно ближе соответствовали рангам дистанций между этими же парами в многомерном пространстве. Эти соображения легли в основу дистанционной модели М.Ричардсона (Richardson, 1938) – первого варианта неметрического многомерного шкалирования. Однако, из-за отсутствия вычислительных возможностей в то время этот метод не мог быть реализован. Поэтому В.Торгерсон предложил рассматривать различия между парами объектов как прямые аналоги расстояний в многомерном пространстве и разработал метод, позволяющий приписывать объектам координаты с сохранением расстояний – метрическая модель Торгерсона (Torgerson, 1952; Торгерсон, 1972). Эту модель уже можно было реализовать на компьютерах, что и было сделано. Но ее условия применимости оказались слишком жесткими, многие меры близости, применяемые психологами, явно не соответствовали аксиомам метрического расстояния, поэтому Р.Шепард и Дж.Крускал вернулись к первоначальным предположениям дистанционной модели М.Ричардсона (Shepard, 1962; Kruskal, 1964a, 1964б; Шепард, 1981). Р.Шепард построил алгоритм

неметрического шкалирования, минимизирующий различия между двумя упорядочениями: различий в исходной матрице данных и дистанций в многомерном пространстве. Особенно обнадежило то обстоятельство, что при неметрических предпосылках алгоритм практически однозначно воссоздавал метрическую структуру данных за счет избыточности числа связей между объектами. Дж.Крускал модифицировал этот алгоритм, предложив использовать квазиметрическую меру различий между двумя упорядочениями ("стресс"), сохраняющуюся при монотонных преобразованиях, и известные градиентные методы минимизации функций многих переменных.

Ситуация значительно улучшилась по сравнению с метрической моделью Торгерсона, однако по трудоемкости вычислений алгоритм Крускала имел четвертый порядок относительно числа объектов. Даже на современных персональных компьютерах это означает обработку не более сотни объектов. Для многих психологических работ этого вполне достаточно, но с многомерным шкалированием случилось то же самое, что и с факторным анализом, – он вышел за пределы психологии и стал применяться в других науках, а там часто требуются другие объемы, например, в молекулярной генетике. Совсем недавно Й.Тагучи и Й.Ооно (Taguchi, Oono, 2005) обнаружили, что возврат к первоначальной схеме Р. Шепарда сокращает время счета более чем на порядок и, соответственно, позволяет обрабатывать тысячи объектов. Это означает резкое расширение потенциальной сферы применимости методов многомерного шкалирования. В ближайшие 10-15 лет следует ожидать взрыва работ по этой тематике, в том числе, и в биологических и психологических исследованиях.

Все эти методы пережили второе рождение с появлением компьютеров, особенно персональных. Сложность вычислительных процедур и объем данных перестали быть ограничением и сейчас классические многомерные методы биометрии входят практически во все профессиональные пакеты статистического анализа данных. Хемометрики активно используют PLS-регрессию, первоначально появившуюся в эконометрике (Boardman et al., 1981; Wold, 1985). Кроме того, за пределами многомерного статистического анализа, наряду с факторным анализом (Иберла, 1980) и многомерным шкалированием (Дэйвисон, 1988), появились специфические компьютерные методы, такие, как самоорганизующиеся карты признаков (Kohonen, 1982) и нейронные сети (Горбань, Россиев, 1996). В отличие от классических методов многомерного анализа, они не опираются ни на какие предположения о распределении данных в генеральной совокупности и не используют расчета достоверности. По строгости теории они значительно уступают методам многомерного статистического анализа. Их прообразом является кластерный анализ (Дидэ, 1985), который тоже появился на заре XX века, однако, вряд ли его можно относить к многомерным методам, так как в нем вообще нет идеи геометрического пространства, в котором расположены объекты. И шкалирование и карты Кохонена как раз дополняют кластерный анализ геометрией взаимного расположения объектов.

С точки зрения практических приложений ситуация выглядит иначе. Очень широко применяются в биологических исследованиях и хорошо поддаются содержательной интерпретации факторный анализ и его разновидность, метод главных компонент, и кластерный анализ, как правило, в виде дендрограмм. Из-за

трудностей в интерпретации практически не используется канонический анализ. Часто применяются множественная регрессия и дискриминантный анализ, однако интерпретировать их с биологических позиций гораздо труднее, чем факторный и кластерный анализы. Карты Кохонена и нейронные сети очень перспективны, однако они только входят в практику обработки биологических данных. Заслуживают большего внимания, хорошо интерпретируются, но редко используются методы многомерного шкалирования. Совсем не используется биологами, и совершенно напрасно, PLS-регрессия.

Вместе с тем, ситуация в биологии и смежных науках продолжает оставаться неудовлетворительной. Во-первых, основная масса биологов недостаточно знакома с математикой и информатикой и предпочитает использовать более простые, хотя и давно устаревшие приемы. В качестве примера можно указать на большую популярность дендрограмм даже среди лидеров современной биологии – молекулярных генетиков, не говоря уж о геоботаниках и систематиках.

Во-вторых, в основном, по историческим причинам, геометрическая суть методов многомерного анализа оказалась скрыта за плотной завесой вероятностно-статистических представлений и понятий. В результате вместо анализа содержательной, биологической стороны дела вопрос все чаще сводится к крайне важному, но все же никак не первичному, определению достоверности полученных результатов. Это не означает, что нужно совсем отказываться от расчета достоверности. Иметь представление о статистической устойчивости получаемых результатов, безусловно, нужно. Хорошим вспомогательным, специфически компьютерным и вполне оправдавшим себя на практике средством, является, например, бутстреп-метод (Efron, 1979, 1982; Диаконис, Эфрон, 1983) (лекция 7). Не нужно только абсолютизировать значимость подобных расчетов.

В-третьих, некоторые из широко распространенных и стандартных методов многомерного статистического анализа, в частности, дискриминантный анализ и множественная регрессия, используют такие линейные преобразования пространства, которые изменяют расстояния между объектами в ходе обработки и, соответственно, искажают содержательный смысл получаемых результатов. Оставаясь безупречно правильными с математической точки зрения, эти методы вместе с рассчитываемой ими достоверностью не совсем адекватны той реальности, для изучения которой предназначены (лекции 5–6).

Таким образом, степень использования многомерных методов в биологии зависит не столько от того, насколько они теоретически обоснованы, сколько от того, насколько они помогают получать биологически интерпретируемые результаты. Это, в свою очередь, зависит от того, насколько биологическая сущность сходства и различия объектов воспроизводится геометрией взаимного расположения отображающих их точек в многомерном пространстве. Наиболее работоспособны те методы, которые в минимальной степени искажают задаваемые исследователем расстояния между объектами.

Особенностью предлагаемого курса является анализ не взаимосвязей между признаками, а расположения объектов в образованном признаками пространстве и направлений изменчивости через корреляции с признаками, а также доведения этого анализа до биологической интерпретации. Главная ценность многомерного анализа заключается не столько в определении достоверности получаемых результатов,

сколько в содержащейся в нем возможности визуализировать промежуточные и окончательные результаты анализа и интерпретировать их с биологической точки зрения. Прежде, чем исследовать гипотезу, ее сначала надо выдвинуть. А до того, как выдвинуть, ее еще надо увидеть. Современная тенденция как раз и заключается в стремлении визуализировать данные, даже в ущерб достоверности и теоретической обоснованности. Хороший результат должен быть представлен в такой форме, чтобы он был очевиден (оче-виден = виден очам) для специалистов в соответствующей предметной области. Когда такой очевидности достигнуть не удается, приходится прибегать к статистическим критериям.

Математическая статистика как наука сформировалась только во второй половине XX века, а представление, что естественно-научные результаты только тогда являются доказательными, когда они обоснованы статистически, стало более или менее общепринятым только в последней четверти XX века. Возникает вполне законный вопрос: а как же наука обходилась без такого обоснования несколько тысяч лет? Архимед не садился в ванну сто раз, чтобы набрать статистику. Согласно легенде, ему хватило одного, чтобы увидеть закон. Обошелся без статистических критериев и Ньютон, когда записал в виде математического выражения закон всемирного притяжения (сам закон принадлежит Гуку). В его время оценка показателя степени при R в формуле

$$F = \gamma \frac{m_1 m_2}{R^2}$$

была возможна с точностью порядка 4%. Но он не усомнился в том, что этот показатель строго равен двум для всей Вселенной, явно и далеко выходя за пределы статистической обоснованности. И оказался прав. Сейчас точность оценки этого показателя составляет около десятка нулей после запятой и он по-прежнему считается равным двум, хотя время от времени и выдвигаются предположения, что он все-таки чуть-чуть отличается от двойки.

Что касается достоверности, то надо ясно понимать ее место. Обычная статистическая практика заключается в том, что мы идеализируем те условия, в которых были получены данные, например, предполагаем существование и многомерную нормальность распределения объектов, отсутствие систематических ошибок, бесконечно большой размер выборки и т.д. В этих идеализированных условиях мы рассчитываем вероятность случайного получения нашего результата и, если она оказывается достаточно мала, делаем вывод, что наша гипотеза статистически подтверждается. Безусловно, это очень важный косвенный довод в пользу гипотезы, но никак не окончательный вердикт. Это примерно то же самое, что предполагать, что чемпион по стрельбе в тире будет самым лучшим охотником в тайге или снайпером на войне. Поэтому главным критерием всегда останется биологический смысл, а окончательное слово всегда принадлежит специалистам в соответствующей предметной области.

В курсе рассмотрен ряд задач, в основном, из области популяционной экологии животных, которые решаются с помощью методов многомерного анализа и которые нельзя было бы решить без этих методов. Спектр задач достаточно широк и хорошо иллюстрирует возможности геометрического подхода к анализу биологических объектов.

Чего нет в этом курсе? Нет дисперсионного анализа и теории планирования

эксперимента. Нет теории проверки гипотез и критических областей, традиционно входящих в курсы математической статистики. Нет проверки нормальности. Для временных рядов нет спектрального анализа, устранения тренда и разложения в ряд Фурье. Все, кому это интересно, отсылаются к специальной литературе.

ЛЕКЦИЯ 2. Предварительная работа с данными

Для проведения многомерного анализа нужно представить исходные данные в виде таблицы "объект–признак", в которой каждый объект характеризуется значениями признаков. Понятие объекта является первичным. Предполагается, что существует некоторая генеральная совокупность объектов и у всех объектов имеются одни и те же свойства (атрибуты, характеристики, параметры) или на них влияют одни и те же факторы, значения которых можно определить для каждого объекта. Множество значений одного свойства или фактора для всей совокупности объектов называется признаком. Обычно мы имеем некоторую выборку объектов, случайную или неслучайную, которая в частных случаях может совпадать со всей генеральной совокупностью. Поскольку каждый реальный объект может характеризоваться необозримым числом свойств, нам приходится выбирать некоторый ограниченный набор признаков, однако понятие выборки к признакам, как правило, не применяется. Объекты должны быть более или менее однородными и обладать некоторым внутренним единством, тогда как признаки могут быть весьма разнокачественными по своей природе.

В некоторых случаях объекты и признаки можно менять местами. Например, если мы рассматриваем смертность мужчин от инфекционных заболеваний за ряд лет по всем экономическим регионам, то за объекты можно принять как регионы, так и годы. Причиной является то обстоятельство, что на самом деле у нас есть один признак – смертность мужчин, измеренный для всех пар «регион–год», которые фактически и есть «настоящие» объекты. В зависимости от целей исследования мы можем принять первые члены пары за объекты, а вторые – за признаки и наоборот. Более сложная ситуация возникает, когда мы рассматриваем смертность от инфекционных заболеваний за ряд лет по всем экономическим районам в зависимости от пола, т.е. фактически имеем тройку «регион–год–пол» в качестве первичного объекта и смертность – в качестве признака. Тогда в качестве объектов мы можем принять и регионы, и годы, и мужчин (женщин), и пары «регионы–годы», «регионы–мужчины (женщины)», «годы–мужчины (женщины)» а в качестве признаков – оставшиеся члены троек.

Признаки делятся на качественные (номинальные), ранговые (порядковые, ординальные) и количественные (интервальные) (Stevens, 1946; Стивенс, 1960). Значения качественных признаков (градации) можно сравнивать только на совпадение. Например, признак «виды» в знаменитых данных Р.Фишера для объектов «ирисы» имеет градации «setosa», «versicol», «virginic» (Fisher, 1936). Качественными могут быть и числовые признаки, например, номера маршрутов городского транспорта.

Отдельного разговора заслуживают ранговые признаки, измеряемые в порядковой шкале. Здесь возможны две ситуации. Значения ранговых признаков могут отражать только отношение порядка в данной выборке объектов. В этом

случае их значения для конкретного объекта зависят от других членов рассматриваемой выборки и могут измениться при добавлении в выборку новых объектов. Эту ситуацию необходимо отличать от ситуации, когда упорядоченным является исходное множество значений признака, например, возраст грызунов, выраженный градациями juvenis, subadultus, adultus, senex, или стадия развития лягушек (Северцов, 2000). При добавлении в выборку новых объектов значения старых уже не изменятся. И в том и в другом случае градациям можно приписать порядковые номера и обращаться с таким признаком, как с количественным. Разница состоит в том, что в первом случае ранги подчиняются равномерному распределению, во втором – распределение произвольно.

Значения количественных признаков получают путем счета (счетные, меристические признаки) или измерения (мерные, метрические, пластические). Значения каждого количественного признака можно представить в виде точек числовой оси и для них, кроме отношения «меньше–больше», имеет смысл вопрос «насколько?». Кроме того, для длин интервалов имеет смысл вопрос «во сколько раз?». Примерами количественных признаков могут служить промеры длины и ширины чашелистика и лепестков и т.п. Говорят, что качественные признаки измерены в номинальной, а количественные -- в интервальной шкале. Иногда среди количественных признаков выделяют признаки, измеренные в шкале отношений, для которых фиксировано начало отсчета и имеет смысл отношение самих значений («во сколько раз?»), но на практике с ними поступают, как с обычными интервальными признаками. Однако, тем не менее, уместно заметить, что широко известный коэффициент вариации имеет смысл только для признаков, измеренных в шкале отношений.

Отнесение признаков к тому или иному типу достаточно условно. Например, счетные признаки при малом числе принимаемых ими значений ведут себя, как качественные, а при большом – как мерные. Такой признак как «зональность», имеющий градации «арктическая тундра», «субарктическая тундра», «лесотундровое редколесье», «северная тайга», «средняя тайга», «южная тайга», «подтаежные леса», «северная лесостепь», «южная лесостепь», «степь» – хотя и выглядит качественным, но его можно рассматривать и как ранговый, так как градации упорядочены в широтном направлении. Любой ранговый признак фактически является счетным, так как его значение для любого объекта равно числу значений меньше него плюс единица. Мерные признаки всегда измеряются с некоторой точностью, поэтому множество принимаемых ими значений можно считать конечным. Из любого количественного признака легко получить ранговый, правда, с потерей информации, упорядочив его значения и взяв в качестве новых значений их порядковые номера. Еще один способ, также с потерей информации, заключается в разбиении значений количественного признака на ряд классов и отнесении каждого из объектов к одному из классов. Например, рост людей можно измерять в сантиметрах, а можно грубо разбить на три класса: низкорослые, среднего роста, высокие. Такой признак можно считать как ранговым, так и качественным. Далее мы увидим, что признаки всех типов можно обрабатывать одними и теми же алгоритмами.

После того, как определены значения признаков для всех объектов выборки, можно заняться статистикой, то есть подсчетом того, сколько и каких объектов

имеется в выборке и представлением этих сведений в обозримом и сжатом виде. Исторически с древнейших времен и до конца XIX века статистика ничем другим и не занималась, а математическая статистика, как наука, сложилась и оформилась только во второй половине XX века. Само слово "статистика" происходит от латинского слова "status" – положение или состояние. От него же происходят и слова "штаты", "государство". Сведения для государственного аппарата собирались еще в глубокой древности, как правило, в целях налогообложения. Известны китайский сборник Шу-Кинг (VI век до н.э.), сообщения Геродота о деятельности Дария и Ксеркса (VI–V век до н.э.), "Политика" Аристотеля (IV век до н.э.), цензы древнего Рима и т.д. На Руси первым примером систематического сбора статистических сведений могут служить переписи населения, проведенные татаро-монголами в XIII веке для упорядочения сбора дани.

Современное название этот предмет получает в середине XVIII века в заглавии книги "Notitia rerum politica vulgo statistica" ("Сведения о делах государственных, в просторечии называемые статистикой"). В XX веке статистикой стали называть учение о методах наблюдений любых массовых явлений (Терентьев, 1971).

Любой способ определения значений признаков, включая визуальный и экспертный, будем называть измерением. Например, глаз опытного специалиста способен различить 120 оттенков черного цвета ткани. Главная цель измерения признаков, которую никогда нельзя упускать из виду – это определение сходства или расстояния между объектами. Признаки нужны не сами по себе, а для различения объектов. Если какой-то даже очень важный признак имеет одно и то же значение для всех объектов, то для обработки он абсолютно бесполезен. Поэтому всегда нужно обращать внимание на то, насколько выбранная шкала отражает те содержательные различия, которые нужно измерить. Например, при использовании ранговых признаков по умолчанию подразумевается, что нам известен только порядок следования объектов и поэтому надежнее всего считать, что расстояние между соседними градациями одинаково. Если же это предположение нас не устраивает, то это значит, что у нас имеется некая явная или неявная дополнительная информация. Но шкалу всегда можно переопределить. Например, в автогонках по Формуле1, а также в командном зачете на Олимпиадах, очки даются за первые шесть мест, причем за первое место 9 очков, за второе – 6, за третье – 4 и далее 3, 2, 1 очко. Это означает, что расстояние между победителем и вторым призером приравнивается к трем условным единицам, а расстояние между седьмым и последним участником – к нулю. {Предельный случай. В средневековом городе N состоялся турнир рыцарей. Победитель получает руку и сердце прекрасной дамы. Участникам, занявшим второе–тридцатое места, предоставлены лучшие места на городском кладбище.} Часто применяемыми способами переопределения шкалы являются логарифмическое преобразование или извлечение корня некоторой степени. Эти преобразования меняют расстояния между объектами. Критерием правильности подбора преобразования служит соответствие полученных расстояний содержательному биологическому смыслу.

Если признак может принимать всего два значения, например, пол, то расстояние между этими значениями всегда одинаково и проще всего кодировать их значениями 0 и 1. В этом случае признак называется бинарным, двоичным,

дихотомическим, индикаторным или характеристическим. Бинарный признак фактически является количественным.

Если номинальный признак может принимать больше двух значений, то расстояние между разными градациями тоже всегда считается одинаковым, но одномерную шкалу в этом случае подобрать нельзя и нужно кодировать такой признак несколькими бинарными, сопоставляя каждой градации отдельный признак и ставя 1, если номинальное значение совпадает с этой градацией, и 0 – в противном случае.

Будем считать, что для рассматриваемой выборки номинальные признаки, если они есть, уже представлены в двоичном виде, значения порядковых признаков заменены их рангами, а для количественных признаков подобраны адекватные шкалы. Это означает, что все признаки можно считать количественными. Тем не менее, остается еще несколько проблем.

Первая: признаки могут быть несопоставимы между собой по единицам измерения, например, вес, длина и пол, или давление и возраст. Вторая – признаки, измеренные в одних и тех же единицах, могут сильно отличаться по абсолютной величине, например, длина черепа и межглазничная ширина. Третья – необходимо измерять расстояние между объектами одновременно по нескольким признакам.

Многомерное пространство. Центрирование и нормирование

Если мы умножим значения любого количественного признака на любую ненулевую константу и прибавим к ним любую константу, то это никак не изменит относительных расстояний между объектами по этому признаку. Поэтому мы можем использовать преобразования сдвига и масштаба для приведения разных признаков в соответствие друг с другом. Преобразование:

$$x_i' = (x_i - \bar{x}), \quad \bar{x} = \sum x_i / N,$$

где $\bar{x}$ – среднее значение, N – число объектов, называется *центрированием*. После центрирования новое среднее признака равно 0:

$$\bar{x}' = \sum x_i' / N = 0.$$

Преобразование:

$$x_i' = \frac{x_i}{s}, \quad s^2 = \sum (x_i - \bar{x})^2 / N$$

где s^2 – дисперсия признака (вместо N часто применяется N-1), называется *нормированием*. После такого преобразования все признаки становятся безразмерными, а новая дисперсия равна 1:

$$s'^2 = \sum (x_i' - \bar{x}')^2 / N = 1$$

Каждый объект через значения измеренных у него признаков можно представить в виде точки в многомерном евклидовом пространстве. Каждый признак является в этом пространстве отдельной координатной осью, ортогональной всем остальным. Все объекты образуют в этом пространстве некоторое "облако". Координатами точек являются значения признаков. До нормировки это "облако" может находиться в стороне от начала координат, которое расположено в точке с нулевыми значениями всех признаков. Как мы уже знаем,

исходные признаки, как правило, центрируются и нормируются. Центрирование геометрически означает перенос начала координат в "центр тяжести облака" – точку со средними значениями всех признаков, которая называется центроидом. Очевидно, что взаимное расположение объектов при центрировании не меняется. Нормировка признаков приводит к изменению масштабов пространства таким образом, что разброс точек вокруг среднего (равного нулю после центрирования) становится одинаковым по каждой оси и равным единице, то есть все признаки уравниваются в правах и приобретают равный вес. Одним из мифов, сложившихся вокруг многомерного анализа, является представление о том, что нормировка – обязательный элемент этого метода. Это не так. Наиболее четко ситуация обрисована в трехтомнике Кендалла и Стьюарта (1976): "Решение о нормировке должно приниматься, исходя из нестатистических соображений". Если по каким-то содержательным причинам нужно придать разные веса исходным признакам или оставить первоначальные (например, работая с частотами), то исследователь вправе это делать по своему усмотрению. Весом признака служит величина разброса вокруг среднего, а не его абсолютные значения.

{Поэтому общепринятые правила судейства в наших КВН являются не совсем объективными. Важность конкурсов задается предельным числом очков, которые можно за него поставить, например, 4 – за разминку и 7 – за домашнее задание. Однако в первом случае судьи (кроме Гусмана), как правило, выбирают между 3 и 4, во втором – между 6 и 7. Это означает, что фактически все конкурсы равноправны и команда, проигравшая разминку с крупным счетом, уже имеет мало шансов отыграться на более важных конкурсах. Правильнее было бы судить все конкурсы из 10 баллов, а их важность оценивать коэффициентами, на которые нужно умножить результаты каждого конкурса.}

Надо всегда учитывать, что любая нормировка заново определяет евклидово расстояние между объектами. На практике количественные признаки, как правило, нормируются, исходя именно из желания исследователя так определить расстояние между объектами, чтобы все признаки участвовали в его определении в равной мере. Однако коррелирующие признаки в какой-то степени дублируют друг друга, и это неизбежно влияет на расстояние между объектами. В качестве попытки решить эту проблему было предложено расстояние Махаланобиса (лекция 3). Возможны и другие нормировки и другие расстояния, которые могут даже не быть расстояниями в том смысле, что для них не выполняются аксиомы метрики. В этом случае они называются различиями.

Возможна ситуация, когда координаты объектов не заданы, а вместо этого сразу дана матрица расстояний (количественный признак на парах объектов) или различий (ранговый признак). (Если задана матрица сходства, то ее всегда можно преобразовать в матрицу различий.) Чтобы приписать объектам координаты, применяются методы многомерного шкалирования (лекция 7).

ЛЕКЦИЯ 3. Линейная алгебра

Основным объектом многомерного анализа является таблица "объект–признак". Все признаки можно считать количественными. Каждый признак отображается на числовую ось и отражает расстояние между объектами. Каждый

признак имеет определенный вес, характеризующий относительную важность этого признака и равный его дисперсии. После стандартной нормировки на среднеквадратичное отклонение все признаки имеют равный вес. Веса объектов считаются равными. Более сложную ситуацию, когда объектам тоже приписываются разные веса, рассматривать не будем. Отметим только, что она не сводится ни к случаю еще одного признака, ни к умножению значений объектов на веса.

Введем следующие определения:

Скаляр – действительное число.

Вектор – набор скаляров.

Матрица – набор векторов одинаковой длины.

Вектор-строка – матрица из одной строки.

Вектор-столбец – матрица из одного столбца.

Операции:

Умножение матрицы на скаляр.

Скалярное произведение векторов x и y: $(x, y) = \sum x_i y_i$

Умножение матрицы на вектор.

Умножение матрицы на матрицу.

Сложение матриц.

Транспонирование матрицы. $(AB)' = B'A'$.

Единичная матрица I. Диагональная матрица L.

Ортогональная матрица. $QQ' = Q'Q = I$. $Q = Q_1 Q_2$.

Будем считать известными понятия скаляра, вектора, матрицы (единичная, диагональная, ортогональная) и операций на ними: умножение матрицы на скаляр, скалярное произведение векторов x и y $(x, y) = \sum x_i y_i$, умножение матрицы на вектор, умножение матрицы на матрицу, сложение матриц, транспонирование матрицы (Ланкастер, 1978).

Таблица "объект–признак" является матрицей, а каждый объект – вектором. Каждый признак тоже является вектором. Геометрическое представление: если в качестве осей выбрать признаки, то каждый объект может быть представлен точкой в этом пространстве. Координатами точки служат значения признаков. Такое пространство будем называть пространством объектов или основным. Если в качестве осей выбрать объекты, то каждый признак может быть представлен точкой в этом пространстве. Будем называть его пространством признаков или двойственным. Оба пространства определены одновременно на основе одной и той же матрицы. Если значения в матрице меняются, то одновременно меняются положения объектов и признаков, как точек в соответствующих пространствах.

Размерность – важнейшее свойство пространства. Размерность основного пространства – число признаков. Размерность двойственного – число объектов. Если размерность равна единице, то точки можно расположить на числовой оси. Если размерность равна двум, то их можно расположить на плоскости. Если размерность равна трем, то совокупность точек еще можно представить наглядно в привычном для наших органов чувств виде, разместив их в пространстве. Если размерность пространства больше трех, то взаимное расположение точек в этом пространстве

можно представить только мысленно, хотя и существуют различные хитроумные приемы для визуального отображения пространств большей размерности: физико-географические карты (цвет), полигоны, лица Чернова и т.д.

Определим в пространстве расстояние между точками по формуле: $d_{xy}^2 = \sum (x_i - y_i)^2$. Такое расстояние является многомерным обобщением обычного пифагорова расстояния и называется евклидовым. Евклидовым называется и все пространство, если в нем определено евклидово расстояние. Каждую точку можно рассматривать как вектор относительно начала координат.

Вычислим скалярное произведение вектора x само на себя $(x,x) = \sum x_i^2$. Показатель $\| x \| = \sqrt{(x,x)}$ называется длиной вектора и является расстоянием до точки x от начала координат. Определим угол α_{xy} между x и y по формуле: $\cos(\alpha_{xy}) = (x,y) / \| x \| \| y \|$. Показатель $r_{xy} = \cos(\alpha_{xy})$ называется коэффициентом корреляции между признаками.

Свойства:

Если ко всем значениям одного признака прибавить или вычесть одно и то же число, то расстояние между объектами не изменится. Произойдет перенос начала координат. Центроид – вектор средних. Центрирование – перенос начала координат в центр тяжести выборки.

Если все значения всех признаков умножить или разделить на одно и то же ненулевое число, то взаимное расположение объектов не изменится. Все расстояния пропорционально возрастут или уменьшатся. Все углы останутся прежними.

После центрирования и нормировки на среднеквадратичные отклонения длины всех признаков одинаковы и равны $\sqrt{N}$, то есть зависят от числа объектов. Разделим все значения всех признаков на $\sqrt{N}$. Тогда в двойственном пространстве все признаки будут расположены на единичной окружности, длины всех признаков равны 1, а $r_{xy} = \cos(\alpha_{xy}) = (x,y) / \| x \| \| y \|$ для любой пары признаков.

Поэтому для одной выборки с матрицей X будем всегда считать, что признаки центрированы и нормированы на их длину. Произведение матриц $R = X'X$ есть матрица коэффициентов корреляции.

Раскроем скобки в определении расстояния между объектами:

$$d_{xy}^2 = \sum (x_i - y_i)^2 = \sum x_i^2 + \sum y_i^2 - 2\sum x_i y_i = (x,x) + (y,y) - 2(x,y).$$

Произведение $D = XX'$ – это матрица, по которой можно легко вычислить расстояния между объектами. Действительно, диагональные элементы равны $D_{xx} = (x,y)$, а недиагональные $D_{xy} = (x,y)$. Поэтому

$$d_{xy}^2 = D_{xx} + D_{yy} - 2D_{xy}.$$

Умножим матрицу X на произвольную ортогональную матрицу Q: $Y = XQ$. Произведение $D = YY' = XQQ'X' = X(QQ')X' = XIX' = XX' = D$ не изменится. Следовательно, не изменятся и расстояния между объектами. Геометрически

умножение на ортогональную матрицу означает поворот в основном пространстве объектов. Так как произведение ортогональных матриц - тоже ортогональная матрица, то последовательность поворотов – тоже поворот.

Матрица $Y = XQ$ – матрица новых признаков. При повороте меняются дисперсии признаков и корреляции между ними. Таким образом, из одного набора старых признаков мы можем с помощью поворотов получить бесконечное число наборов новых признаков. Однако расстояния между объектами и сумма дисперсий признаков *при повороте не меняются*. Если мы хотим, чтобы после нормировки никакие методы обработки не меняли взаимных расстояний между объектами, то такие методы должны базироваться на поворотах. Некоторые стандартные алгоритмы факторного анализа – это нахождение поворотов, удовлетворяющих определенным критериям (варимакс, квартимакс и т.д.).

Почему бинарные признаки можно обрабатывать так же, как количественные

В большинстве статистических учебников вопросы обработки качественных и количественных признаков излагаются раздельно. Алгоритмы и методы выглядят столь различно, что не возникает и мысли об их внутреннем единстве. В настоящем разделе будет показано, что основные формулы, применяемые для анализа качественных признаков, прямо выводятся из соответствующих формул для количественных признаков.

Пусть признак x у N объектов принимает только два значения: 0 и 1. Пусть число единиц равно k. Вычислим среднее и дисперсию признака по формулам для количественных признаков (Васильева, 2000):

$$\bar{x} = \sum x_i / N = k / N = p;$$

$$s^2 = \sum (x_i - \bar{x})^2 / N = (\sum x_i^2 - 2\bar{x}\sum x_i + N\bar{x}^2) / N = (Np - 2Np^2 + Np^2) / N = (p - p^2) = pq,$$

где p – частота признака, $q = 1-p$.

Таким образом, и среднее, и дисперсия признака полностью выражаются через его частоту. Распределение p подчиняется биномиальному закону, который приближенно аппроксимируется нормальным распределением с параметрами p и pq/N Приближение применимо при $Npq \geq 9$ (Корн, Корн, 1970). Поэтому грубое сравнение двух средних для бинарных признаков можно проводить, как и для количественных признаков, с помощью обычного t-критерия. Для более точного сравнения необходимо, конечно, применять φ-преобразование Фишера (Плохинский, 1961).

Пусть теперь значения двух признаков, x и y, у N объектов равны только 0 или 1. Вычислим корреляцию между признаками по формулам для количественных признаков:

$$r_{xy} = (x, y) / \| x \| \| y \| = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} =$$

$$= (\sum x_i y_i - N\bar{x}\bar{y}) / \sqrt{(\sum x_i^2 - N\bar{x}^2)(\sum y_i^2 - N\bar{y}^2)} = (a - Np_x p_y) / \sqrt{(Np_x - Np_x^2)(Np_y - Np_y^2)} =$$

$$= (a - (a+b)(a+c)/N) / \sqrt{(a+b)(1-(a+b)/N)(a+c)(1-(a+c)/N)} =$$

$$= (Na - (a+b)(a+c)) / \sqrt{(a+b)(N-(a+b)(a+c)(N-(a+c)} =$$

$$= ((a+b+c+d)a-(a+b)(a+c))/\sqrt{(a+b)(c+d)(a+c)(b+d)} =$$

$$= (ad-bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)},$$

где a – число объектов со значениями 1 обоих признаков;
b – число объектов со значениями 1 признака x и 0 – признака y;
c – число объектов со значениями 0 признака x и 1 – признака y;
d – число объектов со значениями 0 обоих признаков;
$p_x = (a+b)/N$ – частота признака x;
$p_y = (a+c)/N$ – частота признака y;
$N = a+b+c+d$.

Результат полностью совпадает с формулой тетрахорического коэффициента корреляции для бинарных признаков, который вычисляется по так называемой четырехпольной таблице (табл. 3.1). Впервые это было показано еще К. Пирсоном (Pearson, 1900), автором тетрахорического коэффициента.

Таблица 3.1

Четырехпольная таблица сопряженности двух бинарных признаков

$x \backslash y$	*1*	*0*	Σ
1	a	b	$a+b$
0	c	d	$c+d$
Σ	$a+c$	$b+d$	N

Рассмотрим теперь ситуацию, когда один из признаков, например, y, является количественным и подчиняется нормальному распределению, а другой – x – может принимать только два значения, 0 и 1. Пусть число единиц в x равно N_1, а число нулей – N_0. Фактически выборка разбивается на две с числом объектов N_1 и N_0, $N_1 + N_0 = N$. Вычислим корреляцию между признаками по формулам для количественных признаков:

$$r_{xy} = (x,y)/\|x\|\|y\| = \sum(x_i - x)(y_i - \bar{y})/\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2} =$$

$$= (\sum x_i y_i - N\bar{x}\bar{y})/\sqrt{(\sum x_i^2 - N\bar{x}^2)(\sum y_i^2 - N\bar{y}^2)} =$$

$$(N_1\bar{y}_1 - N(\frac{N_1}{N})\bar{y})/\sqrt{(\frac{N_1}{N} - N(\frac{N_1}{N})^2)(N_0 s_0^2 - N_0\bar{y}_0^2 + N_1 s_1^2 - N_1\bar{y}_1^2 - N\bar{y}^2)}$$

$$\frac{N_0 N_1}{N}(\bar{y}_1 - \bar{y}_0)/\sqrt{\frac{N_0 N_1}{N}(N_0 s_0^2 + N_1 s_1^2 + \frac{N_0 N_1}{N}(\bar{y}_1 - \bar{y}_0)^2)} =$$

$$= (y_1 - \bar{y}_0)/\sqrt{\frac{(N_0 s_0^2 + N_1 s_1^2)N}{N_0 N_1} + (\bar{y}_1 - \bar{y}_0)}$$

где $\bar{y}_1 = \sum_{x=1} y_i / N_1$ – среднее единичной выборки;

$y_0 = \sum_{x=0} y_i / N_0$ – среднее нулевой выборки;

$y = \frac{N_0 y_0 + N_1 y_1}{N}$ – среднее всей выборки

$s_1^2 = \sum_{x=1} (y_i - \bar{y}_1)^2 / N_1$ – дисперсия единичной выборки;

$s_0^2 = \sum_{x=0} (y_i - y_0)^2 / N_0$ – дисперсия нулевой выборки.

Отсюда $(\bar{y}_1 - \bar{y}_0) = \frac{r}{\sqrt{1-r^2}} \sqrt{\frac{(N_0 s_0^2 + N_1 s_1^2)N}{N_0 N_1}}$

$$\frac{(y_1 - \bar{y}_0)}{\sqrt{\frac{(N_0 s_0^2 + N_1 s_1^2)N}{N_0 N_1}}} = \frac{r}{\sqrt{(1-r^2)}}$$

Умножая обе части на $N-2$ получим

$$\frac{\sqrt{N-2}(\bar{y}_1 - y_0)}{\sqrt{\frac{(N_0 s_0^2 + N_1 s_1^2)N}{N_0 N_1}}} = r\sqrt{\frac{N-2}{(1-r^2)}} = t$$

В случае нормального распределения для двух количественных признаков при нулевой гипотезе ρ=0 статистика справа имеет *t*-распределение с *N-2* степенями свободы. Статистика слева есть *t*-критерий Стьюдента равенства средних для двух выборок объема N_0 и N_1 из одного и того же нормального распределения и также подчиняется *t*-распределению с *N-2* степенями свободы (Большев, Смирнов, 1983). Отсюда следует, что в случае бинарного и количественного признаков вычисление коэффициента корреляции между ними и определение его достоверности можно проводить по формулам для двух количественных признаков. В этом случае критерий достоверности коэффициента корреляции совпадает с критерием Стьюдента равенства средних для двух выборок (см. замечание в лекции 5).

Таким образом, вычисление среднего значения бинарного признака по формуле для количественного признака эквивалентно вычислению его частоты. Сравнение двух средних для бинарных признаков можно проводить, как и для количественных признаков, с помощью обычного *t*-критерия. Вычисление корреляции между количественным и бинарным признаками по формулам для количественных признаков эквивалентно сравнению средних по *t*-критерию Стьюдента. Вычисление корреляции между двумя бинарными признаками по формулам для количественных признаков эквивалентно вычислению тетрахорического коэффициента корреляции. Поэтому во всех случаях вычисления можно проводить по формулам для количественных признаков, пользуясь, например, стандартными статистическими пакетами.

Метод главных компонент

Пусть имеется матрица X, содержащая N строк (объектов) и M столбцов (признаков). Обозначим через X' транспонированную матрицу, в которой строки и столбцы меняются местами, и положим $R=X'X$. Пусть Q – матрица собственных векторов матрицы R, Λ – диагональная матрица ее собственных значений и I – единичная матрица. Пусть $U=XQ$. Тогда (Кендалл, Стьюарт, 1976):

$$U'U = Q'X'XQ = Q'RQ = \Lambda$$

$$Q'Q = QQ' = I.$$

Поскольку матрица Q ортогональна, то умножение на нее – это фактически поворот осей в многомерном пространстве, сохраняющий эвклидово расстояние между объектами. Матрица U имеет те же размеры, что и X, однако ее столбцы не коррелируют между собой. Дисперсии λ_j столбцов матрицы U являются собственными значениями матрицы R и диагональными элементами матрицы Λ. Сами столбцы являются линейными комбинациями столбцов матрицы X с суммой квадратов коэффициентов, равной единице, и называются главными компонентами. Каждая компонента имеет дисперсию, максимально возможную из всех линейных комбинаций, ортогональных предыдущим компонентам. Обработка матрицы X методом главных компонент заключается в вычислении матриц U, Q и Λ. Программы вычисления собственных векторов и собственных значений имеются в стандартном математическом обеспечении современных компьютеров (Агеев и др., 1976; Гайдышев, 2001). Если собственные векторы матрицы Q умножить на квадратные корни из собственных чисел λ_j, то мы получим коэффициенты корреляции между компонентами и столбцами матрицы X, достоверность которых можно определять по стандартным статистическим таблицам.

Матрица R называется матрицей вторых или смешанных моментов. Если столбцы матрицы X центрированы $\left(\sum_i x_{ij} / N = 0\right)$, то матрица R называется ковариационной, а если и нормированы на длину $\left(\sum_i x_{ij}^2 / N = 1\right)$, то корреляционной. Как правило, методу главных компонент предшествует центрирование и нормирование матрицы X.

Как мы уже видели, и ранговые и двоичные признаки можно обрабатывать как количественные и вместе с ними, хотя во многих руководствах и учебниках утверждается обратное (Ким, Мьюллер, 1989). Коэффициенты линейной корреляции в этом случае автоматически переходят в соответствующие ранговые, бисериальные и тетрахорические коэффициенты, следовательно, матрица корреляций никогда не будет иметь отрицательных собственных значений, вопреки мнению М.Кендалла и А.Стьюарта (1976).

Если поменять объекты и признаки местами, то получим транспонированную матрицу X'. Ее также можно обрабатывать методом главных компонент. Пусть U^*, Q^* и Λ^* - матрицы, полученные в результате такой обработки и $p = min(N,M)$. Тогда:а) ненулевые собственные значения матриц Λ и Λ^* равны и их не более p;

б) первые p столбцов нормированной матрицы компонент U совпадают с первыми p столбцами (собственными векторами) матрицы Q^*;

с) первые p столбцов нормированной матрицы компонент U^* совпадают с первыми p столбцами (собственными векторами) матрицы Q.

Удобнее обрабатывать матрицу, у которой число столбцов меньше, чем число строк.

Главные компоненты не коррелируют между собой. Каждая из них отвечает за свою долю изменчивости. Так как сумма дисперсий при поворотах не меняется, то смысл имеет только доля каждой компоненты. Обычно она выражается в процентах. Главный смысл применения главных компонент заключается в том, что первые компоненты могут взять на себя значительную часть общей дисперсии и выявить реальную размерность данных.

Поворот с помощью матрицы Q, очевидно, не меняет расстояний между объектами. А вот нормировка главных компонент собственными значениями λ_i приводит к изменению расстояний. Новое пространство $Y = XQ\Lambda^{-1/2}$ называется пространством Махаланобиса, а расстояние в нем – расстоянием Махаланобиса. Для чего это делается?

Корреляция между исходными признаками означает, что признаки в некоторой мере дублируют друг друга. Если, например, корреляция равна 1, то дублируют полностью. В этом случае у нас фактически один признак, повторенный дважды. Очевидно, что второй признак никакой новой информации не несет. Однако он дает вклад как в расстояние между объектами, так и в дисперсии главных компонент, в которые входит. Кроме того, порождается еще одна компонента с нулевой дисперсией, которая начинает приносить неприятности в множественной регрессии и дискриминантном анализе. Можно, конечно, его просто отбросить, как рекомендуется в некоторых статистических руководствах. Но, если корреляция между признаками по модулю меньше 1, то, отбрасывая один из них, мы, наряду с дублируемой, лишаемся и какой-то уникальной информации. Дублирование информации выражается в увеличении дисперсий первых главных компонент и в появлении новых компонент с малыми и нулевыми дисперсиями. Если пронормировать главные компоненты, то мы уберем это дублирование, сохранив всю необходимую информацию.

К сожалению, у этой красивой идеи есть очень большой недостаток – при переходе в пространство Махаланобиса нормируются все компоненты, а не только имеющие большие дисперсии. Это приводит к тому, что неоправданно большой вес получают дальние компоненты с малыми дисперсиями. Мы искусственно увеличиваем масштаб изменчивости по направлениям, которые совершенно этого не заслуживают, фактически умножаем «шум». На сегодняшний день эта проблема, несмотря на несколько десятков лет исследований, еще не имеет приемлемого решения. Практическая рекомендация заключается в том, что нужно вообще выбросить из анализа дальние компоненты с малыми или нулевыми дисперсиями. Другим практическим выходом из положения является PLS-регрессия (лекция 6).

Метод нелинейных главных компонент

В ситуации, когда множество точек в многомерном пространстве на самом деле укладываются в подмножество меньшей размерности, применимы методы, изложенные в (Principal Manifolds..., 2007).

Поворот осей. Факторный анализ

Иногда распределение объектов на плоскости главных компонент, особенно в

случаях, сильно отклоняющихся от нормального распределения, удобнее анализировать, если плоскость повернуть на некоторый угол. Однако надо сразу оговориться, что в этом случае оси в общем случае перестают быть ортогональными. В факторном анализе этот прием является основным, так как в нем ортогональности не требуется изначально. Именно поэтому многие алгоритмы факторного анализа начинают с метода главных компонент, а потом добавляют к нему поворот, исходя из каких-либо соображений наподобие простоты структуры нагрузок, как в известных критериях "варимакс " или "квартимакс". Однако соображения могут быть и любые другие, например, расположение оси в направлении некоторой интересной или отклоняющейся группы объектов и т.д. Технически это осуществляется следующим образом. Пусть α – угол поворота, а v_1 и v_2 – векторы нагрузок (собственные векторы, вклады признаков, веса), соответствующие осям плоскости u_1 и u_2. Определим новые оси и новые векторы нагрузок через формулы:

$$v'_1 = v_1 cos\,\alpha + v_2 sin\,\alpha$$
$$v'_2 = -v_1 sin\,\alpha + v_2 cos\,\alpha$$
$$u'_1 = u_1 cos\,\alpha + u_2 sin\,\alpha$$
$$u'_2 = -u_1 sin\,\alpha + u_2 cos\,\alpha$$

Так как $v_1^2 = v_2^2 = 1$ и $v_1 v_2 = 0$, то легко видеть, что и новые векторы нагрузок будут удовлетворять этим же соотношениям. Поэтому после поворота можно анализировать вклады признаков в новые оси точно так же, как и в старые (Однако в общем случае после поворота корреляции новых компонент с исходными признаками уже не будут пропорциональны коэффициентам векторов нагрузок.) Дисперсии новых компонент будут равны:

$$u'^2_1 = u_1^2 cos^2\,\alpha + u_2^2 sin^2\,\alpha = \lambda_1 cos^2\,\alpha + \lambda_2 sin^2\,\alpha$$
$$u'^2_2 = u_1^2 sin^2\,\alpha + u_2^2 cos^2\,\alpha = \lambda_1 sin^2\,\alpha + \lambda_2 cos^2\,\alpha$$

ЛЕКЦИЯ 4. Внутривыборочная изменчивость

Цель настоящей лекции – подробнее разобраться в том, насколько полезен многомерный анализ как средство поиска биологического смысла при анализе изменчивости биологических объектов. Прежде всего, у нас есть объекты, есть признаки и есть значения признаков для каждого объекта, сведенные в таблицу "объект – признак". Что такое объекты – объяснять не надо. От них требуется, чтобы они были биологическими по своей природе, иначе ни о каком биологическом смысле говорить не придется, и обладали некоторым внутренним единством. Для определенности – пусть это будут черепа полевок, взятых в одной или нескольких географических точках. Что такое признаки, подробно разбиралось в лекции 2. Примерами количественных признаков могут служить промеры черепов, сделанные штангенциркулем: кондилобазальная длина, длина мозговой части, межглазничная ширина и т.п. Признаки получаются путем применения к объектам некоторой измерительной процедуры, например, сравнения с некоторым эталоном, и называются исходными.

Но как быть с признаками, которые получаются путем вычислений из исходных признаков, например, с очень широко распространенными среди морфологов индексами? Мы знаем, что у биологических объектов размеры сильно

варьируют, и хотим избавиться от их влияния, чтобы изучать форму в чистом виде. Поэтому берем отношение одного признака к другому, например, длины мозговой части к кондилобазальной длине, и считаем его новым признаком, имеющим те же права, что и исходные признаки. (Примечание 1. Еще лучше взять логарифм отношения, тогда абсолютная величина нового признака не будет зависеть от того, берем ли мы отношение меньшего признака к большему или наоборот. Примечание 2. Эта операция применима только к признакам, измеренным в шкале отношений, то есть, имеющим фиксированное начало отсчета. Примечание 3. По мнению академика А.Д.Александрова (1987) само понятие вещественного (действительного) числа выросло из отношения длин отрезков).

Имеет ли признак, полученный таким образом, биологический смысл? Любой систематик ответит на этот вопрос утвердительно, исходя из многолетнего опыта своей науки. Можно ли вычислить, например, его наследуемость? А почему бы и нет, ответит любой генетик, конечно, можно, если нас интересует наследуемость именно формы. Является ли новый признак "математической переменной", непосредственно не измеряемой никаким инструментом? Вне всякого сомнения. Откуда же у "математической переменной" взялся биологический смысл? И вслед за этим второй вопрос – а был ли биологический смысл у исходных признаков и какой именно? Ведь то, что мы умеем что-то измерять, вовсе не означает, что само измерение осмысленно. Ответ зависит от того, для какой цели мы проводим измерения или вычисляем новые признаки.

В случае с черепами ответ более или менее очевиден. Основной причиной вариабельности промеров черепа в выборках из природных популяций является возрастная изменчивость. Но не единственной. Даже если брать только взрослых особей, например, перезимовавших полевок, или даже строго одновозрастных лабораторных крыс (Atchley et al., 1981), то наиболее заметными все равно будут различия в размерах. Растущему организму проще всего среагировать на любые внешние влияния или проявить внутренние отличия от других особей замедлением или ускорением развития организма в целом или отдельных его частей. Вычисление индексов исходит из не всегда осознаваемого предположения, что с увеличением размеров все промеры увеличиваются пропорционально, не меняя формы объекта, которая, таким образом, является инвариантом. (Более аккуратные рассуждения учитывают возможную аллометрию, но суть от этого не меняется). А разница в индексах означает разницу именно в форме, которая часто более интересна биологам, особенно систематикам, чем разница в размерах. Получается, что "математическая переменная" может иметь больший биологический смысл, чем те исходные признаки, из которых она вычислена. А они, в свою очередь, служат лишь вспомогательными, промежуточными звеньями для достижения цели.

Перейдем теперь к многомерному анализу. Каждый объект через значения измеренных у него признаков можно представить в виде точки в многомерном евклидовом пространстве. Каждый признак является в этом пространстве отдельной координатной осью, ортогональной (перпендикулярной) всем остальным. Все объекты образуют в этом пространстве некоторое "облако". Координатами точек являются значения признаков. В случае промеров черепа, которые всегда положительны, это "облако" находится в стороне от начала координат, которое расположено в точке с нулевыми значениями всех признаков. Кроме того,

некоторые признаки имеют заведомо бо'льшие значения, чем другие, например, кондилобазальная длина всегда больше межглазничной ширины, и "облако" вдоль таких признаков будет более сдвинутым и более вытянутым. Возможна ситуация, когда они отличаются и по размерности, например, если мы будем дополнительно брать вес черепа и нижней челюсти. Поэтому исходные признаки, как правило, центрируются и нормируются (лекция 2).

После центрирования и нормирования все объекты получают новые координаты – значения центрированных и нормированных признаков. Эти признаки обладают следующим математическим свойством: если взять скалярное произведение любых двух признаков (сумму попарных произведений координат объектов), то она будет равна линейному коэффициенту корреляции между ними (лекция 3). Положительный коэффициент корреляции означает, что с увеличением одного признака, как правило, увеличивается и другой, отрицательный – что другой уменьшается. Иными словами, поведение одного признака дает нам некоторую информацию о поведении другого. Нулевой коэффициент корреляции означает отсутствие линейной статистической связи между признаками, то есть при увеличении или уменьшении одного из признаков, другой изменяется произвольным образом. Обычно достоверность коэффициента корреляции определяется, исходя из предположения о двумерной нормальности распределения объектов по исследуемым признакам. Однако еще одним из распространенных мифов о методе главных компонент является представление о том, что он обязательно требует многомерной нормальности распределения. Это абсолютно не так. Объекты могут быть распределены как угодно, образовывая одно "облако" или несколько любой нелинейной формы, например, в виде подков или петель, скалярное произведение признаков все равно будет являться линейным коэффициентом корреляции. Нормальность требуется только при определении достоверности коэффициента корреляции и то только потому, что мы не умеем ее вычислять в случае других распределений. Еще она желательна, но не обязательна, когда мы трактуем компоненты как действие независимых причин.

Теперь переходим к самому главному – а есть ли биологический смысл в распределении "облака" точек, представляющих наши объекты в признаковом пространстве, и их взаимном расположении? То, что такой смысл есть в распределении объектов по каждому отдельному признаку, никто не сомневается, так как обычно именно по отдельным признакам ведется содержательный анализ. Однако математически оба представления эквивалентны. Геометрически каждый признак представляет собой всего лишь некоторое направление, на которое спроецировано многомерное "облако" объектов. И, наоборот, из распределений объектов по всем признакам "облако" восстанавливается однозначно. Следовательно, содержательный смысл у обоих представлений абсолютно одинаков. Разница заключается только в том, что, анализируя признаки по отдельности, мы не видим того общего, что их объединяет, а многомерный анализ позволяет охватить всю картину разом, предоставляя для этого некоторые дополнительные возможности. Например, если рассмотреть любое другое направление в многомерном пространстве и спроецировать на него это "облако", то мы можем изучать распределение объектов и по нему точно так же, как и по любому исходному признаку. Каждый объект получит некоторое значение, являющееся его

координатой на новой оси, т.е. фактически мы получим новый признак. То, что мы непосредственно не измеряли его значений, а вычислили их из значений исходных признаков, как мы уже видели выше на примере индексов, никакой роли не играет. Более того, мы можем рассмотреть любой набор взаимно перпендикулярных направлений, число которых равно числу исходных признаков, и спроецировать на них наше "облако". Геометрически это означает поворот в пространстве, который не меняет расстояний и взаимного расположения объектов. Такой поворот называется ортогональным. Наглядным примером может служить перемещение точек изображения на экране дисплея относительно осей комнаты, когда мы его поворачиваем для того, чтобы лучше рассмотреть изображение.

Какой из возможных поворотов выбрать, зависит исключительно от целей, которые мы перед собой ставим. Например, если у нас есть две группы объектов, о которых у нас есть некоторая дополнительная информация, мы можем одну из осей провести через них и посмотреть, как на этой оси расположатся другие объекты. Мы можем выбрать поворот и из соображений удобства. (Кстати, именно это и делает факторный анализ (лекция 3).) Ведь иметь дело с исходными признаками как раз не очень удобно. Так как дисперсия каждого признака после нормировки равна единице, каждый признак вносит равную долю информации. Однако визуально анализировать расположение точек в более чем трехмерном пространстве наши органы чувств не приспособлены. Даже перебор всех сочетаний по два или три признака достаточно утомителен, хотя никому и не возбраняется. Но охватить всю картину и увидеть ее целостность, если она, конечно, есть, мы все равно не в состоянии.

Вот здесь и выступает на сцену коррелированность признаков. Если признаки коррелируют между собой, то это означает, что "облако" объектов в многомерном пространстве вытянуто вдоль некоторого направления, не совпадающего ни с одной из осей, и тем больше вытянуто, чем сильнее они коррелируют. И распределение объектов по этому направлению имеет дисперсию больше единицы, то есть формально содержит больше изменчивости, чем любой исходный признак. Поэтому мы можем поставить математическую задачу: найти направление, на котором достигается максимальная дисперсия проекции "облака". Именно эта задача решается в методе главных компонент (Pearson, 1901; Hotelling, 1933). Точнее говоря, в методе главных компонент ищется весь набор направлений, сохраняющий всю информацию об "облаке" и обладающий одним дополнительным свойством: если упорядочить направления по величине из изменчивости, то суммарная дисперсия любого числа первых компонент максимальна. На практике нередки случаи, когда, например, на первые две-три компоненты приходится 70-80% всей дисперсии. Поскольку с геометрической точки зрения переход к новым признакам означает всего лишь поворот всего "облака" в пространстве без изменения расстояний и взаимного расположения объектов, то информация не добавляется и не исчезает, просто это же "облако" объектов мы видим в несколько ином ракурсе, наиболее удобном для обозрения. Каждый новый признак является линейной комбинацией старых (суммой с некоторыми коэффициентами). Верно и обратное, каждый старый признак является линейной комбинацией новых, что лишний раз подчеркивает сохранность исходной информации при этих преобразованиях.

(Кстати, попутно развеем еще пару мифов, сложившийся вокруг главных

компонент. Для метода главных компонент совершенно необязательно вычислять корреляционную или ковариационную матрицу и поэтому число объектов абсолютно не обязано быть таким, чтобы коэффициенты корреляции были достоверными. Тем более не обязательно, чтобы число объектов превышало число признаков. Другое дело, что расчет через корреляционную матрицу технически очень удобен и излагается во всех статистических руководствах и применяется во всех статистических пакетах. Некоторые пакеты даже не умеют вычислять главные компоненты, если число объектов меньше числа признаков. Однако эквивалентные результаты можно получить и прямым вычислением главных компонент без всякой корреляционной матрицы, так называемым "разложением по сингулярным числам" (SVD). При этом корреляции между компонентами все равно окажутся равными нулю, причем независимо от формы "облака". Второй миф заключается в том, что, поскольку коэффициент корреляции является линейным, т.е. измеряет только степень приближения "облака" к прямой линии, то вся информация, содержащаяся в возможном нелинейном расположении объектов, пропадает. На самом деле эта информация никуда не девается, при ортогональных поворотах "облака" взаимное расположение объектов полностью сохраняется и при проекции на главные компоненты нелинейность очень хорошо визуализируется. Естественно, коэффициенты корреляции, в силу своей линейности, ее не отражают и требуются какие-нибудь специальные нелинейные меры, если нужно выразить ее численно, но главные компоненты этому ничуть не мешают, скорее наоборот. Часто достаточно самой визуализации через компоненты, чтобы правильно понять биологический смысл наблюдаемой нелинейности.)

После расчета компонент мы можем оставить для рассмотрения только первые две-три из них и потерять при этом всего лишь 20-30% общей изменчивости. Если в распределении объектов в исходном признаковом пространстве был какой-то биологический смысл, то мы его потеряли в минимально возможной степени. Причем, скорее всего, потеряли не столько информацию, сколько "шум", неизбежно присутствующий в реальных данных из-за ошибок измерения и действия малозначимых или случайных причин. А что приобрели взамен? А приобрели возможность анализировать распределение объектов в дву-трехмерном пространстве вместо *M*-мерного, где *M* зачастую равняется нескольким десяткам или даже сотням признаков и с которым мы все равно не могли справиться. А так как компоненты, как мы уже видели, являются новыми признаками, то получается, что два-три таких новых признака заменяют все старые с минимальной потерей информации. Но у таких мощных признаков очень даже может быть биологический смысл!

Вот тут-то и нужна компетентность биолога. Трактовка результатов всегда зависит от природы объектов, с которыми мы имеем дело, и от задачи, которую мы перед собой поставили. Интерпретация компонент предъявляет довольно высокие требования к квалификации биолога, так как нужно одновременно понимать геометрический смысл проведенных преобразований и биологический смысл получаемых результатов. Например, нулевые корреляции между компонентами, как мы уже знаем, означают статистическую независимость, т.е. поведение одной компоненты статистически ничего нам не говорит о поведении другой. Поэтому вполне осмысленным и часто оправдывающимся на практике является

предположение, что и биологически эти новые признаки достаточно автономны, например, отражают разные процессы или фазы развития особей. Само собой разумеется, что это только предположение и его каждый раз нужно обосновывать биологически, например, анализируя вклады признаков в компоненты (Васильев и др., 2003).

В случае с черепами грызунов "облако" объектов, как правило, имеет эллипсовидную форму, а первая компонента всегда представляет общие размеры, так как почти все или даже все признаки дают в нее вклад одного знака, часто близкий по величине. Но поскольку в ней участвует много признаков, то эти размеры определены надежнее и универсальнее, чем если бы использовали только какой-то один из них, например, кондилобазальную длину. Во вторую компоненту, тоже практически всегда, основной вклад вносит межглазничная ширина, а большинство вкладов остальных признаков противоположно ей по знаку. Это означает, что когда значения второй компоненты увеличиваются, то увеличивается и межглазничная ширина, а значения большинства остальных уменьшаются и наоборот. Межглазничная ширина в процессе онтогенеза ведет себя очень самостоятельно (Виноградов, 1921; Европейская рыжая ..., 1981). Во-первых, она раньше остальных останавливается в росте, вероятно, потому, что растущий организм не может позволить себе роскошь отложить на потом формирование системы зрения, как, например, половое созревание. Видеть надо всегда, а глазная система слишком тонкий инструмент, чтобы успевать постоянно подстраиваться под слишком сильные изменения межглазничного расстояния. Поэтому дальнейшее развитие черепа происходит таким образом, чтобы в минимальной степени затрагивать уже сформировавшуюся систему зрения. Во-вторых, относительная изменчивость межглазничной ширины, если ее измерять, например, коэффициентом вариации, заметно больше изменчивости других признаков, что, вкупе с ранней остановкой в росте и, следовательно, меньшей зависимостью от среды, позволяет предполагать проявление в ней наследственных отличий между особями (Ковалева, 1999). По литературным данным, наследуемость промеров черепа, включая межглазничную ширину, колеблется в диапазоне 0.4–0.6 (Atchley et al., 1981). Но генетической трактовке мешает то, что, хоть и в меньшей степени, межглазничная ширина тоже участвует в росте и в ее изменчивости наследственные различия между особями могут быть смешаны с размерной изменчивостью. Здесь-то и помогает многомерный анализ. Как мы уже видели, главные компоненты обладают одной важной особенностью: они статистически независимы по построению. Поэтому, если в первой компоненте сосредоточена вся размерная изменчивость, то во второй и последующих она "снята". И поэтому предположение о значительной доле наследственных факторов в изменчивости второй компоненты имеет больше шансов оказаться справедливым, чем такое же предположение об изменчивости самой межглазничной ширины. Следовательно, и первая и вторая компоненты, рассматриваемые как новые признаки, могут нести в себе содержательную информацию в некотором "очищенном", в отличие от обычных признаков, виде.

Что касается биологического смысла самой межглазничной ширины, то этот вопрос довольно подробно исследовался еще академиком С.С.Шварцем. По его мнению, одним из существенных факторов, определяющих различия в пропорциях черепа, является скорость роста животных. "У медленно растущих животных

кондилобазальная длина черепа, скуловая ширина, длина зубного ряда, и лицевой части больше, а высота черепа и ширина межглазничного промежутка меньше, чем у растущих быстро" (Шварц, 1980). Поэтому, учитывая, что первая компонента "снимает" размерную изменчивость, а основной вклад во вторую компоненту, противоположный по знаку вкладам большинства других признаков, дает межглазничная ширина, есть все основания полагать, что изменчивость по второй компоненте отражает различия между особями по скорости роста, причем в значительной степени обусловленные наследственными факторами, т.е. различия в генетических программах развития особей.

Следует заметить, что точками в многомерном пространстве можно представлять не только особей, но и любые другие биологические объекты, например, популяции. В работах (Косова и др., 1992; Галактионов и др., 1995) исследованы 50 выборок половозрелых особей остромордой лягушки. Совокупность средних значений 14 морфометрических признаков по всем выборкам обработана методом главных компонент. Все признаки внесли в первую компоненту положительный вклад. Поэтому логично назвать эту компоненту размерно-возрастной. Подобная трактовка первой компоненты достаточно универсальна, так как преобладающая изменчивость общих размеров проявляется практически во всех морфометрических исследованиях (Галактионов, 1981; Животовский, 1984; Акимов и др., 1993). Какой-либо связи с ландшафтами и географией в распределении выборок по первой компоненте не обнаруживается, что, по-видимому, свидетельствует о том, что выборки брались достаточно рандомизированно по отношению к их средним размерам.

Ландшафтная специфика выборок раскрылась в пространстве II и III главных компонент (рис. 4.1). Выборки сгруппированы по их географическому положению: северная группа – выборки Поозерской провинции, включая выборку 48 и тяготеющие к ним выборки 12, 13; центральная – выборки Белорусской Возвышенной (без 12, 13), Предполесской (без 50) и Восточно-Белорусской провинций; южная – выборки Полесской провинции. Южная группа, в свою очередь, разбита на две группы выборок, относящихся к Брестской (включая выборку 50) и Гомельской областям (Косова и др., 1992).

Все выборки, относящиеся к центральной группе, сместились вниз по третьей компоненте (рис. 4.1). Выборки из южной группы занимают крайнее левое положение. Очевидно, вторая компонента отражает направление изменчивости «юг-север», т.е. связь с температурным градиентом среды. Выборки центральной группы отличаются от всех остальных тем, что они приурочены к возвышенной части территории Беларуси, а наиболее отклоняющиеся по третьей компоненте выборки 18, 15, 16 располагаются ближе других к самой ее высокой точке. Для понимания фенотипических различий между ландшафтными популяциями следует учесть, что северная и южная группы занимают низменные, наиболее заболоченные части территории Беларуси. Таким образом, третья компонента отражает изменчивость, связанную с направлением «возвышенность» – «низменность», т.е. с высотным градиентом среды.

В третью компоненту фактически дали вклады только промеры бедра и голени и, с обратным знаком, длина внутреннего пяточного бугра (рис. 4.2). Это означает, что даже приведенные к равным общим размерам за счет отбрасывания

первой компоненты выборки из центра Белоруссии дополнительно отличаются длинноногостью. Морфогенетическая интерпретация напрашивается сама собой - адаптация к более сухопутной жизни. (По этой логике длина внутреннего пяточного бугра должна означать адаптацию к плаванию.)

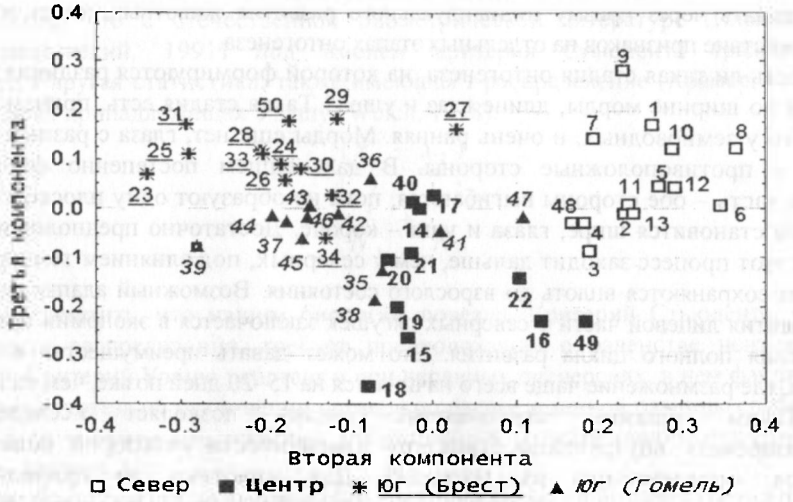


Рис. 4.1. Расположение ландшафтных выборок остромордой лягушки в пространстве II и III главных компонент изменчивости средних значений выборок

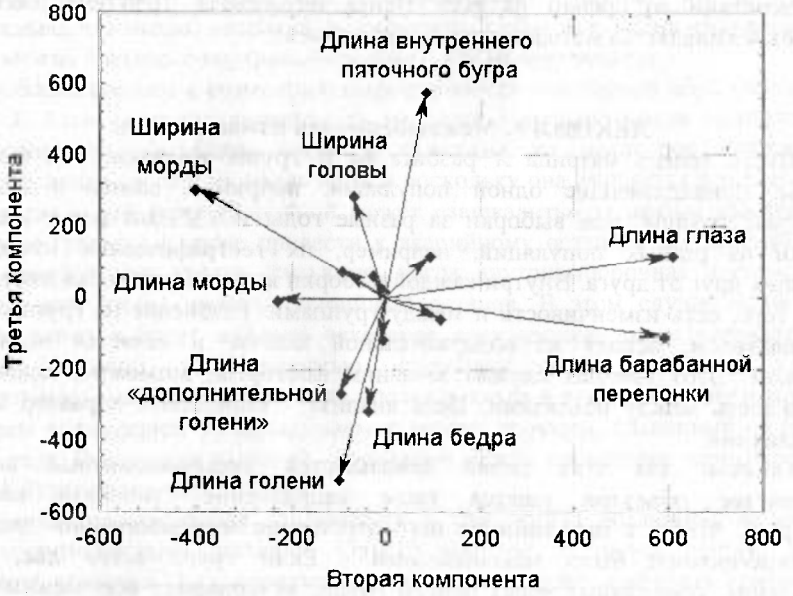


Рис. 4.2. Вклады признаков во II и III главные компоненты изменчивости средних значений ландшафтных выборок остромордой лягушки

Интерпретация второй компоненты не столь очевидна. У северных лягушек, при равных размерах, более узкая и короткая морда и более длинные глаза и барабанные перепонки по сравнению с южными. Причины сопряженного варьирования признаков по каждой из главных компонент могут быть установлены, если вклады признаков рассматривать через призму индивидуального развития животных, через возможное взаимодействие признаков на отдельных этапах онтогенеза.

Есть ли такая стадия онтогенеза, на которой формируются различия главным образом по ширине морды, длине глаз и ушей? Такая стадия есть, причем у всех, а не только у земноводных, и очень ранняя. Морды еще нет, глаза с разных сторон и глядят в противоположные стороны. В дальнейшем постепенно формируется лицевая часть – обе стороны выгибаются, пока не образуют одну плоскость. Морда при этом становится шире, глаза и уши – короче. Достаточно предположить, что у южных этот процесс заходит дальше, чем у северных, под влиянием температуры, и различия сохраняются вплоть до взрослого состояния. Возможный адаптивный смысл недоразвития лицевой части у северных лягушек заключается в экономии времени для завершения полного цикла развития. Это может давать преимущество в северных районах, где размножение чаще всего начинается на 15–20 дней позже, чем на юге.

Таким образом, многомерный анализ позволяет исследовать и визуализировать внутреннюю структуру изменчивости некоторой совокупности объектов, определяемую их природой. Для животных из природных или лабораторных популяций, как хорошо известно биологам, основными факторами фенотипической изменчивости особей являются их генетическая изменчивость и эволюционно сформировавшиеся закономерности развития особей, вытекающие из необходимости поддерживать целостность и функционирование организма при взаимодействии со средой на всех этапах онтогенеза. Действие именно этих факторов и выявляется методом главных компонент.

ЛЕКЦИЯ 5. Межвыборочная изменчивость

Пусть теперь матрица X разбита на K групп объектов. Это могут быть объекты, принадлежащие одной популяции, например, самцы и самки, или возрастные группы, или выборки за разные годы и т.д. Это могут быть также выборки из разных популяций, например, их географических точек, далеко отстоящих друг от друга. Внутри каждой выборки имеется некоторая изменчивость. Кроме того, есть изменчивость и между группами. Разбиение на группы задается исследователем, исходя из содержательной задачи, и является номинальным признаком. Этот признак служит внешним фактором, возможно, влияющим на изменчивость между объектами. Цель анализа – определить характер и степень этого влияния.

Обычно для этих целей используется дискриминантный анализ. В пространстве объектов ищется такое направление (линейная комбинация признаков), чтобы в проекции на него отношение межвыборочной дисперсии к внутривыборочной было максимальным. Если групп всего две, то одно направление, проходящее через центры групп, исчерпывает всю межвыборочную изменчивость. Если при этом признак только один, то дискриминантный анализ сводится к хорошо известному *t*-критерию Стьюдента:

$$t = \frac{N - 2N_1N_2(x_1 - x_2)}{(N_1s_1^2 + N_2s_2^2)N}$$

Заметим, что в отечественной биометрической литературе (Плохинский, 1961; Животовский, 1991) под именем критерия Стьюдента традиционно используется другая статистика, также имеющая *t*-распределение (Крамер, 1975) и, на самом деле, принадлежащая Уэлшу (Welch, 1938):

$$t = \frac{(\bar{x}_1 - x_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Надо заметить, что нашим биологам повезло. Критерий Стьюдента, кроме нормальности распределения, требует предположения о равенстве неизвестных дисперсий. Критерий Уэлша работает и при неравных дисперсиях, в нем фактически проверяется гипотеза об отклонении нормально распределенной разницы средних от нуля. Таблица критических значений для различных уровней значимости одна и та же в обоих случаях.

Если групп три, то их центры образуют плоскость, на которой сосредоточена вся межвыборочная изменчивость, при условии, что они не лежат на одной прямой, и т.д. В общем случае таких направлений насчитывается *min(K-1, M)*.

С формально-математической точки зрения дискриминантный анализ безупречен. На первый взгляд, и с содержательной стороной все в порядке. Действительно, что можно возразить против того, чтобы межвыборочная дисперсия была как можно больше, а внутривыборочная – как можно меньше.

Проблема состоит в возможной вырожденности или плохой обусловленности матрицы *X*. Если реальная размерность матрицы *X* меньше числа признаков, то может возникнуть ситуация, когда в проекции на некоторое направление внутривыборочная дисперсия очень мала, а поскольку она находится в знаменателе, то отношение к ней межвыборочной может «зашкалить» за любые мыслимые и немыслимые границы и даже привести к аварийному останову вычислений. Еще хуже, с нашей точки зрения, ситуация, когда внутривыборочная дисперсия не настолько мала, чтобы вызвать аварийный останов. В этом случае вычисления дойдут до конца и будет найдено некоторое дискриминирующее направление с формально высокой достоверностью, лишенное, тем не менее, всякого содержательного смысла. Эта ситуация вполне реальна и возникает, например, если мы изучаем асимметрию и закладываем в анализ промеры, сделанные на разных сторонах тела. Вследствие высокой корреляции между промерами парных органов матрица *X* будет плохо обусловлена.

Обычная рекомендация заключается в том, чтобы исключить из анализа высоко коррелирующие признаки. Однако уместно поставить вопрос: а чем провинились признаки? Исследователь должен иметь право подбирать признаки по своему усмотрению, исходя из поставленной им задачи, а если математический аппарат не срабатывает, то, возможно, дело в самом аппарате?

Чтобы ответить на этот вопрос, надо поставить другой: что происходит с расстояниями между объектами в дискриминантном анализе? Алгоритм дискриминантного анализа может быть представлен в виде следующей последовательности действий (Уилкс, 1967). Сначала каждая группа центрируется своими средними для исключения межвыборочной изменчивости. Геометрически это означает совмещение центров тяжести всех выборок с началом координат и объединение выборок. К объединенной выборке X_c, характеризующей после центрирования только внутривыборочную изменчивость, применяется метод главных компонент, то есть ищется ортогональная матрица поворота Q, приводящая к некоррелированным осям. Найденный поворот применяется к обеим матрицам, $Y=XQ$ и $Y_c=X_cQ$. Компоненты Y_c нормируются своими дисперсиями λ_i и ими же нормируются компоненты Y. В полученных пространствах все направления имеют одну и ту же внутривыборочную дисперсию, равную единице, и для нахождения направления с максимальной межвыборочной дисперсией достаточно еще раз применить к Y метод главных компонент.

Рассмотрим подробно каждый шаг. Поворот с помощью матрицы Q не меняет расстояний между объектами. А вот нормировка собственными значениями λ_i приводит к изменению расстояний. Новое пространство Y называется пространством Махаланобиса, а расстояние в нем – расстоянием Махаланобиса (лекция 3). Деление на собственные числа приводит к тому, что все главные компоненты внутривыборочной матрицы приобретают равный вес. Математически это очень удобно, но содержательно совершенно бессмысленно. Наряду с несколькими первыми компонентами, содержащими действительно полезную информацию, (которые, возможно, и стоит нормировать), в анализ на равных правах включаются и все остальные. Однако дальние компоненты содержат, в основном, «шум», причем их тем больше, чем больше число исходных признаков и чем сильнее эти признаки коррелируют между собой. Расстояние Махаланобиса заглушает полезную информацию, умножая «шумы». Именно оно является слабым звеном дискриминантного анализа.

Стоит подчеркнуть, что в силу своего внутреннего устройства дискриминантный анализ всегда, в большей или меньшей степени, искажает реальную информацию. Возможно, именно поэтому дискриминантные оси труднее интерпретировать через вклады признаков, чем главные компоненты, и дело обычно сводится к констатации достоверности различий, чему очень способствует умножение «шумов». Кроме того, при возврате в исходное пространство признаков дискриминантные оси становятся неортогональными, а это очень неудобно для интерпретации.

Очевидно, нужны другие алгоритмы дискриминантного анализа. Они должны максимизировать различия между межвыборочной и внутривыборочной дисперсиями и при этом не искажать расстояния между объектами. Построение таких алгоритмов – дело будущего, а пока можно рекомендовать предварительную обработку исходной матрицы данных методом главных компонент, отсечение дальних компонент с малыми дисперсиями и применение дискриминантного анализа к оставшимся первым нескольким главным компонентам. Достоверности различий между выборками резко упадут, но им можно будет верить.

Другим способом анализа величины и направления межвыборочной изменчивости может служить помещение всех выборок в компонентное

пространство одной из них. Прообразом этого способа обработки является метод профилей С.Р.Царапкина (Zarapkin, 1934; Царапкин 1960). В этом методе одна из групп (обычно самая представительная) принимается за стандарт. Средние значения других групп нормируются средними и среднеквадратичными отклонениями стандарта по формуле:

$$x'_i = \frac{x_i - x_{cm}}{s_{cm}}$$

что равносильно помещению центров тяжести других групп в центрированное и нормированное признаковое пространство стандарта.

После поворота признаков к главным компонентам, который, как мы знаем, не меняет расстояние между объектами, центры тяжести других групп оказываются в компонентном пространстве стандарта, что позволяет изучать не только величину, но и направление межпопуляционной изменчивости, а также взаимное расположение групп. Если за стандарт принять объединенную внутривыборочную матрицу, то рассматриваемый способ сводится к первому шагу дискриминантного анализа, но без преобразования Махаланобиса, что отнюдь не является недостатком.

Кроме того, никто не запрещает вычислить матрицу центроидов групп, рассматривая их как новые объекты, и исследовать ее методом главных компонент.

Необходимо особо отметить, что направления изменчивости в многомерном пространстве можно выбирать не только из статистических, но и непосредственно из биологических соображений, например, генетических. Для примера рассмотрим метод, позволяющих находить линейные комбинации признаков с максимальной наследуемостью в узком смысле. Метод основан на исследовании взаимного расположения родителей и гибридов первых двух поколений в многомерном пространстве и выделении направлений, обусловленных гетерозиготностью, эпистатическим и аддитивным действием генов (Efimov et al., 2005).

Хорошо известно, что фенотипическая изменчивость гибридов F_1 от скрещивания двух чистых линий является ненаследственной и только начиная с F_2 в изменчивости проявляется расщепление комплексов генов, полученных от обоих родителей. Пусть имеются две чистых линии, P_1 и P_2, и F_1 - первое поколение гибридов между ними, у которых измерены значения M признаков. В простейшей, аддитивно-доминатной модели без межаллельного взаимодействия средние значения каждого признака у F_1 равны $x^{F1}_i = m_i + h_i$, где $m_i = (x^{P1}_i + x^{P2}_i)/2$ - среднее между родителями, h_i - отклонение, обусловленное доминированием (Мазер, Джинкс, 1985).

В результате расщепления в следующем поколении средние значения гибридов F_2 будут равны (Мазер, Джинкс, 1985) $x^{F2}_i = m_i + h_i/2 = (m_i + x^{F1}_i)/2$ и в n-ом - $x^{Fn}_i = m_i + h_i f(n)$, где $f(n)$ – доля гетерозигот на локус в зависимости от системы скрещивания (самооплодотворение, инбридинг и т.д.)

Обозначим через $x^F = (x^F_1, x^F_2, \ldots, x^F_m)$ точку в многомерном пространстве, образованную средними значениями признаков для каждого поколения ($F = P_1, P_2, F_1, F_2, \ldots, F_n$). Из простых геометрических соображений следует, что точки x^{P1}, m, x^{P2} и x^{F1} образуют треугольник, в котором точки x^{Fi} расположатся на прямой линии, проходящей через точку x^{F1} и точку $m = (x^{P1} + x^{P2})/2$ - середину отрезка, соединяющего родительские средние. Точка x^{F2} попадет на середину отрезка,

соединяющего точки x^{P1} и m, а остальные точки x^{F1} будут стремиться к точке m со скоростью, зависящей от системы скрещивания (рис. 5.1).

При отклонении от аддитивно-доминантной модели наследования, например, вследствие межлокусного взаимодействия – эпистаза в широком смысле – ситуация усложняется и x^{F2}, вообще говоря, может оказаться в любой другой точке признакового пространства, в том числе и выходя за пределы плоскости, проходящей через x^{P1}, x^{P2} и x^{F1}. В этом случае приходится анализировать взаимное расположение выборок в трехмерном пространстве. Однако направление x^{F1}–x^{F2} в многомерном пространстве все равно будет обладать следующими свойствами. По мере расщепления гибридов будут исчезать все эффекты, связанные с гетерозиготностью, то есть, эффекты доминирования и все эпистатические эффекты, определяемые гетеро-гомозиготными и гетеро-гетерозиготными межлокусными взаимодействиями (Мазер, Джинкс, 1985). Поэтому направление x^{F1}–x^{F2} с полным основанием можно назвать "осью гетерозиготности". Аддитивное действие генов и оставшиеся эпистатические эффекты, определяемые гомо-гомозиготными межлокусными взаимодействиями, проявятся в проекции на плоскость, ортогональную x^{F1}–x^{F2} и проходящую через x^{P1} и x^{P2}. На этой плоскости центроиды F_1 и F_2 образуют одну точку. При справедливости аддитивно-доминантной модели эта точка должна совпасть с точкой m – серединой центроидов P_1 и P_2.

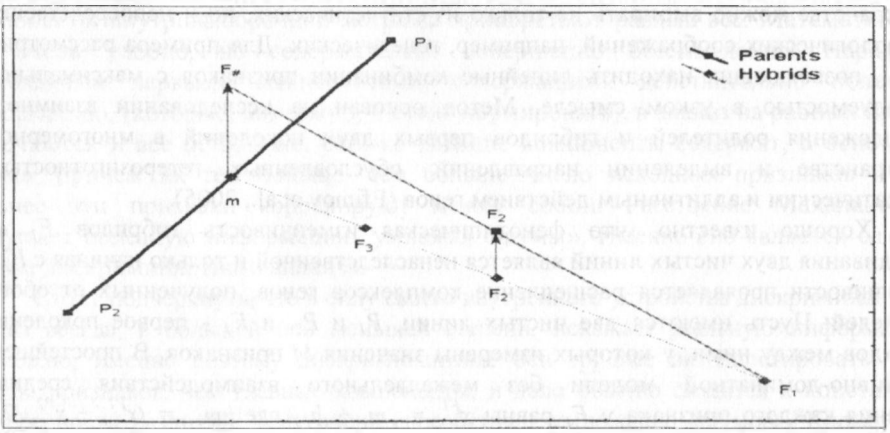


Рис. 5.1. Расположение центроидов родительских и гибридных выборок в многомерном пространстве. ***F1 - m*** – ось гетерозиготности в рамках аддитивно-доминантной модели наследования количественных признаков. Общий случай (HIA-модель): ***F1 - F#*** – ось гетерозиготности ***H***; ***P1 - P2*** – ось аддитивности ***A***; ***m - F#*** – ось эпистаза ***I***

Поэтому отклонение от нее в этой плоскости можно, в первом приближении, рассматривать как проявление эпистатических взаимодействий и, соответственно, назвать "эпистатическим направлением". Оставшееся ортогональное направление, в проекции на которое точка $x^{F1}(x^{F2})$ уже совпадает с точкой m, также с большой долей условности, можно именовать "аддитивным". Возможная аддитивность должна проявиться в увеличении дисперсии F_2 по сравнению с F_1.

Направления, обусловленные аддитивным действием генов, наиболее

подходят для отбора и его можно вести уже в F_2, не дожидаясь дальнейшего расщепления. Кроме того, найденные направления изменчивости можно использовать как новые перспективные признаки при изучении природных популяций.

ЛЕКЦИЯ 6. Внешние факторы как возможные причины изменчивости

Пусть теперь матрица X разбита на две группы признаков – внутренних Y и внешних Z. Внешние признаки называются факторами. Разбиение задается исследователем, исходя из содержательной задачи. Факторы, возможно, влияют на изменчивость между объектами. Цель анализа – определить характер и степень этого влияния, а также выяснить возможности предсказания характеристик объектов – значений внутренних признаков – по значениям внешних факторов.

Множественная линейная регрессия и ее проблемы

Обычно внутренние признаки рассматриваются поодиночке. В множественной регрессии такой внутренний признак называется зависимой переменной, а совокупность внешних факторов – независимыми переменными. Имеется некоторая совокупность объектов, которую можно считать генеральной. Для всех объектов мы знаем значения внешних факторов. Для части объектов (обучающая выборка) мы знаем значения внутренних признаков. Мы хотим узнать, можно ли по этой информации вычислить с приемлемой точностью значения внутренних признаков для остальных объектов. В классической постановке задача сводится к поиску линейной комбинации независимых переменных, в максимальной степени аппроксимирующих зависимую. Традиционно задача решается методом наименьших квадратов.

Здесь есть одно принципиальное обстоятельство. Применяя метод наименьших квадратов, мы наилучшим образом аппроксимируем линейную зависимость на обучающей выборке и почему-то думаем, что она останется наилучшей и для остальных объектов. Практика показывает, что это не так. Если проверять результаты расчетов на контрольной выборке, где на самом деле нам известны значения зависимых переменных, но мы не использовали их для расчета уравнения регрессии, то всегда оказывается, что до определенного числа параметров точность предсказания растет, а затем падает, хотя аппроксимация обучающей выборки становится все лучше и лучше. Чем больше факторов и чем выше корреляция между ними, тем хуже работает метод наименьших квадратов.

Как и в дискриминантном анализе, проблема состоит в возможной вырожденности или плохой обусловленности матрицы Z (проблема мультиколлинеарности независимых переменных). Точно так же возможен аварийный останов вычислений или окончание работы с непредсказуемым искажением результатов. На самом деле, это не очень удивительно, так как дискриминантный анализ формально можно рассматривать как частный случай множественной регрессии. Так же, как и в дискриминантном анализе, обычная рекомендация заключается в том, чтобы исключить из анализа высоко коррелирующие признаки. Например, это можно сделать с помощью пошагового анализа (Боровиков, Боровиков, 1997). И точно так же можно поставить вопрос: а

может, дело не в признаках, а в самом анализе?

Например: имеется несколько сотен образцов бензина, для которых известны результаты лабораторных анализов качества (октановое число, содержание свинца и т.д.). Каждый образец можно легко и быстро проанализировать с помощью инфракрасного спектрометра и получить значения нескольких тысяч факторов. Так как число факторов превышает число объектов, матрица обязательно Z будет вырожденной и классическая множественная линейная регрессия просто не сработает. Рекомендация "выбросить признаки" тоже неприемлема, так как означает выбросить почти все признаки.

Один из методов решения проблемы – регрессия на главные компоненты. Применим к Z метод главных компонент, то есть найдем матрицу $U=ZQ$. Очевидно, что матрицу U можно рассматривать как матрицу новых независимых переменных. С вычислительной точки зрения это даже очень удобно, так столбцы матрицы U (главные компоненты) не коррелируют между собой, и регрессия распадается на сумму регрессий зависимой переменной от каждого столбца матрицы U, которые можно вычислять независимо друг от друга. При этом в методе наименьших квадратов обязательно происходит нормировка каждого столбца матрицы U его дисперсией, а в случае плохой обусловленности или вырожденности матрицы Z часть этих дисперсий мала или равна нулю. Как и в дискриминантном анализе, такие столбцы не несут содержательного смысла и могут рассматриваться, как заглушающие полезную информацию. Очевидно, их можно и нужно выбросить. Что считать малой дисперсией, решает исследователь. Число оставшихся компонент всегда меньше числа объектов, но в каждую из них теоретически могли внести вклад все факторы.

Дальнейшим развитием этой идеи является PLS-регрессия (проекция на латентные структуры). Основная идея заключается в том, чтобы позаботиться о хороших предсказательных свойствах уравнения регрессии заблаговременно. А для этого учитывается не только качество аппроксимации, но и дисперсия линейной комбинации, на основе которой делается предсказание. Чем выше ее дисперсия, тем надежнее работает линейная регрессия. Это означает, что максимизируется не коэффициент корреляции (что эквивалентно методу наименьших квадратов), а коэффициент ковариации между зависимой переменной и аппроксимирующей ее линейной комбинацией независимых переменных. Если предварительно преобразовать матрицу Z методом главных компонент, то решение получится в виде суммы одиночных регрессий зависимой регрессии на компоненты с весами, пропорциональными дисперсиям компонент. Для компонент с нулевыми дисперсиями это эквивалентно их исключению из анализа, для компонент с малыми дисперсиями – малое влияние на окончательный результат.

Хемометрики активно используют PLS-регрессию последние два десятка лет, в том числе, и на производстве. Нефтяные и пивоваренные компании, применившие эту технологию, получили экономию в сотни тысяч евро в год. Биологи, к сожалению, в большинстве своем даже не знакомы с таким вариантом множественной регрессии.

Однако и PLS-регрессия представляется небезупречной. Создается впечатление, что ее практический успех обусловлен, прежде всего, тем, что она оказалась явно лучше классической линейной регрессии. Но проведем мысленный

эксперимент. Представим себе, что мы берем один из внешних факторов и размножаем его в большом количестве. Никакой новой информации, очевидно, не добавляется. Однако веса компонент, в которые входит этот фактор, будут расти и, следовательно, будет расти вклад этого фактора в окончательное уравнение регрессии независимо от того, насколько он лучше остальных. По-видимому, нужно каким-то образом ограничить предельный вес дисперсии компоненты в уравнении регрессии.

В любом случае необходимо разбиение объектов на обучающую и контрольную выборки, например, с помощью бутстреп-методов (Efron, 1979, 1982; Диаконис, Эфрон, 1983) (лекция 7). Любая зависимость, установленная на обучающей выборке, должна проверяться на контрольной. Только так можно обеспечить надежность содержательных выводов.

ЛЕКЦИЯ 7. Нелинейные методы, неевклидовы расстояния

Все методы, рассмотренные в предыдущих лекциях, относятся к числу линейных, то есть объекты предполагаются размещенными в евклидовом пространстве, а направления задаются линейными комбинациями исходных признаков. Однако, даже если каждый объект и задается своими значениями в пространстве признаков, расстояние между ними не обязано быть евклидовым, а направления – линейными. Кроме того, нелинейной может быть и регрессия, как функция зависимой переменной от нескольких независимых. В качестве варианта нелинейной регрессии можно рассматривать нейронные сети.

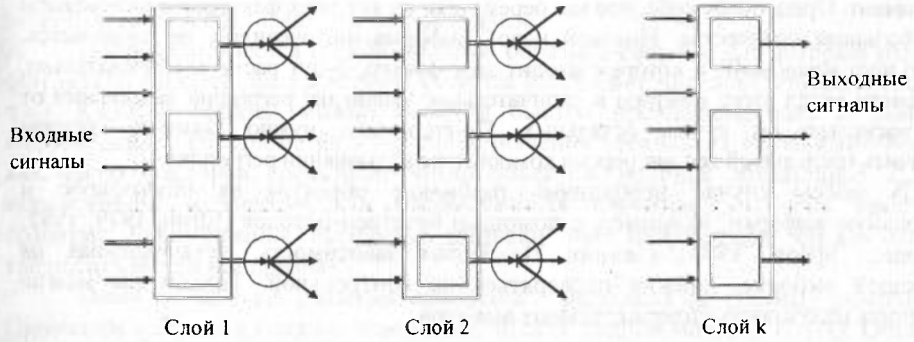
Нейронные сети

В последние годы интерес к искусственным нейронным сетям необычайно высок. Однако, несмотря на обилие описательной информации, библиотек программ для моделирования нейронных сетей не так уж много.

Под искусственной нейронной сетью понимается некоторое вычислительное устройство обработки информации, состоящее из большого числа параллельно работающих простых процессорных элементов – нейронов, связанных между собой линиями передачи информации – связями или синапсами. У нейронной сети выделена группа связей, по которым она получает информацию из внешнего мира, и группа выходных связей, с которых снимаются выдаваемые сетью сигналы. Нейронная сеть обучается решению задачи на основании некоторой обучающей выборки – "задачника", состоящего из набора пар "вход – требуемый выход", проверяется на контрольном наборе данных, имеющем ту же структуру, и далее способна решать примеры, не входящие в обучающую выборку (Горбань, 1990; Горбань, Россиев, 1996; Principal Manifolds, 2007). Именно структурные аналогии с устройством реального мозга и наличие процесса адаптации к предъявляемым ситуациям (обучение) дали нейроинформатике название, основные идеи и термины, заимствованные, в основном, из нейробиологии и нейрофизиологии.

Архитектура нейронных сетей

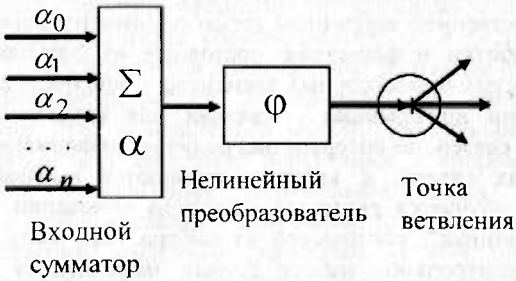
Описаны только слоистые нейронные сети как наиболее простые среди всего множества нейросетевых архитектур.



В слоистых сетях нейроны расположены в несколько слоев. Нейроны первого слоя получают входные сигналы, преобразуют их и через точки ветвления передают нейронам второго слоя. Далее срабатывает второй слой и т.д. до ***k***-го слоя, который выдает выходные сигналы. Обычно каждый выходной сигнал ***i***-го слоя подается на вход всех нейронов ***i+1***-го. Число нейронов в каждом слое может быть любым и никак заранее не связано с количеством нейронов в других слоях. Стандартный способ подачи входных сигналов: каждый нейрон первого слоя получает все входные сигналы. Особое распространение получили трехслойные сети, в которых каждый слой имеет свое наименование: первый – входной, второй – скрытый, третий – выходной.

Нейрон

Нейроны, используемые в большинстве нейронных сетей, имеют следующую структуру.



Веса адаптивных связей при создании сети принимают случайные значения и при обучении сети могут изменяться в диапазоне [-1,1].

В качестве нелинейного элемента нейрона часто используется нелинейный сигмоидный преобразователь $\varphi(A)=A/(c+|A|)$, где A - выход сумматора нейрона, а константа c – параметр крутизны сигмоиды. Параметр крутизны можно задавать отдельно для каждого слоя сети.

Каждый слой рассчитывает нелинейное преобразование от линейной комбинации сигналов предыдущего слоя. Отсюда видно, что линейная функция активации может применяться только для тех моделей сетей, где не требуется последовательное соединение слоев нейронов друг за другом. Для многослойных сетей функция активации должна быть нелинейной, иначе можно построить

эквивалентную однослойную сеть, и многослойность оказывается ненужной. Если применена линейная функция активации, то каждый слой будет давать на выходе линейную комбинацию входов. Следующий слой даст линейную комбинацию выходов предыдущего, а это эквивалентно одной линейной комбинации с другими коэффициентами, и может быть реализовано в виде одного слоя нейронов.

Многослойная сеть может формировать на выходе произвольную нелинейную многомерную функцию при соответствующем выборе количества слоев, диапазона изменения сигналов и параметров нейронов за счет поочередного расчета линейных комбинаций и нелинейных преобразований. Поэтому многослойные сети оказываются универсальным инструментом аппроксимации функций (Горбань, 1998).

В общем виде, задачи, которые решают нейронные сети, можно разбить на два основных вида: *классификация* и *прогнозирование*. В задачах *классификации*, как правило, нужно определить, к какому из нескольких заданных классов принадлежит данный входной набор. Примером может служить медицинский диагноз, который ставится на основании анализов. В задачах *прогнозирования* требуется предсказать значение переменной, принимающей, как правило, непрерывные числовые значения, например, заболеваемость туберкулезом на следующий год. В этом случае в качестве выходных данных требуется одна количественная переменная.

Нейросеть может решать одновременно несколько задач как прогнозирования (предсказания значений нескольких количественных признаков), так и задачи классификации (предсказания состояний нескольких качественных признаков), так и задачи прогнозирования и классификации одновременно.

Для каждой из задач могут быть установлены свои требования к точности. Для прогнозируемого качественного признака точность означает максимально допустимое отклонение прогноза сети от истинного значения признака. Желательно задавать как можно менее жесткие требования к точности. Это ускорит как процесс обучения, так и процесс упрощения сети. Также задачу можно будет решить на основе нейронной сети с меньшим числом слоев или нейронов, и, обычно, на основании меньшего числа входных сигналов. Требуемая точность ни в коем случае не должна превосходить погрешностей получения сигнала (погрешностей измерительных приборов, погрешностей огрубления значений при вводе их в компьютер). Так, если значение признака изменяется в диапазоне [0,10] и измерительный прибор имеет собственную точность <0.1, то нельзя требовать от сети предсказания с точностью <0.01.

Для качественного признака точность (надежность) означает уверенность в принадлежности качественного признака тому или иному дискретному состоянию. Чем больше уровень требуемый уверенности, тем более надежно должна сеть диагностировать отличия каждого дискретного состояния от других.

Примеры применений нейронных сетей

Методы нейронных сетей можно использовать в любой ситуации, где требуется найти значения неизвестных переменных или характеристик по известным данным наблюдений или измерений (сюда относятся различные задачи регрессии, классификации и анализа временных рядов), причем этих исторических данных должно быть достаточное количество, а между известными и неизвестными

значениями действительно должна существовать некоторая связь или система связей (нейронные сети довольно устойчивы к помехам). Подробное обсуждение теоретических аспектов вопроса о том, когда применение нейронных сетей должно быть успешным, можно найти, например, в электронном учебнике по статистике *STATISTICA Neural Networks*. Далее приводится достаточно представительный, однако далеко не исчерпывающий набор примеров из разных областей, где применяются нейронные сети:

– оптическое распознавание символов, включая распознавание подписи (в частности, разработана система идентификации подписи, учитывающая не только окончательный ее рисунок, но и скорость ручки на различных участках, что значительно затрудняет подделку чужой подписи);

– обработка изображений (например, система сканирует видеоизображения станций лондонского метро и определяет, насколько станция заполнена народом, причем работа системы не зависит от условий освещенности и движения поездов);

– прогнозирование финансовых временных рядов (компания LBS Capital Management объявила о значительных успехах в финансовых операциях, достигнутых за счет прогнозирования цен акций с помощью многослойных персептронов);

– геологоразведка: анализ сейсмических данных, ассоциативные методики поиска полезных ископаемых, оценка ресурсов месторождений.

Нейросети используются фирмой Amoco для выделения характерных пиков в показаниях сейсмических датчиков. Надежность распознавания пиков - 95% по каждой сейсмо-линии. По сравнению с ручной обработкой скорость анализа данных увеличилась в 8 раз;

– медицинская диагностика (например, прогнозирование эпилептических припадков, определение размеров опухоли простаты).

Группа НейроКомп из Красноярска (под руководством Александра Николаевича Горбаня) совместно с Красноярским межобластном офтальмологическом центром им. Макарова разработали систему ранней диагностики меланомы сосудистой оболочки глаза. Этот вид рака составляют почти 90% всех внутриглазных опухолей и легко диагностируется лишь на поздней стадии. Метод основан на косвенном измерении содержания меланина в ресницах. Полученные данные спектрофотометрии, а также общие характеристики обследуемого (пол, возраст и др.) подаются на входные синапсы 43-нейронного классификатора. Нейросеть решает, имеется ли у пациента опухоль, и если да, то определяет ее стадию, выдавая, кроме этого, процентную вероятность своей уверенности (http://www.chat.ru/~neurocom/);

– синтез речи (знаменитая экспериментальная система Nettalk, способная произносить фонемы из написанного текста);

– прогнозирование хаотических временных рядов (целый ряд исследований продемонстрировал хорошие способности нейронных сетей к прогнозированию хаотических временных данных);

– автоматизация производства: оптимизация режимов производственного процесса, комплексная диагностика качества продукции (ультразвук, оптика) мониторинг и визуализация многомерной диспетчерской информации, предупреждение аварийных ситуаций, робототехника.

Ford Motors Company внедрила у себя нейросистему для диагностики двигателей после неудачных попыток построить экспертную систему, т.к. хотя опытный механик и может диагностировать неисправности, он не в состоянии описать алгоритм такого распознавания. На вход нейро-системы подаются данные от 31 датчика. Нейросеть обучалась различным видам неисправностей по 868 примерам. "После полного цикла обучения качество диагностирования неисправностей сетью достигло уровня наших лучших экспертов, и значительно превосходило их в скорости";

– лингвистический анализ (пример: сеть с неконтролируемым обучением используется для идентификации ключевых фраз и слов в языках туземцев Южной Америки).

Из приведенного списка видно, что специфика объекта не играет никакой роли и не накладывает никаких предметных ограничений на применение нейронных сетей. В то же время пока они сравнительно мало используются в биологических, экологических и медицинских исследованиях. В ближайшее время надо ожидать бурного роста работ по применению нейронных сетей и в этих научных областях.

Неевклидовы расстояния

Исследователь вправе выбрать любое расстояние (меру сходства или различия), которое считает нужным, исходя из содержательных соображений. Например, в зоогеографических исследованиях часто применяется индекс сходства Жаккара-Наумова между вариантами населения. Большой список индексов сходства и мер различия приведен в работе Ю.А.Песенко (1982). Уместно заметить, что мера сходства между признаками – коэффициент корреляции Браве-Пирсона – тоже неевклидова, если рассматривать их как объекты в двойственном пространстве. Однако методы работы с неевклидовыми расстояниями разработаны гораздо хуже.

Термокарты (heatmaps) и иерархическая кластеризация

Пусть имеется таблица "объект – признак". Простейший способ получить визуальное представление о всей таблице сразу – это ее раскрасить (рис. 7.1, 7.2, пример условный). Раскраска осуществляется следующим образом. Каждому значению таблицы сопоставляется отдельная клетка. Клетка раскрашивается в зеленый цвет, если значение меньше среднего (по столбцу), и в красный, если значение больше. Причем, чем больше значение по абсолютной величине, тем цвет ярче. В черный (или серый, или белый) красятся клетки, значении в которых близки к среднему. Иногда вместо зеленого и красного используются синий и желтый цвета – для лиц с ограниченным цветовосприятием (дальтоников). Раскраска таблиц широко применяется в работах молекулярных генетиков, однако нет никаких причин не применять ее в других областях биологии, где требуется кластерный анализ.

Но раскрашенная таблица выглядит очень пестро, если ее не структурировать. Для этого используется кластерный анализ. Кластерный анализ – это разбиение исходного множества объектов на классы таким образом, чтобы близкие объекты попали в одни и те же классы, а далекие – в разные. Мера сходства или различия может быть измерена в количественной или даже ранговой шкале. Один из самых популярных способов структурирования – иерархическая классификация. "Иерархическая" означает, что каждый класс вложен в некоторый

другой. Самый известный и часто используемый алгоритм иерархической классификации – алгоритм ближайшего соседа или единственной связи. Вначале каждый объект считается отдельным классом. На следующем шаге ищется пара самых близких объектов, которая объединяется в новый класс. Расстояния (или меры сходства) для нового класса со старыми пересчитываются по следующему правилу: расстоянием между классами считается расстояние между ближайшими объектами в этих классах (отсюда и название). Далее все повторяется до тех пор, пока не останется ровно один класс, содержащий все объекты. Если за расстояние между классами принять расстояние между самыми далекими объектами в этих классах, то получим метод дальнего соседа или полной связи. Можно также за расстояние между классами принять среднее расстояние между объектами этих классов, тогда получим метод UPGMA или средней связи. Несколько особняком стоит метод Уорда, в котором учитывается еще и разброс объектов внутри кластера.

Общепринятым способом отобразить иерархическую классификацию является дендрограмма (рис. 7.1, 7.2). Объекты играют роль листьев и расположены каждый на своей ветке. Если объекты объединяются в один класс, то и их ветви объединяются в одну, причем длина равна расстоянию (или сходству) между классами. Чтобы дендрограмму можно было нарисовать, объекты надо переставить местами. Если одновременно переставить строки таблицы и термокарты, то результат будет более нагляден (рис. 7.1). (При этом не следует думать, что получившаяся дендрограмма хоть каким-то образом отражает линейное упорядочение объектов. Любые две объединяющиеся ветви можно всегда поменять местами (вместе со всеми подветками и листьями), а это приведет совсем к другому упорядочению (рис. 7.2).) Классифицировать можно и признаки, в этом случае надо переставлять столбцы таблицы и термокарты.

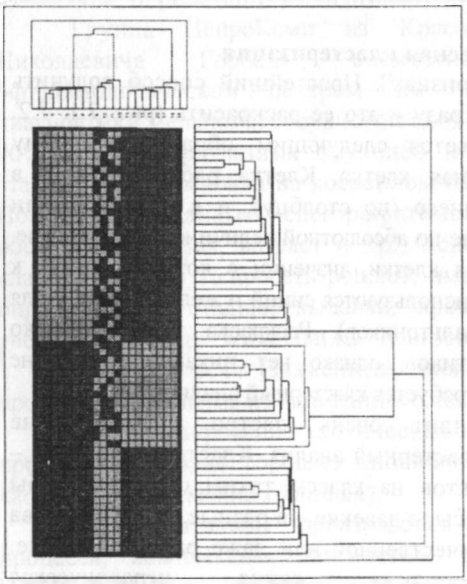


Рис. 7.1. Термокарта и дендрограммы (условный пример)

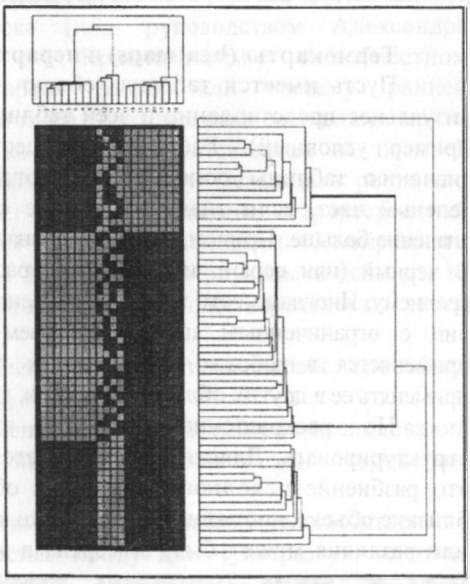


Рис. 7.2. Термокарта и дендрограммы после перестановки строк

Практика показывает, что дендрограммы, полученные различными методами на одних и тех же данных, могут не слишком походить друг на друга.

Алгоритм К-средних

На начальном этапе случайным образом выбирается K объектов (K задается исследователем). Они объявляются центрами классов. Остальные объекты разносятся по классам по следующему правилу: каждый объект попадает в тот класс, к центру которого он находится ближе всего. После этого в каждом классе определяется новый центр. Снова все объекты разносятся по классам и так до тех пор, пока процесс не сойдется.

В отличие от иерархической классификации, все классы равноправны и находятся на одном уровне. Еще одна особенность – классы не обязательно удовлетворяют так называемому условию "компактности", т.е. не являются "хорошими" или "естественными" в том смысле, что ближайшими к некоторым объектам одного класса, могут быть объекты из другого класса. Поэтому некоторые авторы предпочитают называть его алгоритмом группировки, а не классификации. Есть критерии, позволяющие оценить удачность разбиения на классы (аналогичные методу Уорда). Если разбиение оказалось не очень удачным, K необходимо изменить и весь процесс повторить с другим К. Рекомендуется для начала брать K равным квадратному корню из числа объектов, однако это сильно зависит от исследуемого множества.

Многомерное шкалирование

В кластерном анализе активно эксплуатируется понятие близости между объектами. По существу, весь анализ базируется на том, что одни объекты ближе друг к другу, чем другие. При этом основные принципы кластерного анализа могут слегка нарушаться. Например, во многих алгоритмах, таких, как метод К-средних или метод Уорда, дополнительно вычисляется центр кластера как среднее координат входящих в него объектов. При этом неявно предполагается, во-первых, что усреднение не выводит центр за пределы кластера и он тоже может считаться равноправным с другими объектом, и, во-вторых, что такой центр в некотором смысле минимизирует максимальное расстояние от себя до объектов кластера и поэтому может считаться наилучшим представителем всего кластера. Вообще говоря, ни то, ни другое ниоткуда не следует. Теоретически можно придумать и такие множества объектов и такие меры близости, что оба эти предположения будут нарушаться, причем как угодно сильно. Однако на практике эти алгоритмы довольно успешно работают.

Большим недостатком кластерного анализа является то, что он не дает информации о взаимном расположении объектов и образованных ими кластеров. Это резко сужает возможности исследователя по интерпретации получаемых результатов. На самом деле, такая информация, как правило, присутствует в исходных данных, просто кластерный анализ ее игнорирует. Однако существуют другие методы, которые активно используют геометрические представления для решения стоящих перед исследователями содержательных задач. В частности, к ним относятся методы многомерного шкалирования.

В этих методах исходные координаты объектов используются только для

того, чтобы вычислить матрицу коэффициентов различия между объектами. Возможна ситуация, когда координаты объектов не заданы, а вместо этого сразу дана матрица расстояний (количественный признак на парах объектов) или различий (ранговый признак). Если задана матрица сходства, то ее всегда можно преобразовать в матрицу различий, например, взяв с обратным знаком. На выходе требуется получить небольшое число латентных переменных, описывающих объекты в некотором удобном пространстве с хорошей метрикой, удовлетворяющем аксиомам расстояния: рефлексивности, симметричности, аксиоме треугольника – например, в метрике Минковского,

$$d_{xy} = (\sum | x_k - y_k |^p)^{1/p},$$

частным случаем которой является евклидова метрика (при $p = 2$). Критерием служит соответствие между расстояниями в этом пространстве и исходной матрицей сходства-различия между объектами.

Хотя в литературе огромное внимание уделено метрическому шкалированию, на сегодня можно смело утверждать, что этот подход устарел. Неметрические оценки сходства-различия в экспериментальных ситуациях получить гораздо проще. Достаточно просто определить любую содержательно подходящую меру сходства между объектами, не заботясь о формальном соответствии свойствам расстояния, и неметрическое шкалирование все равно метризует пространство объектов.

Поэтому последние несколько десятков лет, в основном, используется неметрическое шкалирование в квазиметрическом варианте, восходящем к Крускалу (Kruskal, 1964a, 1964б), хотя оно требует очень много машинного времени и поэтому число объектов, которое можно обработать этим методом на персональных компьютерах с помощью профессиональных статистических пакетов не превышает сотни.

Пусть имеется конечное множество объектов и матрица R расстояний или мер сходства между ними, а также произвольное представление объектов этого множества в виде точек в метрическом пространстве размерности K с метрикой d. Определим критерий различия между множеством и его представлением ("стресс" по Крускалу) в виде

$$H(R,D) = \sum_{i,j} (f(r_{ij}) - d_{ij})^2,$$

где f – некоторое монотонное преобразование.

В алгоритме Крускала ищется такое представление, для которого функция H принимает наименьшее возможное значение. Это приводит к задаче минимизации H как функции многих переменных от координат, например, методом сопряженных градиентов.

Популярность этого метода объясняется исключительно тем, что ему не было альтернативы. Ситуация радикально изменилась после появления работы Й.Тагучи и Й.Ооно (Taguchi, Oono, 2005), в которой произошел возврат к первоначальной идее Р.Шепарда (Shepard, 1962) и неметрическому шкалированию, образно выражаясь, вернули права гражданства. Теперь речь идет об обработке тысяч и десятков тысяч объектов без потери качества метризации, что открывает огромные перспективы для исследователей во всех областях знаний.

Алгоритм Шепарда-Тагучи-Ооно работает следующим образом. Исходные оценки различия ранжируются. Выбирается размерность и метрика результирующего пространства. В этом пространстве случайным образом помещается совокупность *N* точек, каждая из которых соответствует одному объекту. Между ними вычисляется матрица расстояний, которая также ранжируется. Каждой из *N*(N-1)/2* пар объектов соответствует два ранга, в одной и другой ранжировке. Если ранжировки полностью соответствуют друг другу, то первый этап работы алгоритма закончен. Если нет, то имеется пара объектов, для которых ранги в двух ранжировках различны. Если ранг расстояния в результирующем пространстве больше ранга различия той же пары объектов в исходной матрице, то точки, представляющие объекты, чуть-чуть сдвигаются друг к другу, если меньше – раздвигаются. После прохождения всех пар объектов расстояния между точками результирующего пространства пересчитываются и ранжируются заново. Процесс продолжается до тех пор, пока сходство между ранжировками, например, ранговый коэффициент корреляции Спирмена, не перестанет расти. Если оно слишком мало, размерность пространства увеличивается на единицу и весь процесс повторяется. Скорость этого алгоритма оказалась, по меньшей мере, на порядок больше, чем алгоритма Крускала, что позволяет обрабатывать значительно большее число исходных данных.

Почему ранговые оценки сходства различий позволяют с такой большой точностью восстановить метрическую структуру данных? На этот вопрос лучше всего ответил сам автор неметрического шкалирования. "Парадоксальная возможность восстановления количественной структуры из качественных данных связана с тем обстоятельством, что число пар точек и, следовательно, число порядковых ограничений на их расстояния возрастает приблизительно как квадрат числа определяемых количественных координат точек. Такие методы называются «неметрическими», поскольку в этом случае используются только порядковые свойства входных данных. Однако выход может достигать большой метрической точности и всегда будет метричным в смысле соответствия аксиомам расстояния" (Шепард, 1980).

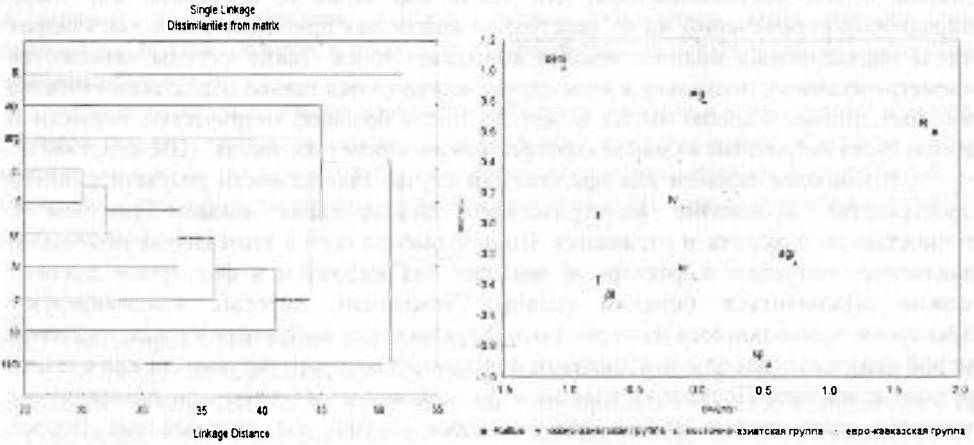
В наиболее важном для приложений случае евклидовости результирующего пространства алгоритмы неметрического шкалирования выдают решение с точностью до поворота и отражения. Вопрос выбора осей в этом случае полностью аналогичен ситуации в факторном анализе. Так же, как и в факторном анализе, можно ограничиться поиском главных компонент, которые максимизируют дисперсии, приходящиеся на первые оси. Можно также выбрать оси с максимальной мерой сходства с исходными шкалами для лучшей интерпретируемости или сделать ручное вращение. Поскольку взаимное расположение объектов при поворотах не меняется, исследователь вправе принять любое удобное для него решение. Вопрос выбора метрики результирующего пространства и его размерности – тоже его личное дело. Размерность можно задавать в явном виде, а можно через величину коэффициента сходства ранжировок, которую необходимо достигнуть в ходе вычислений.

Следует отметить, что алгоритмы неметрического шкалирования обладают одним весьма важным свойством. Если в качестве исходной меры близости между объектами-объектами взять евклидово расстояние, то при размерности результирующего пространства, равной реальной размерности исходного пространства, алгоритм воспроизведет исходную конфигурацию объектов (с

точностью до поворота и отражения). Применение метода главных компонент позволяет однозначно определить оси и их дисперсии, которые будут совпадать с результатом прямого применения метода главных компонент к исходным данным. Его можно заложить прямо в алгоритм многомерного шкалирования, что обычно и делается. В этом случае, метод главных компонент выглядит как частный случай неметрического шкалирования. Поскольку неметрическое шкалирование обладает гораздо большей общностью, следует ожидать, что в конечном итоге оно может полностью вытеснить прямой метод главных компонент из практики обработки и последний останется только как способ однозначного выбора осей в результирующем евклидовом пространстве после многомерного шкалирования. Исключение может быть для случая малого числа признаков и очень большого числа объектов.

Если же размерность результирующего пространства меньше размерности исходного пространства, то произойдет "вминание" множества точек из пространства большей размерности в пространство меньшей, но с максимально возможным сохранением расстояний между ними. (В методе главных компонент точки проецируются из пространства большей размерности на пространство меньшей.) Какие возможности предоставятся исследователям, и какие ограничения возникнут, еще предстоит исследовать.

Содержательная интерпретация полученных результатов, в силу наглядности представления, обычно не вызывает особенных затруднений. Рассмотрим, например, матрицу дивергенции (%) 402 пн участка гена цитохрома *b* мтДНК лесных и полевых мышей рода *Apodemus*, приведенную в статье Челоминой и др. (Генетика, 1998, т.34, №5, 650-661) .



В статье на основании дендрограммы приводятся доводы в пользу выделения в роде Apodemus трех групп видов. Если обработать ту же матрицу дивергенции методами *K*-средних и двумерного шкалирования, то результаты получаются гораздо убедительнее.

Существуют и другие, более сложные, модели неметрического шкалирования, основанные на одновременном рассмотрении целого ряда матриц (Caroll, Chang, 1970; Caroll, 1976), на которых мы не будем останавливаться.

Бутстреп

Bootstrap (англ.) – ремешок на заднике ботинка, облегчающий его надевание. В английском языке существует идиома – lift oneself by one's own bootstrap – дословно, поднять самого себя за ремешок собственного ботинка. В переносном смысле – выбиться в люди благодаря собственным усилиям; самому пробить себе дорогу; быть всем обязанным самому себе. В статистике так называется процедура, предложенная Б. Эфроном (Efron, 1979, 1982; Диаконис, Эфрон, 1983). Предположим, что у нас есть данные и некоторая последовательность вычислительных действий, например, с использованием нейронных сетей или многомерного шкалирования или любых других эвристических алгоритмов. Мы хотим иметь представление о статистической устойчивости результатов расчета.

Если бы у нас было достаточно много случайных выборок из одной и той же генеральной совокупности, то задача решалась бы просто. Мы бы провели этот расчет на каждой выборке и получили бы распределение, а, следовательно, среднее значение, дисперсию и доверительные интервалы для каждой характеристики, которая нас интересует. Проблема состоит в том, что у нас, как правило, имеется только одна выборка. Обычно в этом случае, за неимением лучшего, статистики предполагают (неизвестно откуда) известным генеральное распределение характеристики и считают неизвестными только значения параметров распределения, которые и оценивают по выборке.

Б.Эфрон нашел другой путь. Он предложил размножать исходную выборку. Пусть она состоит из N элементов. Новую выборку получим следующим образом. С помощью датчика случайных чисел с равными вероятностями выберем любой элемент исходной и включим его копию в новую выборку. Повторим процесс N раз. Выборка сформирована.

Новая выборка почти наверняка будет отличаться от исходной, потому что одни элементы исходной выборки случайно несколько раз попадут в новую выборку, другие – ни разу. Поэтому можно получить столько новых выборок, сколько потребуется. Подавляющее большинство из них будет отличаться от исходной выборки и друг от друга.

За прошедшие десятилетия бутстреп-метод изучен вдоль и поперек. Основные выводы заключаются в следующем. Среднее значение, полученное по совокупности новых выборок, будет смещено по сравнению с генеральным средним и не будет его оценкой, так как оно, естественно, будет колебаться вокруг среднего исходной выборки. А вот форма распределения и его дисперсия будут очень близки к генеральным для произвольного вида распределения. Поэтому для выборочного среднего можно получить достаточно надежные оценки доверительных интервалов.

ЛЕКЦИЯ 8. Временные ряды

Устойчивость статистических связей

Основной проблемой при использовании статистических методов для анализа динамик численности животных, урожайности грибов и растений, метеофакторов и других временных рядов различной природы является значительная корреляция между соседними значениями, что не позволяет относиться к ним, как к независимым наблюдениям. По этой причине необходимо исследовать достаточно

длинные интервалы, так как часто наблюдаемая корреляция между разными рядами на коротких интервалах наблюдений может быть обусловлена наложением фаз при близких, но все же не совпадающих периодах колебаний и в дальнейшем рассыпаться и даже сменить знак на обратный. Для временных рядов с высокой автокорреляцией, к которым относятся, например, все циклические динамики численности, обычно применяются схемы авторегрессии типа

$$x_{t+1} = Ax_t + Bx_{t-1} + \ldots + Cz_t + Dz_{t-1} + \ldots + e_t \qquad (1)$$

где x_t - показатель численности в момент t,

z_t - внешний по отношению к популяции фактор,

e_t - остаток или "шум" без автокорреляции (Moran, 1953).

Кроме того, выбор класса схем и подгонка коэффициентов должны осуществляться на одной части статистического материала, а оценка соответствия – на другой (Колмогоров, 1933, 1986; Дрейпер, Смит, 1987). Для временных рядов обычно "зажимают" несколько последних элементов ряда, оценивают коэффициенты по оставшимся и проверяют степень расхождения на последней части ряда. По существу, мерой качества статистической модели является ее предсказательная сила. В качестве критериев адекватности избираются различные показатели: минимум дисперсии прогноза, близость спектральных и автоковариационных функций и т.д. (Кашьяп, Рао, 1983). Одним из возможных критериев является коэффициент корреляции.

Будем использовать следующую процедуру. Прогнозируемый ряд и предполагаемый предиктор (ряд, используемый для предсказания) разобьем на две части: обучающую и проверочную. Длину проверочной части будем брать порядка двух десятков отсчетов, чтобы можно было воспользоваться аппроксимацией Р.Фишера для выборочного коэффициента корреляции (Большев, Смирнов, 1983). Подгонку параметров линейной регрессионной модели будем проводить на обучающей части рядов по минимуму суммы квадратов отклонений, что влечет максимизацию коэффициента корреляции между исходными и расчетными данными. Адекватность модели будем оценивать величиной коэффициента корреляции между проверочной частью исходных данных и соответствующими расчетными данными, вычисленными по подогнанным на обучающей части параметрам. В случае одной независимой переменной речь идет просто о вычислении коэффициентов корреляции между прогнозируемый рядом и предиктором на обучающей и проверочной частях отдельно. Предиктор проходит статистический тест и остается в списке для содержательного рассмотрения, если коэффициенты корреляции с прогнозируемым рядом по обеим частям отдельно достаточно велики и имеют одинаковый знак. Если каждый из коэффициентов корреляции превышает величину, соответствующую некоторому уровню значимости β, то вероятность для обоих коэффициентов превысить его одновременно, имея одинаковые знаки, равна $\beta^2/2$. Выбирая стандартный уровень значимости α (0.05; 0.01; 0.001), получим, что $\beta = \sqrt{2\alpha}$ (0.3162; 0.1414; 0.0447). Само собой разумеется, что значимой на уровне α должна быть и корреляция между предиктором и прогнозируемым рядом на всем интервале наблюдений. Уровень значимости одного коэффициента корреляции при заданном числе наблюдений n будем оценивать по уровню значимости нормально распределенной случайной

величины $z = 0.5\ln((1+r)/(1-r))$ $n-3$ (Большев, Смирнов, 1983).

В исследованиях динамик численности животных принято логарифмировать данные учетов или заготовок шкурок (Уильямсон, 1975). Это связано с тем, что при отсутствии внутренних и внешних ограничений рост популяции описывается уравнением:

$$\dot{X}_t = k()x_t,$$

где k – коэффициент воспроизводства, $\dot{X}_t$ – производная по времени.

Это же уравнение можно записать в виде $\dot{L}_t = f()$, где $L = \ln(x), f() = \ln(k)$ (Бигон и др., 1989). Дискретный аналог этого уравнения имеет вид

$$L_{t+1} = L_t + f(),$$

что совпадает с уравнением (1), если положить

$$f() = (A-1)L_t + BL_{t-1} + ... + Cz_t + Dz_{t-1} + ... + e_t.$$

Полученные ряды лучше описываются нормальным распределением.

Фазовые портреты. Теорема Такенса. Метод главных компонент для временных рядов

Другим способом анализа временных рядов для выявления внутренне присущих им закономерностей является разложение Карунена-Лоэва (метод главных компонент, разложение на естественные ортогональные составляющие, сингулярный спектральный анализ) (Ефимов и др., 1988; Главные компоненты ..., 1997; Бобрецов и др., 2000). Этот метод применим к любому временному ряду, не требует его стационарности, как, например, спектральный анализ, автоматически выявляет тренды, если они имеются, без каких-либо предположений об их природе и форме, и (последнее по счету, но не по важности) позволяет получать многомерные представления временного ряда – фазовые портреты – дающие возможность визуального изучения траектории ряда в многомерном пространстве его состояний.

Сущность метода заключается в следующем. Если временной ряд порождается некоторой динамической системой с конечным числом параметров, то совокупность его отрезков можно рассматривать как точки многомерного фазового пространства. Соединяя их последовательно, например, сплайнами, получим траекторию ряда в этом пространстве, которая, как следует из знаменитой теоремы Такенса (Takens, 1981), воспроизводит многомерный фазовый портрет динамической системы, если длина отрезков превышает удвоенное число параметров. Для редукции размерности динамических систем обычно используют преобразование Пуанкаре (Балеску, 1978). Однако можно применить метод главных компонент, заключающийся в поиске координатных осей, в проекции на которые дисперсия траектории ряда максимальна (Ефимов, Галактионов, 1983; Ефимов и др., 1988; Главные компоненты ..., 1997). Максимизация автоковариации вместо дисперсии приводит к методу гладких компонент. Оба метода оказываются исключительно полезны при анализе внутренних закономерностей и прогнозе

динамики численности и структуры популяций животных и влияющих на них факторов.

Кроме того, в последнее время получили некоторую популярность вэйвлет-методы (wavelet methods), которые близки по своим принципиальным подходам к фильтрации рядов в методе главных компонент. В этих методах сначала выбирается так называемая «материнская волна», например, «мексиканская шляпа», зависящая от параметров сдвига и сжатия, а потом эта волна применяется в качестве фильтра к исходному ряду при всех возможных значениях этих параметров. Получающаяся при этом поверхность над двумерной плоскостью анализируется визуально. Литературы на русском языке по вэйвлет-методам практически нет, но их описание и программное обеспечение на английском языке доступно через Интернет (поиск по ключевому слову "wavelet").

Обработка одного временного ряда методом главных компонент

Пусть имеется последовательность $x_1, x_2, ..., x_N$ наблюдений некоторого показателя в равноотстоящие моменты времени $1, 2, ..., N$. Выберем в качестве многомерной характеристики процесса в момент времени t $(N = t > T)$ вектор $(x_t, x_{t-1}, ..., x_{t-T})$, именуемый предысторией процесса за время T. Параметр T называется лагом (запаздыванием). Сведем полученные векторы в таблицу, имеющую $N-T$ строк (объектов) и $T+1$ столбец (признак) (табл. 8.1).

Таблица 8.1

Сдвиг временного ряда на T отсчетов

X_1					
X_2	X_1				
X_3	X_2	X_1			
...	...	...	...		
X_t	X_{t-1}	X_{t-2}	...	X_1	
X_{t+1}	X_t	X_{t-1}	X_{t-2}	...	X_1
X_{t+2}	X_{t+1}	X_t	X_{t-1}	...	X_2
...	...	...	...	...	...
...	...	...	...	...	...
X_{N-1}	X_{N-2}	X_{N-3}	X_{N-4}	...	X_{N-T-1}
X_N	X_{N-1}	X_{N-2}	X_{N-3}	...	X_{N-T}
	X_N	X_{N-1}	X_{N-2}	...	X_{N-T-1}
		X_N	X_{N-1}	...	X_{N-T-2}
			X_N	...	...
				X_N	X_{N-1}
					X_N

Обработка полученной матрицы методом главных компонент приводит к появлению новой матрицы тех же размеров. Новые признаки (компоненты) являются линейными комбинациями старых

$$U_{ij} = \sum a_{ij} x_{t-i}, \quad \sum a_{ij}^2 = 1, \quad j = 0,...,T; \quad t = T+1,...,N$$

и не коррелируют между собой. Первая компонента имеет максимально возможную из всех линейных комбинаций дисперсию, вторая – максимально возможную из всех линейных комбинаций, ортогональных первой, и так далее.

Так как каждая из полученных компонент, в свою очередь, является новым временным рядом, то ее поведение можно исследовать в зависимости от любой другой компоненты, получая фазовые портреты. В последнем случае каждое состояние представляется точкой на плоскости, образованной соответствующей парой компонент, и состояния соединяются последовательно (например, сплайнами), образуя траекторию процесса в проекции на плоскость данных компонент.

Кроме того, компоненты имеет смысл использовать в качестве предикторов, так как за каждой компонентой предположительно стоит порождающая ее самостоятельная и статистически независимая от других причина (Ефимов и др., 1988).

В исследования по динамике численности животных фазовые портреты с $T=1$ впервые применил Моран (Уильямсон, 1975). Более сложный случай рассмотрен в работе (Schaffer, 1984). С помощью компьютерной графики исследовалась траектория (x_t, x_{t-i}, x_{t-2i}) заготовок шкур канадской рыси в трехмерном пространстве, где i выбиралось таким образом, чтобы выйти за пределы значимой корреляции между x_t и x_{t-i}. Многомерная траектория $(x_t, x_{t-1}, \ldots, x_{t-1})$ динамики заготовок шкурок водяной полевки и ее представление в виде фазового портрета с помощью метода главных компонент впервые рассмотрены нами в публикациях (Ефимов, Галактионов, 1982, 1983; Галактионов и др., 1987) и монографиях (Ефимов и др., 1988) и (Бобрецов и др., 2000).

Очевидно, что сфера применения метода выходит далеко за пределы динамики численности животных и он может быть применен к временным рядам любой природы. Однако история его появления достаточно запутана. Первыми публикациями, относящимися к этому методу, считаются статьи (Colebrook, 1978; Broomhead, King, 1986a, 1986b). Однако его идеи неоднократно и независимо появлялись (и появляются до сих пор) в различных областях знаний, связанных с обработкой временных рядов. Один из обзоров публикаций на эту тему можно найти в сборнике (Главные компоненты … , 1997). Имеются две монографии: элементарное введение в метод (Elsner, Tsonis, 1996) и содержащая его теоретическое обоснование книга (Golyandina et al, 2001).

Обработка нескольких временных рядов методом главных компонент

Методом главных компонент можно обрабатывать и совокупности взаимосвязанных временных рядов. В этом случае информация представляется в виде матрицы, в которой объектами являются отсчеты, например, годы, а признаками служат исследуемые временные ряды. После обработки полученной матрицы методом главных компонент большая часть информации оказывается сосредоточенной в первых компонентах. Любую из компонент можно анализировать как новый временной ряд.

Так как каждая компонента отражает существующую по какой-либо причине общность временных рядов, постоянную, временную или даже случайную,

проявляющуюся в коррелированности рядов, и компоненты не коррелируют между собой, то очень часто является осмысленным предположение, что эти причины также независимы. Если совокупность временных рядов представляет собой регистрацию одного показателя, относящегося к различным точкам или районам территории, то метод главных компонент можно использовать для районирования этой территории по каждой компоненте и, следовательно, отдельно по каждой причине, порождающей общность временных рядов (Ефимов, Галактионов, 1983; Гусев, Ефимов, 1985; Ефимов и др., 1988). Если эта совокупность объединяет группу близких по смыслу показателей, например, динамики урожайности нескольких видов культур, то с помощью главных компонент можно выявить, во-первых, общие для всех или частные для некоторых подгрупп факторы, а, во-вторых, расположить виды в соответствии с чувствительностью и направлением действия этих факторов. Правда, необходимо отметить, что метод главных компонент не предоставляет автоматической интерпретации получаемых факторов и об их смысле приходится догадываться отдельно, что в некоторых случаях представляет непростую задачу и предъявляет довольно высокие требования к квалификации интерпретатора.

Если обрабатывается транспонированная матрица, то временными рядами являются собственные векторы, а вклады признаков отражены в компонентах.

Метод гладких компонент

Еще одним способом выбора ортогональной матрицы, осуществляющей поворот к новым (не ортогональным) осям, является метод гладких компонент. Не умаляя ценности главных компонент следует заметить, что, кроме максимальной дисперсии, нас часто интересует возможность прогнозирования получаемой комбинации и в этом случае целесообразнее максимизировать не дисперсию, а автоковариацию (произведение дисперсии на коэффициент автокорреляции). Это приводит к новому методу обработки временных рядов, который мы назвали *методом гладких компонент* (Бобрецов и др., 2000).

Пусть $X = \{x_{tj}\}, \quad t = 1,...,N; \quad j = 1,...,M,$

где N – число лет наблюдений, M – число временных рядов. Пусть X уже центрирована и нормирована. Обозначим через X_1 матрицу X без первой строки, через X_N - матрицу X без последней строки. Тогда

$$r = \sum_{t=1}^{t=N-1} [\sum_{j=1}^{j=M} (a_j x_{tj}) \sum_{j=1}^{j=M} a_j x_{t+1,j})] = (X_N a)' X_1 a = a' X'_N X_1 a = a' V a,$$

где $V = X'_N X_1$.

Очевидно, что точно так же $r = (X_1 a)'(X_N a) = a' X'_1 X_N a = a' V' a$ и, следовательно, $r = a' W a$, где $W = (V + V')/2$, но матрица W уже симметрична.

Максимизируем r при условии $\sum a_j^2 = a'a = 1$.

Обычными методами дифференциального исчисления (Кульбак, 1967) получим, что вектор a удовлетворяет матричному уравнению

$$Wa = \lambda a.$$

Так как матрица W симметрична, для ее решения достаточно применить стандартный метод нахождения собственных векторов и значений. В результате получим

$$W = Q\Lambda Q',$$

где Q – ортогональная матрица собственных векторов,

Λ – диагональная матрица собственных значений.

Умножая X на Q, получим $U=XQ$ - матрицу гладких компонент. По существу, этот метод подобен вычислению канонической корреляции между матрицами X_I и X_N , но при дополнительном условии совпадения набора коэффициентов внутри каждой пары дискриминантных функций.

Еще раз напомним, что, в отличие от главных компонент, которые всегда ортогональны друг другу, гладкие компоненты не обязаны быть ортогональными, несмотря на ортогональность Q. Однако на практике корреляции между ними обычно невелики, что позволяет относиться к ним, как достаточно независимым составляющим матрицы X. Второе отличие заключается в том, что собственные значения могут быть отрицательными, если отрицательны соответствующие автоковариации.

Существует довольно глубокое и неожиданное сходство между методами главных и гладких компонент для анализа временных рядов и многомерным генетическим анализом. Одним из его приемов является многомерный анализ фенотипической ковариационной матрицы P (Thorpe, Leamy, 1983; Falconer, 1989). Более двадцати лет назад была введена матрица генетических ковариаций G (Lande, 1979). По смыслу – это коэффициенты корреляции(ковариации) между признаками родителей и их потомков. После ее введения оказалось возможным оценивать аддитивную наследуемость любой линейной комбинации признаков, в том числе, главных компонент матриц G и P (Atchley et al., 1981). Однако направленный поиск комбинированных признаков с максимальной аддитивной наследуемостью предложен и проведен только в недавнее время. Например, Ott&Rabinowitz (1999) для максимизации аддитивной наследуемости предложили разложение матрицы GP^{-1} на собственные вектора. Klingenberg&Leamy (2001) с помощью такого разложения получили линейную комбинацию промеров нижней челюсти с аддитивной наследуемостью 0.73, не совпадающую ни с одной из главных осей матриц G и P. При этом наследуемость общего размера нижней челюсти на этом же материале равна 0.42, что по порядку совпадает с оценками наследуемостей других краниометрических признаков (Leamy, 1974; Atchley et al., 1981). Таким образом, разложение соответствующих ковариационных матриц может привести к комбинированным признакам с существенно более высокими коэффициентами наследуемости, чем у исходных признаков.

В случае временных рядов матрицы R и W играют роль матриц P и G, соответственно, так состояние системы на следующий год является «потомком» по отношению к ее текущему состоянию. Аналогом наследуемости является предсказуемость или прогнозируемость, аналогом ДНК – инерционность системы. В широком смысле и ДНК и инерцию можно рассматривать как формы памяти – нечто инвариантное, наследуемое следующим поколением. Продолжая аналогию дальше, можно ставить вопрос о поиске линейных комбинаций во временных рядах с максимальной прогнозируемостью через разложение матрицы WR^{-1} и пытаться выяснить их содержательный смысл.

ЛИТЕРАТУРА

С литературой по многомерному анализу дело обстоит плохо. Есть много учебников и пособий, требующих глубоких математических знаний и не слишком доступных для биологов. На сегодня лучшим источником информации является Интернет. Однако везде есть неточности и ошибки, поэтому все надо перепроверять по другим источникам.

Рекомендуемая литература (основная):

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. –М.: Финансы и статистика, 1985. –487с.

Боровиков В.П., Боровиков И.П. STATISTICA® – Статистический анализ и обработка данных в среде Windows®. –М.: «Филинъ», 1997. –600с.

Васильева Л.А. Биологическая статистика. –Новосибирск: ИЦиГ СО РАН, 2000. –123с.

Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. –Новосибирск: Наука, 1996. –276с.

Дрейпер Н., Смит Г. Прикладной регрессионный анализ. В 2-кн. –М.: Финансы и статистика, 1987. –351с.

Дэйвисон М. Многомерное шкалирование. –М.: Финансы и статистика, 1988. –254с.

Иберла К. Факторный анализ. –М.: Статистика, 1980. –398 с.

Кендалл М., Стьюарт А. Статистические выводы и связи. –М.: Наука, 1973. –899с.

Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. –М.: Наука, 1976. –736с.

Ланкастер П. Теория матриц. –М.: Наука, 1978. –280с.

Песенко Ю.А. Принципы и методы количественного анализа в фаунистических исследованиях. –М.: Наука, 1982. –287с.

Плохинский Н.А. Биометрия. –Новосибирск: Изд-во СО АН СССР, 1961. –364с.

Родионова О.Е., Померанцев А.Л. Хемометрика: достижения и перспективы // Успехи химии, 2006. –Т.75, –С.302-317.

Уильямсон М. Анализ биологических популяций. –М.: Мир, 1975. –271с.

Principal Manifolds for Data Visualisation and Dimension Reduction (Eds. Gorban A., Kegl B., Wunsch D., Zinovyev A.). –Berlin–Heidelberg–New York: Springer, 2007. –330p.

Рекомендуемая литература (дополнительная):

Агеев М.И., Алик В.П., Марков Ю.И. Библиотека алгоритмов 51б-100б. -М.: Сов. радио, 1976. -136с. (Справочное пособие; Вып.2).

Акимов И.А., Гробов О.Ф., Пилецкая И.В., Барабанова В.В., Ястребцов А.В., Горголь В.Т., Залозная Л.М., Галактионов Ю.К., Ефимов В.М., Непомнящих В.А. Пчелиный клещ Varroa Jacobsoni. –Киев: Наукова думка, 1993. –256с.

Александров А.Д., 1987. Основания геометрии. –М: Наука. –288с.

Балеску Р. Равновесная и неравновесная статистическая механика. Т.2. –М: Мир, 1978. –478с.

Бигон М., Харпер Дж., Таунсенд К. Экология. Особи популяции и сообщества. –М.: Мир, 1989. –Т.2. –477с.

Бобрецов А.В., Бешкарев А.Б., Басов В.А., Васильев А.Г., Ефимов В.М., Кудрявцева Э.Н., Мегалинская И.З., Нейфельд Н.Д., Сокольский С.М., Теплов В.В., Теплова В.П. Закономерности полувековой динамики биоты девственной тайги Северного Предуралья. –Сыктывкар: Госкомстат республики Коми. 2000. –206с.

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. –М.: Наука. 1983. –416с.

Бурлачук Л.Ф. Психодиагностика: Учебник для вузов. 2002. // www.lbvs.kiev.ua/psydiag.

Васильев А.Г., Фалеев В.И., Галактионов Ю.К., Ковалева В.Ю., Ефимов В.М., Епифанцева Л.Ю., Поздняков А.А., Дупал Т.А., Абрамов С.А. Реализация морфологического разнообразия в природных популяциях млекопитающих. –Новосибирск: Издательство СО РАН. 2003. –232с.

Вернадский В.И. Размышления натуралиста. пространство в неживой и живой природе. –М.: Наука, 1975. –175с.

Виноградов Б.С. Процесс роста и возрастная изменчивость черепа Arvicolidae // Изв. Петроградск. обл. станции защиты растений от вредителей. 1921. –Петроград. –Т.3. –С. 71-81.

Гайдышев И. Анализ и обработка данных: специальный справочник. –СПб: Питер, 2001. – 752с.

Галактионов Ю.К. Дискретный полиморфизм по скорости роста в природной популяции водяной полевки / Научн.-техн. бюлл. СО ВАСХНИЛ, 1981. –Вып.37. –С.17-26.

Галактионов Ю.К., Ефимов В.М., Гусев В.М. Некоторые особенности анализа агрометеорологических рядов методом главных компонент. Метеорология и гидрология, 1987. №9, –С.92–97.

Галактионов Ю.К., Ефимов В.М., Пикулик М.М., Косова Л.В. Онтогенетические механизмы морфометрической адаптации остромордой лягушки *Rana arvalis* (ANURA, RANIDAE) к физико-географическим градиентам среды // Вестник зоологии, 1995. №1. –С.55–61.

Главные компоненты временных рядов: метод "Гусеница". (ред. Д.Л.Данилов, А.А.Жиглявский). –СПб: СПбГУ, 1997. –308с.

Горбань А.Н. Обучение нейронных сетей. –М.: изд. СССР-США СП "ПараGraph", 1990. –160 с.

Горбань А.Н. Функции многих переменных и нейронные сети // Сорос. образ. журн., 1998. – №12. –С.105-112.

Гусев С.М., Ефимов В.М. Районирование сельскохозяйственных культур по урожайности в Новосибирской области.//Вестник с.-х. наук, 1985. №3(342). –С.37–41.

Диаконис П., Эфрон Б. Статистические методы с интенсивным использованием ЭВМ. // В мире науки, 1983, 7. С.60–73.

Дидэ Э. Методы анализа данных. –М.: Финансы и статистика, 1985. –357с.

Дирак П.А.М. Воспоминания о необычайной эпохе. –М.: Наука, 1990. –208с.

Европейская рыжая полевка (ред. Башенина Н.В.). –М.: Наука. 1981. –351с.

Ефимов В.М., Галактионов Ю.К. Основы прогноза динамики численности водяной полевки. – Научн.-техн.бюл. //ВАСХНИЛ, Сиб.отд-ние, СибНИИЗХим. –Новосибирск, 1982. – Вып.22. –С.11–26.

Ефимов В.М., Галактионов Ю.К. О возможности прогнозирования циклических изменений численности млекопитающих // Ж. общ. биол., 1983. №3, –С.343–352.

Ефимов В.М., Галактионов Ю.К., Шушпанова Н.Ф. Анализ и прогноз временных рядов методом главных компонент. –М.: Наука, 1988. –70с.

Ефимов В.М., Ковалева В.Ю. Многомерный анализ биологических данных: учебное пособие. – Горно-Алтайск: РИО ГАГУ, 2007. –75с.

Животовский Л.А. Интеграция полигенных систем в популяциях (проблемы анализа комплекса признаков). –М.: Наука, 1984. –184 с.

Животовский Л.А. Популяционная биометрия. –М.: Наука, 1991. –271с.

Кашьяп Р.Л., Рао А.Р. Построение динамических стохастических моделей по экспериментальным данным. –М.: Наука, 1983. –383с.

Ким Дж. О., Мьюллер Ч.У., Клекка У.Р. и др. Факторный, дискриминантный и кластерный анализ. –М.: Финансы и статистика, 1989. –215с.

Ковалева В.Ю. Краниоодонтологическая изменчивость полевок // Автореф. дисс ... канд. биол. наук. –Новосибирск: ИСЭЖ СО РАН, 1999. –24с.

Колмогоров А.Н. К вопросу о пригодности найденных статистическим путем формул прогноза // Журн. геофиз., 1933. –Т.3. –С.78–82. (Переизд.: Колмогоров А.Н. Теория вероятностей и математическая статистика. –М.: Наука, 1986. –С.161–167.)

Колмогоров А.Н. Основные понятия теории вероятностей. –М–Л.: ОНТИ, 1936. (2-е изд. – М.: Наука, 1974. –122с.)

Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. –М.:

Наука, 1970. –720с.

Косова Л.В., Пикулик М.М., Ефимов В.М., Галактионов Ю.К. Внутривидовая изменчивость морфометрических признаков остромордой лягушки *Rana arvalis* (ANURA, RANIDAE) Беларуси // Зоол.журн., 1992. –Т.71, №4, –С.34–44.

Крамер Г. Математические методы статистики. –М.: Мир, 1975. –648с.

Кульбак С. Теория информации и статистика. –М.: Наука, 1967. –408с.

Любищев А.А. Проблемы формы, систематики и эволюции организмов. –М.: Наука, 1982. –280с.

Мазер К., Джинкс Дж. 1985. Биометрическая генетика. М.: Мир.–464 с.

Миронов Б.Н. История в цифрах. Математика в исторических исследованиях. –Л.: Наука, 1991. –167с.

Пуанкаре А. О науке. –М.: Наука, 1983. –560с.

Родионова О.Е. Интервальный подход к анализу больших массивов физико-химических данных // Автореф. дисс ... докт. физ.-мат. наук. М.: ИФХ РАН, 2007. –48с.

Северцов А.С. Контрбаланс векторов движущего отбора как причина эволюционного стазиса //Экология в России на рубеже XXI века. –М.: МГУ, 2000. –С.27–53.

Терентьев П.В. Истоки биометрии // Из истории биологии. Вып. 3. –М.: Наука, 1971. –С.124–134.

Уилкс С. Математическая статистика. –М.: Наука. 1967. –632 с.

Фейнман Р., Лейтон Р., Сэндс М. Фейнмановские лекции по физике. –М.: Мир, 1978. –524с.

Царапкин С.Р. Анализ дивергенции признаков между двумя географическими расами и двумя видами // Применение математических методов в биологии. –Л.: Изд-во ЛГУ. 1960. Вып. 1. –С.65–74.

Шараф М.А., Иллмэн Д.Л., Ковальски Б.Р. Хемометрика. –М.: Мир, 1987. –272с.

Шараф М.А., Иллмэн Д.Л., Ковальски Б.Р. Хемометрика. –Л., Химия, 1989. –272с.

Шварц С.С. Экологические закономерности эволюции. –М.: Наука, 1980. –278с.

Шепард Р.Н. Многомерное шкалирование и безразмерное представление различий. // Психологический журнал, 1980, Т. 1, № 4, с. 72—83.

Шепард Р. Многомерное шкалирование и неметрические представления. // Нормативные и дескриптивные модели принятия решений. –М.: Наука, 1981.

Шмальгаузен И.И. Организм как целое в индивидуальном и историческом развитии. –М.: Наука, 1982. –383с.

Atchley W.R., Rutledge J.J., Cowley D.E. Genetic components of size and shape. 2.Multivariate covariance patterns in the rat and mouse skull //Evolution, 1981. –V.35. –N6. –P.1037–1055.

Boardman A.E., Hui B.S., Wold H. The Partial Least Squares – Fix-Point Method of Estimating Interdependent Systems With Latent Variables Communication in Statistics // Theory Meth. 1981. –Vol. A10, –No. 7. –P. 613-639.

Broomhead D.S., King G.P. Extracting qualitative dynamics from experimental data // Physica D. 1986a. –Vol. 20. –P.217–236.

Broomhead D.S., King G.P. On the qualitative analysis of experimental dynamical systems // Nonlinear Phenomena and Chaos / Ed. by S. Sarkar. Bristol: Adam Hilger. 1986b. –P.113–144.

Caroll J.D., Chang J.-J. Analysis of individual differences in multidimensional scaling via N-way generalization of «Ekart-Young» decomposition // Psychometrika, 1970. Vol. 35, P. 283-319.

Carroll J.D. Spatial, non-spatial and hybrid models for scaling // Psychometrika, 1976, v. 41, p. 439––463.

Cattell J. Mc-K. Mental Test and Mesurement // Mind, 1890. V.15. P.373-381.

Colebrook J.M. Continuos plankton records – zooplankton and environment, northeast Atlantic and North Sea, 1948-1975. Oceanol. Acta. N1. 1978. –P.9–23.

Efimov VM, Kovaleva VY and Markel AL. A new approach to the study of genetic variability of complex characters // Heredity, 2005. –V.94. –P.101–107.

Efron B. Bootstrap methods: another look at the jackknife // Ann. Statist. 1979. –V.7. –P.1–26.

Efron B. The jackknife, the Bootstrap and other resampling plane. –Philadelphia.Pa: SIAM, 1982. – 92p.

Elsner J., Tsonis A. Singular Spectrum Analysis. A New Tool in Time Series Analysis. –New York: Plenum Press, 1996. –163p.

Falconer, D. S. Introduction to Quantitative Genetics, 3rd ed. – New York: Longman, 1989. –438 p.

Fisher R.A. The use of multiple measurements in taxonomic problems // Annals of Eugenics, 1936. –V.7,–P.179-188.

Galton F. Psychometric experiments // Brain: A Journal of Neurology, 1879. V. II, P.149-162.

Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. –Boca Raton: Chapman & Hall/CRC. 2001. –305 p.

Hotelling H. Analysis of a complex of statistical variables into principal components. J. Ed. Psych., 1933. 24. 417–441, 489–520.

Hotelling H. Relations between two sets of variables. Biometrika, 1936. 28. 321–377.

Klingenberg C.P., Leamy L. Quantitative genetics of geometric shape in the mouse mandible. Evolution, 55(11), 2001, pp. 2342–2352.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59–69.

Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis// Psychometrika, Vol. 29, 1964a. P. 1-27.

Kruskal J.B. Nonmetric multidimensional scaling: a numerical method// Psychometrica, Vol. 29, 1964б. P. 115-130.

Lande R. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution, 1979. 33, 402–416.

Leamy L. Heritability of osteometric traits in a random bred population of mice. J. Hered., 1974. 65: 109–120.

Moran P.A.P. The statistical analysis of the Canadian lynx cycle //Aust.J.Zool., 1953. –V.1. –P.163–173,291–298.

Ott J., Rabinowitz D. A Principal-Components Approach Based on Heritability for Combining Phenotype Information. Human Heredity, 1999. 49(2), 106–111.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling // Philosoph. Mag., 1900. –V. 50. –P.157-175.

Pearson K. On lines and planes of closest fit to systems of points in space // Philosoph. Mag., 1901. –V. 2(6). –P. 559.

Richardson MW. 1938. Multidimensional psychophysics. Psychological Bulletin, 35, 659-660.

Shepard R.M. The analysis of proximities: multidimensional scaling with an unknown distance function.—Psychometrika, 1962, v. 27, N 2-3, p. 125-139, 219-246.

Schaffer W.M. Stretching and folding in lynx fur returns: evidence for astrange attractor in nature? //Am. Nat., 1984. –V.124. –N6. –P.798–820.

Spearman, C. E. (1904a). "General intelligence" objectively determined and measured. American Journal of Psychology, 15, 201-293.

Spearman, C. E. (1904b). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Spearman C.E. The abilities of man, their nature and measurement. –New York: Macmillan, 1927. 457p.

Stevens S.S. On the theory of scales of measurement // Science. 1946. V. 103. P. 677—680. [Стивенс С.С. Математика, измерение и психофизика // Экспериментальная психология, т. 1. М.: ИЛ, 1960. С. 19-89.]

Taguchi Y.-h. and Oono Y. Relational patterns of gene expression via non-metric multidimensional scaling analysis // BIOINFORMATICS, 2005. Vol. 21, No. 6, P. 730–740.

Takens F. Dynamical Systems and Turbulence. Lecture Notes in Mathematics. –Heidelberg: Springer-Verlag, 1981. P.366–381.

2

62 norpe RS, Leamy L. Morphometric studies in inbred and hybrid House mouse (Mus sp.): Multivariate analysis of size and shape. J. Zool. Lond., 1983. 199: 421–432.

Th hurstone LL. 1927. A law of comparative judgment/ Psychological Review, 34, 273-286.

hurstone, L. L., The Vectors of Mind - Multiple-Factor Analysis for the Isolation of Primary Traits. Th Chicago: University of Chicago Press, 1935. 266 p.

Th hurstone LL. Primary mental abilities. 1938. Primary Mental Abilities: By L.L. Thurstone. Chicago: University of Chicago Press, 1938. 116 p.

Th orgerson W.S. Multidimensional scaling: I. Theory and method. Psychometrika, 1952, v. 17, N 3, p. 401-419. [Торгерсон У.С. Многомерное шкалирование: теория и метод // To Статистическое измерение качественных характеристик. – М.: Статистика, 1972. – С. 95-118.]

/elch B.L. The significance of the difference between two means when the population variances are unequal // Biometrika, 1938. Vol. XXIX, Parts III and IV. P. 350–-362.

W old H. Partial least squares / Encyclopedia of statistical sciences (S.Kotz and N.L.Johnson, eds.). – New York: Wiley, 1985. –Vol. 6. –P. 581-591.

W arapkin S.R. Zur Phanoanalyse von geographischen Rassen und Arten. Arch. Naturgesch. N.F. 1934. Bd. 3. Z. 161–186.

Z

ЗАДАНИЯ ДЛЯ ПРАКТИЧЕСКИХ РАБОТ
и методические указания по их выполнению, контрольные вопросы и варианты ответов для студентов биологических специальностей

ВВЕДЕНИЕ

Программа дисциплины "МНОГОМЕРНЫЙ АНАЛИЗ БИОЛОГИЧЕСКИХ ДАННЫХ" предусматривает изучение многомерных методов исследования массовых биологических процессов и явлений; их математического аппарата. В курсе излагаются основные понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации, свертки и обработки многомерных статистических данных с целью их удобного представления, интерпретации, получения научных и практических выводов. Курс нацелен на оснащение студентов знаниями и навыками в области основ выявления и биологической интерпретации многомерных данных, их прикладного статистического анализа, построения, идентификации и верификации статистических моделей анализируемых явлений, компьютерной реализации излагаемых приемов и методов.

Задачи учебного курса

В результате изучения дисциплины студенты должны знать основные методы многомерного анализа данных: метод главных компонент, факторный анализ, дискриминантный анализ, регрессионные методы, многомерное шкалирование, нейронные сети. Должны иметь представление об операциях над матрицами и об их соответствии геометрическим преобразованиям в многомерном пространстве.

Требования к уровню освоения курса

Студенты должны уметь использовать стандартные пакеты статистических программ при построении интегральных показателей и отборе наиболее информативных переменных и снижении размерностей анализируемых моделей. Должны уметь применять как линейные, так и нелинейные методы анализа взаимного расположения объектов в многомерном пространстве и интерпретировать получаемые результаты с биологической точки зрения.

Методические указания к выполнению заданий

Предполагается, что у каждого студента имеется собственная рабочая директория UserDir, в которой хранятся результаты всех расчетов. Вычисления проводятся с помощью пакетов Statistica и Excel. Учебными данными служат файлы IrisDat.Sta и Sunspots.Sta, имеющиеся в пакете Statistica (Program Files\StatSoft\Examples\Datasets). Обработка студентами собственных данных всячески приветствуется. Для нейросетевой обработки данных используется демо-версия программы Neural Network Wizard (NNW) BaseGroup Labs, для обработки временных рядов – свободно распространяемая версия программы «Гусеница» СПбГУ.

ЗАДАНИЕ №1

Построение графиков. Работа с признаками

1. Запустить программу Statistica.

2. Открыть файл IRISDAT.STA (File\Open\Datasets\IrisDat.Sta)

В файле находятся промеры длины и ширины чашелистиков и лепестков трех различных видов ирисов (Fisher, 1936).

3. Сохранить как файл Excel (File\Save As..\\UserDir)

(Тип файла: Excel Workbook; Сохранить)

(Опция: *Put variable names in first row – Yes;* Опция: *Use text labels – Yes*). OK.

4. Построить категоризованный график по первым двум параметрам:

(Graphs\Categorized Graphs\Scatterplots\)

Опция: Overlaid. Нажать: Variables.

Выбрать в столбце Scatterplot X: 1-SEPALLEN;

Выбрать в столбце Scatterplot Y: 2-SEPALWID;

Выбрать в столбце X-Category: 5-IRISTYPE.

OK; OK.

5. Сохранить рисунок как файл Statistica с расширением *.stg:

(WorkBook\Extract as stand-alone window\Copy;

File\Save As..\\UserDir\Iris1.stg)

6. Сохранить рисунок как файл Jpeg с расширением *.jpg:

(File\Save As..\\UserDir\Iris**1**. jpg)

7. Построить категоризованный график по следующим двум параметрам:

(внизу: 2D Categorized Scatt..) Опция: Overlaid. Нажать: Variables.

Выбрать в столбце Scatterplot X: **3**-PETALLEN;

Выбрать в столбце Scatterplot Y: **4**-PETALWID;

Выбрать в столбце X-Category: 5-IRISTYPE.

OK; OK.

8. Сохранить рисунок как файл с именем Iris2 с расширениями *.stg и *.jpg (аналогично пунктам 5 и 6).

9. Закрыть все рисунки.

10. Построить категоризованный график по трем параметрам:

(Graphs\Categorized Graphs\ 3D XYZ Plots\)

Опция: Graph type – Scatterplot;

Нажать: Codes – none. Выбрать IRISTYPE, OK; All, OK.

Нажать: Variables.

Выбрать в столбце X: 2-SEPALWID;

Выбрать в столбце Y: 3-PETALLEN;

Выбрать в столбце Z: 4-PETALWID;

Выбрать в столбце Category: 5-IRISTYPE. OK. . OK.
По открывшемуся графику щелкнуть правой кнопкой мыши. Выбрать (Graphs Properties\Categorization) Опции: Categories layout-Overlaid) OK.

11. Сохранить рисунок как файл с именем Iris3 с расширениями *.stg и *.jpg (аналогично пунктам 5 и 6).
12. Щелкнуть по графику правой кнопкой мыши. Выбрать (Graphs Properties\Plot: General) Опция: Spike line – No) OK.
Щелкнуть по графику правой кнопкой мыши. Выбрать (Graphs Properties\Point of View). Покрутить график вручную.
Нажать: Analytic exploratory spin options.
OK.
13. Закрыть все рисунки. Не сохранять рабочую область.
14. Запустить программу Excel. **(координаты ячеек набирать латиницей)**
15. Открыть файл IRISDAT.xls (Файл\Открыть\\ UserDir \IrisDat.xls).
16. Упорядочить данные по видам
Выделить массив A2:E151. (Данные\Сортировка\Сортировать по IRISTYPE\OK)
17. Выделить ячейки A2:C158 (Формат\Ячейки\ЧисловойФормат-Числовой \Число десятичных знаков-1\OK)
18. Для совместимости с программой Statistica сохранить как файл Excel 4.0. (Файл\Сохранить как..\\UserDir \IrisDat1.xls).
19. Развернуть номинальный признак IRISTYPE в набор бинарных признаков.
Поместить в ячейку F1 текст SETOSA.
Поместить в ячейку G1 текст VERSICOL.
Поместить в ячейку H1 текст VIRGINIC.
Заполнить массив F2:H151 нулями.
Заполнить массив F2:F51 единицами
(напротив значений SETOSA в признаке IRISTYPE).
Заполнить массив G52:G101 единицами
(напротив значений VERSICOL в признаке IRISTYPE).
Заполнить массив H102:H151 единицами
(напротив значений VIRGINIC в признаке IRISTYPE).
Сохранить файл IrisDat1.xls.
20. Вычислить для каждого вида средние и среднеквадратичные отклонения по каждому признаку:
Разместить в ячейках A153-158 формулы:
=СРЗНАЧ(A2:A51);
=СТАНДОТКЛОН(A2:A51);
=СРЗНАЧ(A52:A101);
=СТАНДОТКЛОН(A52:A101);

=СРЗНАЧ(A102:A151);

=СТАНДОТКЛОН(A102:A151);

Скопировать ячейки A153-158

в ячейки B153-158, C153-158, D153-158, F153-158, G153-158, H153-158.

(Копирование: выделить ячейки откуда, Ctrl+Insert,

выделить ячейки куда через Shift+Insert, Enter)

21. Центрировать признаки по каждому виду ирисов отдельно:

Скопировать ячейки A1-D1 в ячейки I1-L1. Отредактировать новые названия (SEPALLEN заменить на SEPALLEN1 и т.д.)

Разместить в ячейке I2 формулу: =A2-A$153

Скопировать ячейку I2 в ячейки I2- L51.

Разместить в ячейке I52 формулу: =A52-A$155

Скопировать ячейку I52 в ячейки I52- L101.

Разместить в ячейке I102 формулу: =A102-A$157

Скопировать ячейку I102 в ячейки I102- L151.

Скопировать ячейки A153-158 в ячейки I153-158, J153-158, K153-158, L153-158.

Сохранить файл IrisDat1.xls.

22. Выделить массив формул I2:L151. Копировать в буфер Windows (Правка\Копировать). Вставить на те же места как значения (Правка\Специальная вставка\Значения\OK).

23. Удалить строки A153-158: Выделить ячейки A153-158. (Правка\Удалить..\строку)\OK.

Сохранить как файл Excel4 с именем IrisDat2.xls (Файл\Сохранить как..\\UserDir \IrisDat2.xls).

24. Закрыть Excel. Открыть в программе Statistica файл IrisDat2.xls. Тип файла: Excel. (File\Open\\UserDir \ IrisDat2.xls)

Нажать: Import selected sheet to a SpreedSheet.

Опция: Get variable names from first row – Yes. OK.

23. Сохранить IrisDat2 как файл Statistica с расширением *.sta.

24. Построить категоризованный график по трем параметрам: SEPALWID1, PETALLEN1, PETALWID1

(п.**10-13**).

ЗАДАНИЕ №2

Главные компоненты, факторный анализ.

1. Запустить программу Statistica.

2. Открыть файл IRISDAT2.STA (File\Open**UserDir**\IrisDat2.Sta)

Признаки SEPALLEN1, SEPALWID1, PETALLEN1, PETALWID1 отражают объединенную внутривыборочную изменчивость трех различных видов ирисов по длине и ширине чашелистиков и лепестков (межвыборочная устранена центрированием).

3-18. ***Вычислить главные компоненты для объединенной внутривыборочной матрицы ирисов***

3. (Statistics\Multivariate Exploratory Techniques\Principal Components & Classification Analysis\) (Вошли в стартовую панель главных компонент)

4. Кнопка:Variables

В столбце Variables for Analysis выбрать:

9- SEPALLEN1;
10-SEPALWID1;
11-PETALLEN1;
12-PETALWID1;

OK. (Вернулись в стартовую панель главных компонент). OK.

5. (Вошли в панель результатов. Корешок Variables)

6. Сохранить матрицу собственных значений (Eigenvalues)(WorkBook\Extract as stand-alone window\Copy; File\Save As..\\UserDir\ Eigenvalues.sta)Сохранить.
Закрыть два последних окна (крестики вверху). Опция: Save changes.. – Нет.

7. Сохранить матрицу собственных векторов (Eigenvectors)(WorkBook\Extract as stand-alone window\Copy; File\Save As..\\UserDir\ Eigenvectors.sta)Сохранить.
Закрыть два последних окна (крестики вверху). Опция: Save changes.. – Нет.

8. Щелкнуть внизу кнопку Principal components (Вернулись в панель результатов главных компонент. Корешок Variables)

9. Сохранить матрицу корреляций признаков с компонентами (Factor & variable correlations)(WorkBook\Extract as stand-alone window\Copy; File\Save As..\\UserDir\ Factor-variable.sta)Сохранить.
Закрыть два последних окна. Опция: Save changes.. – Нет.

10. Щелкнуть внизу кнопку Principal components (Вернулись в панель результатов главных компонент. Корешок Variables)

11. Посмотреть график собственных векторов (Plot of factor coordinates, 2D) Опции: Factor1, Factor2. OK

Закрыть последнее окно (с графиком). Опция: Save changes.. – Нет.

12. Щелкнуть внизу кнопку Principal components

(Вернулись в панель результатов главных компонент. Корешок Cases)

13. Сохранить таблицу главных компонент

Опция: Factor coordinates –Yes;

Save case statistics. Select all; OK

File\Save As..\\UserDir\ Factors.sta)Сохранить.

Закрыть и вернуться в панель результатов главных компонент. Корешок Cases.

14. Посмотреть график главных компонент (без сохранения)

(Plot of factor coordinates, 2D) Опции: Factor1, Factor2. OK

Закрыть и вернуться в панель результатов главных компонент.

15. Закрыть все окна. Вернуться в стартовую панель Statistica.

16. Открыть файл Factors.STA (File\Open\..\\UserDir\ Factors.sta)

17. Построить категоризованный график по первым двум компонентам (Factor1 – Factor2, без сохранения).

18. Построить категоризованный график по первым трем компонентам (Factor1 – Factor3, без сохранения).

19-24. ***Выполнить факторный анализ для объединенной внутривыборочной матрицы ирисов***

19. (Statistics\Multivariate Exploratory Techniques\ Factor Analysis\)

(Вошли в стартовую панель факторного анализа)

20. Кнопка:Variables\

В окне Select Variables for Factor Analysis выбрать

9- SEPALLEN1;

10-SEPALWID1;

11-PETALLEN1;

12-PETALWID1;

OK. (Вернулись в стартовую панель факторного анализа)

OK. (Вошли в панель выбора метода.

21. Корешок Advanced)

Опция: Extraction metod: Principal components – Yes.

Опция: Max no. of factors – 4.

Опция: Mini. eigenvalue – 0.000.

OK. (Вошли в панель результатов. Корешок Loadings)

22. Щелкнуть кнопку Summary: Factor loadings.

Убедиться, что таблица факторных нагрузок (при опции Unrotated) тождественна матрице корреляций признаков с главными компонентами (File\Open\Factor-variable.sta\Открыть.

Закрыть два последних окна и вернуться в панель результатов факторного

анализа (кнопка внизу). Корешок Loadings.

23. В окне Factor rotation выбрать опцию Quartimax raw.
Щелкнуть кнопку Summary: Factor loadings.
Проанализировать изменения в таблице факторных нагрузок.
Закрыть окно и вернуться в панель результатов факторного анализа (кнопка внизу).

24. Корешок Scores. Кнопка Save factor scores.
Кнопка Select all. OK.
Сохранить файл под именем IrisDat3.sta.

25. Закрыть все окна и выйти из программы Statistica.

ЗАДАНИЕ №3

Дискриминантный анализ

1. Запустить программу Statistica.

2. Открыть файл IRISDAT3.STA (File\Open\ **UserDir** \IrisDat3.Sta)

3-13. ***Провести дискриминантный анализ трех видов ирисов.***
(Statistics\Multivariate Exploratory Techniques\Discriminant Analysis\) (Вошли в стартовую панель дискриминантного анализа)

3. Кнопка:Variables (Вошли в панель).
В окошке Grouping Variable выбрать: IRISTYPE
В окошке Independent Variable list выбрать:
1- SEPALLEN;
2-SEPALWID;
3-PETALLEN;
4-PETALWID;
OK. (Вернулись в стартовую панель дискриминантного анализа)

4. Кнопка: Codes for grouping variable. (Вошли в панель).
All; OK. (Вернулись в стартовую панель дискриминантного анализа).

5. OK. (Вошли в панель результатов).

6. Корешок: Advanced.
Кнопка: Perform canonical analysis (Вошли в панель канонического анализа).

7. Корешок: Canonical scores.
Кнопка: Scatterplot of canonical scores (Вошли в рисунок).

8. Сохранить рисунок как файл Statistica с расширением *.stg:
(WorkBook\Extract as stand-alone window\Copy;
File\Save As..\\UserDir\IrisRoots.stg)

9. Сохранить рисунок как файл Jpeg с расширением *.jpg:
(File\Save As..\\UserDir\IrisRoots. jpg)
Закрыть два последних окна (крестики вверху). Опция: Save changes.. – Нет.

10. Щелкнуть внизу кнопку: Canonical analysis
(Вернулись в панель результатов канонического анализа)

11. Кнопка: Save canonical scores (Вошли в панель).
Select All; OK. (Открылась таблица исходных данных, дополненная дискриминантными осями).

12. Сохранить файл под именем IrisDat4.sta.
File\Save As..\\UserDir\IrisDat4.sta)Сохранить.

13. Сохранить таблицу как файл Excel с расширением *.xls:
File\Save As..\\UserDir\ IrisDat4.xls)Сохранить.

14. Закрыть программу Statistica.

ЗАДАНИЕ №4

Множественная регрессия

1. Запустить программу Statistica.

2. Открыть файл IRISDAT4.STA (File\Open\ **UserDir** \IrisDat4.Sta)

3-10. ***Вычислить множественную регрессию на исходные признаки***
(Statistics\Multiple Regression \) (Вошли в стартовую панель множественной регрессии).

3. Кнопка:Variables (Вошли в панель).
В окошке Dependent Variable выбрать: ROOT_1
В окошке Independent Variable list выбрать:
1- SEPALLEN;
2-SEPALWID;
3-PETALLEN;
4-PETALWID;
OK. (Вернулись в стартовую панель множественной регрессии). OK.
Появится предупреждение об исчерпании дисперсии. OK.
Появится предупреждение о невозможности обращения матрицы. OK.

4. Корешок: Advanced. Опция Ridge regression – Yes. OK.

5. Корешок: Quick. Кнопка: Summary. Regression results.

6. Сохранить таблицу результатов как файл Statistica с расширением *.sta:
(WorkBook\Extract as stand-alone window\Copy;
File\Save As..\\UserDir\Ridge.sta)
Закрыть последнее окно (крестик вверху). Опция: Save changes.. – Нет.

7. Кнопка внизу: Multiple regression.
Корешок: Residuals/Assumption/prediction.
Кнопка: Perform residual analysis.
Корешок: Scatterplots.
Кнопка: Predicted vs Observed.

8. Сохранить рисунок как файл Statistica с расширением *.stg:
(WorkBook\Extract as stand-alone window\Copy;
File\Save As..\\UserDir\Predicted.stg)

9. Сохранить рисунок как файл Jpeg с расширением *.jpg:
(File\Save As..\\UserDir\ Predicted. jpg)

10. Закрыть все окна без сохранения изменений.
Открыть файл IRISDAT4.STA

11-25. ***Вычислить множественную регрессию на главные компонент.***

11. (Statistics\Multivariate Exploratory Techniques\Principal Components & Classification Analysis\) (Вошли в стартовую панель главных компонент)

12. Кнопка:Variables\В окне: Variables for Analysis выбрать
1- SEPALLEN;
2-SEPALWID;
3-PETALLEN;
4-PETALWID;
OK. (Вернулись в стартовую панель главных компонент)
OK. (Вошли в панель результатов).

13. Сохранить таблицу главных компонент
(Save case statistics; Опция: Factor coordinates-Yes)
Select all; OK
File\Save As..\\UserDir\IrisDat5.sta)Сохранить.

14. Закрыть все окна без сохранения изменений.
Открыть файл IRISDAT5.STA

15. Вычислить корреляционную матрицу.

16. Statistics\Basic statistic.Tables\Correlation matrices\OK\Summary\
В окне First variable list выбрать Select all. OK.
Проанализировать полученную матрицу. Обратить внимание на корреляции между ROOT_1 и Factor1-Factor4 (последними).

17. Закрыть все окна без сохранения изменений.
Открыть файл IRISDAT5.STA

18. Вычислить множественную регрессию на главные компоненты.
(Statistics\Multiple Regression \) (Вошли в панель множественной регрессии).

19. Кнопка:Variables (Вошли в панель).
В окошке Dependent Variable выбрать: ROOT_1
В окошке Independent Variable list зыбрать:
23 - Factor1;
24 - Factor2;
25 - Factor3;
OK.OK.

20. Корешок: Quick. Кнопка: Summary. Regression results.

21. Сохранить таблицу результатов как файл Statistica с расширением *.sta:
(WorkBook\Extract as stand-alone window\Copy;
File\Save As..\\UserDir\Ridge2.sta)
Закрыть последнее окно (крестик вверху). Опция: Save changes.. – Нет.

22. Кнопка внизу: Multiple regression.
Корешок: Residuals/Assumption/prediction.
Кнопка: Perform residual analysis.
Корешок: Scatterplots. Кнопка: Predicted vs Observed.

23. Сохранить рисунок как файл Statistica с расширением *.stg:
(WorkBook\Extract as stand-alone window\Copy;
File\Save As..\\UserDir\Predicted2.stg)

24. Сохранить рисунок как файл Jpeg с расширением *.jpg:
(File\Save As..\\UserDir\ Predicted2. jpg)

25. Закрыть все окна без сохранения изменений
и выйти из программы Statistica.

ЗАДАНИЕ №5

Нейронные сети

1. Запустить программу Statistica.

2. Открыть файл IRISDAT5.STA (File\Open\ **UserDir** \IrisDat5.Sta).
Сохранить как файл Excel с именем IrisDat5.xls .
Опция: Put variable names in first row – Yes. OK.

3. Закрыть программу Statistica.

4. Открыть в Excel файл IrisDat5.xls.

5. Удалить столбцы I..Z (SEPALLEN1 … Factor4).
Удалить столбец E (IRISTYPE).

6. Сохранить с именем IrisDat6.csv (разделители - запятые).
Опция: Сохранить книгу в этом формате – Да.
Закрыть Excel. Опция: Сохранять изменения – Нет.

7. Открыть файл IrisDat6.csv в текстовом редакторе (например, WordPad).
Заменить все символы ";" (точка с запятой) на пробелы. Сохранить файл с именем IrisDat6.txt. Закрыть редактор.

8. Открыть программу NNW\bin\wizard.exe

9. Кнопка: Обзор. Открыть файл IrisDat6.txt. Кнопка: Далее>>.
(*Имена полей должны содержать только буквы и цифры. Подчерки и пробелы не допускаются. Значения полей должны быть только числовыми.*)

10. В списке доступных полей SEPALLEN..PETALWID пометить как входные, SETOSA..VIRGINIC – как целевые. Кнопка: Далее>>.

11. Число нейронов задать 1. Кнопка: Далее>>.

12. Опция: Прошло –Yes. Число эпох оставить 10000. Кнопка: Далее>>.

13. Панель: Конфигурация нейросистемы. Кнопка: Далее>>.

14. Кнопка: Пуск обучения.

15. Если результаты обучения заметно отличаются от 100%, кнопками «Назад» вернуться и увеличить число нейронов (в крайнем случае – слоев). Повторить обучение.

16. Кнопка: Далее>>.

Задать набор входных параметров (5 3.3 1.4 0.2). Кнопка: Расчет.

Задать набор входных параметров (6.5 2.8 4.6 1.5). Кнопка: Расчет.

Задать набор входных параметров (6.4 2.8 5.6 2.2). Кнопка: Расчет.

Сохранить как файл NeuralWizard с именем IrisNeuro.

17. Кнопка Отмена. Выйти из NeuralWizard.

18. Открыть программу NNW\bin\wizard.exe

19. Кнопка: Обзор. Открыть файл IrisNeuro (файлы Neural Network Wizard). Кнопка: Далее>>.

20. Задать набор входных параметров (5 3.3 1.4 0.2). Кнопка: Расчет.

21. Кнопка Отмена. Выйти из NeuralWizard.

ЗАДАНИЕ №6

Многомерное шкалирование

1. Набрать в Excel таблицу. Сохранить как файл Excel4 с именем Chelomin1.xls. Закрыть Excel.

div	R	agr	sem	Arg	pt	Fv	s	f	ur	sp	Species
R	0	52	66	64	66	60	63	67	57	64	R.norvegicus
agr	52	0	63	50	55	48	56	54	45	45	A.agrarius
sem	66	63	0	51	57	57	55	54	51	64	A.semotus
arg	64	50	51	0	48	44	54	56	50	49	A.argenteus
pt	66	55	57	48	0	33	36	25	33	54	A.ponticus
fv	60	48	57	44	33	0	37	42	40	52	A.fulvipectus
s	63	56	55	54	36	37	0	39	41	61	A.sylvaticus
f	67	54	54	56	25	42	39	0	27	54	A.flavicollis
ur	57	45	51	50	33	40	41	27	0	41	A.uralensis
sp	64	45	64	49	54	52	61	54	41	0	A.speciosis

В файле содержатся данные числа замен 402 пн участка гена цитохрома В лесных и полевых мышей рода Apodemus (Челомина и др., Генетика, 1998. Т.34, №5. С.650-661.)

2. Запустить программу Statistica.
Открыть файл Chelomin1.xls (File\Open\UserDir\ Chelomin1.xls) как файл Excel.
Кнопка: Import selected sheet to a Spreadsheet.
Опция: Get case names from first column – Yes.
Опция: Get variable names from first row – Yes.
OK.

3. Сохранить как файл Statistica с именем Chelomin1.sta (File\Save as..\ Chelomin1.sta).
(*Шаги **4-8** делаются только для получения правильного матричного формата Statistica.)*

4. (Statistic\Multivariate Exploratory Techniques\Cluster Analisis).
Выбрать: Joining (tree clustering). OK.
Кнопка: Variables. Выбрать все, кроме 11-Species.OK.
OK. Кнопка: Matrix.

5. Открыть в Excel файл Chelomin1.xls (не выходя из Statistica).
Выделить клетки B2:K11. Установить для них числовой формат (Формат\Ячейки\Числовой.) Число десятичных знаков – 0. OK.
Скопировать массив в буфер Windows (Правка\Копировать).

6. Перейти в окно Statistica. Вставить (курсор в левый верхний угол, Shift+Insert).

7. Сохранить как матричный файл Statistica с именем Chelomin1.smx (File\Save as..\ Chelomin1.smx). Закрыть Excel.

8. Закрыть файл Chelomin1.sta. Ответить – Да.

9. Открыть панель многомерного шкалирования (Statistic\Multivariate Exploratory Techniques\Multidimensional Scaling).
Кнопка: Variables. Выбрать все (Select all).OK.
Number of dimensions – 2. OK.OK.
Graf final configuration. OK.
Сохранить рисунок в форматах stg и jpg с именем Chelomin1. Закрыть рисунок.
Кнопка внизу: Results. Корешок: Review&save.
Кнопка Save final configuration. OK.

10. Сохранить как файл Statistica с именем Chelomin2.sta (File\Save as..\ Chelomin2.sta).

11. Закрыть Statistica.

ЗАДАНИЕ №7

Анализ и прогноз временных рядов

1. Запустить программу Statistica.
2. Открыть файл SUNSPOT.STA (File\Open\Datasets\SunSpot.Sta)
   В файле содержится динамика количества солнечных пятен с 1749 по 1924 год.
3. Добавить новый столбец. Щелкнуть по столбцу SPOTS. Кнопка View\Add variables (Вошли в панель).
   В окошке Name задать: Years.
   В окошке Display format выбрать: Number. OK.
4. Щелкнуть правой кнопкой мыши по столбцу с годами. Выбрать Case Names Manager. Опция: To – Yes. В окошке Variables щелкнуть левой кнопкой два раза. Выбрать Years. OK.OK.
5. Сохранить таблицу в директории пользователя как файл Statistica с расширением *.sta: File\Save As..\\UserDir\ SunSpot.sta)Сохранить.
   Сохранить как файл Excel. Опция: Put variable name in first row – Yes. OK.
6. Построить график по первым двум параметрам:
   (Graphs\Scatterplots\) Кнопка Variables:
   В окошке над X выбрать 1-Years.
   В окошке над Y выбрать 2-SPOTS.
   OK; OK.
   Щелкнуть правой кнопкой мыши по любому кружочку.
   Выбрать Properties. Опция Line –Yes. OK.
7. Сохранить рисунок как файл Statistica с расширением *.stg:
   (WorkBook\Extract as stand-alone window\Copy;
   File\Save As..\\UserDir\ SunSpot.stg)
8. Сохранить рисунок как файл Jpeg с расширением *.jpg:
   (File\Save As..\\UserDir\ SunSpot. jpg)
9. Закрыть все рисунки. (Save changes.. – Нет.)
   Закрыть все панели, кроме таблицы SunSpot.
10. ***Спектральный анализ***. Statistics\ Advanced_Linear/Nonlinear_Models \Time_Series/Forecasting. (Вошли в панель спектрального анализа). Выбрать SPOTS.
    Кнопка Spectral (Fourier) analysis. (Вошли в панель). Кнопка Single series Fourier analysis. (Вошли в панель). В рамке Plot by задать: Period-Yes. Кнопка Spectral density.
11. Сохранить рисунок как файл Statistica с расширением *.stg:
    (WorkBook\Extract as stand-alone window\Copy;
    File\Save As..\\UserDir\ Spectral.stg)
12. Сохранить рисунок как файл Jpeg с расширением *.jpg:

(File\Save As..\\UserDir\ Spectral. jpg)

13. Закрыть все рисунки. (Save changes.. – Нет.)

Закрыть все панели, кроме таблицы SunSpot.

14. ***Авторегрессия.*** Statistics\ Advanced_Linear/Nonlinear_Models \Time_Series/Forecasting. Щелкнуть SPOTS.

Кнопка ARIMA&autocorrelation functions. (Вошли в панель Single series ARIMA).

Estimate constant – Yes. p-Autoregressive – 1. q-Moving_aver. – 1.

OK. (Вошли в панель).

Кнопка: Advanced.

Number of cases – 12. Start at case – 165. Append forecast … on EXIT – No.

Кнопка Plot series & forecast. (Рисунок).

Закрыть рисунок. Cancel. (Вошли в панель Single series ARIMA).

Самостоятельно подобрать параметры p-Autoregressive и q-Moving_aver. с наилучшим прогнозом. Сохранить рисунок с наилучшим прогнозом под именем Forecasts и расширениями *.stg и *.jpg.

15. Закрыть все панели.

Закрыть программу Statistica.

16. Запустить Excel. Открыть файл SunSpot.xls (Файл\Открыть\\ UserDir\ SunSpot.xls.

17. Сохранить как текстовый файл (с разделителями табуляции) с расширением *.dat (Файл\Сохранить как..\\UserDir \ SunSpot.dat). На запрос – Нет. Закрыть Excel.

Сохранить изменения – Нет.

18. Запустить программу \GUS\Caterplr (Гусеница).

19. Открыть файл SunSpot.dat (Файл\Открыть\\ UserDir \ SunSpot.dat)OK.

Панель: Формат исходных данных.

Опции: Структурированный – Yes. Названия переменных – Да.

Разделитель – точка. Кодировка – OEM.OK.

Выберите переменную – SPOTS. OK.

Кнопка: Опции\Графика\Максимальное кол-во точек на экран – 200. OK.

(В случае неполадок со шрифтами посмотреть файл readme.txt в директорииGUS)

20. Кнопка: Разложение временного ряда (картинка). (Вошли в панель). Длина гусеницы – 11. Нормировать – Нет. Центрировать – Да. OK.

21. Кнопка: Восстановление временного ряда (картинка). (Вошли в панель).

Кнопка: >>.. Выделить 4-ю компоненту.

Кнопка: <. Нажимать, пока в правом окне не останутся первые 3 компоненты. OK.

22. Сохранить (Файл\Сохранить_данные\Кодировка-OEM\OK\ SunSpot1.dat\OK).

23. Кнопка: ◄. Опция: Графика 1D.

Максимальное количество точек – 200.OK. График 2D.

24. Закрыть программу «Гусеница».

25. Запустить Excel. Открыть файл SunSpot.xls (Файл\Открыть\\ UserDir\ SunSpot.xls.

26. В ячейки C1-M1 занести имена новых переменных S1-S11.

27. Выделить блок B2-B177. Скопировать в ячейки C3-C178, D4-D179, E5-E180, F6-F181, G7-G182, H8-H183, I9-I184, J10-J185, K11-K186, L12-L187, M13-M188.

28. Удалить неполные строки (сначала 178-188, потом 2-12).

29. Выделить ячейки A2-N166. Формат\Ячейки\ЧисловойФормат-Числовой \Число десятичных знаков-0\OK)

30. Сохранить как файл Excel 4.0 (Файл\Сохранить как..\\UserDir \ SunSpot1.xls).

31. Удалить строки 157-166.

32. Сохранить как текстовый файл с разделителями табуляции (Файл\Сохранить как..\\UserDir \ SunSpot2.txt). Запрос – Да.

33.. Закрыть Excel. Сохранить изменения – Нет.

34. Открыть программу NNW\bin\wizard.exe

35. Кнопка: Обзор. Открыть файл SunSpot2.txt. Кнопка: Далее>>.
(*Имена полей должны содержать только буквы и цифры.*
Подчерки и пробелы не допускаются.
Значения полей должны быть только числовыми.)

36. В списке доступных полей S1..S11 пометить как входные, SPOTS – как целевое, YEARS – не использовать. Кнопка: Далее>>.

37. Число нейронов задать 1. Кнопка: Далее>>.

38. Опция: Прошло –Yes. Число эпох оставить 10000. Кнопка: Далее>>.

39. Панель: Конфигурация нейросистемы. Кнопка: Далее>>.

40. Кнопка: Пуск обучения.

41. Если результаты обучения заметно отличаются от 100%, кнопками «Назад» вернуться и увеличить число нейронов (в крайнем случае – слоев). Повторить обучение.

42. Задать число нейронов 3. Повторить обучение. Кнопка: Далее>>.

43. Открыть в Excel файл SunSpot1.xls (не закрывая NNW).
Вставить пустой столбец в колонку C. (Курсор в ячейку C1. Вставка\Столбцы).
Занести в ячейку C1 текст: PREDICT.

44. Перейти в окно NNW (не закрывая Excel).
Задать (вручную) набор входных параметров S1..S11 из строки с годом 1915, т.е. 96 14 36 57 186 439 485 620 538 635 420.
Кнопка: Расчет. Занести полученное значение в Excel в ячейку C157.
Задать в окне NNW набор входных параметров из строки с годом 1916.
Кнопка: Расчет. Занести полученное значение в Excel в ячейку C158.
Задать в окне NNW набор входных параметров из строки с годом 1917.
Кнопка: Расчет. Занести полученное значение в Excel в ячейку C159.

... (и так далее...)

Задать в окне NNW набор входных параметров из строки с годом 1924.

Кнопка: Расчет. Занести полученное значение в Excel в ячейку C166.

45. Сохранить обученную нейросеть как файл NeuralWizard с именем SunSpotNeuro.

Кнопка Отмена. Выйти из NeuralWizard.

46. Сохранить SunSpot1.xls как файл Excel4.

Закрыть Excel.

47. Запустить программу Statistica.

48. Открыть файл SunSpot1.xls (File\Open\UserDir\ SunSpot1.xls)

Кнопка: Input selected spreadsheet as Spreadsheet.

Опция: Put variable names from first row – Yes.

Сохранить как файл Statistica.(File\Save As..\\UserDir\ SunSpot1.sta)Сохранить.

49. Построить график. (Graphs\Scatterplots).

Опция: Graph type – Multiple. Опция: Linear fit – No.

Кнопка: Variables.

В окне над X выбрать Years. В окне над Y – SPOTS, PREDICT. OK.

Кнопка: Select cases. Опция: Enable Selection Conditions – Yes.

Опция: Specific, selected by.

В окне "or case number" набрать 130:166. OK.OK.

Щелкнуть правой кнопкой по красному объекту.

Кнопка: Properties.. Опция: Line – Yes. OK.

Щелкнуть правой кнопкой по синему объекту.

Кнопка: Properties.. Опция: Line – Yes. OK.

Сохранить рисунок с именем Forecast2.jpg.

Выйти из программы Statistica.

Контрольные вопросы и варианты ответов к курсу «Многомерный анализ биологических данных»

1. В чем суть геометрического подхода в биологических исследованиях?

1. В представлении объектов фигурами в многомерном евклидовом пространстве.
2. В представлении объектов линиями в многомерном евклидовом пространстве.
3. В представлении признаков векторами в многомерном евклидовом пространстве.
4. В представлении объектов точками в многомерном евклидовом пространстве.

2. Что служит моделью биологического объекта при геометрическом подходе?

1. Линия. 2. Фигура. 3. Точка.4. Отрезок.

3. Что служит моделью различий между объектами при геометрическом подходе?

1. Угол между векторами.
2. Несовпадение геометрических фигур при наложении.
3. Евклидово расстояние между объектами.
4. Кратчайшее расстояние между линиями.

4. Чем объекты отличаются от признаков?

1. Признаки однородны, объекты необязательно.
2. Объекты однородны, признаки необязательно.
3. Объекты являются строками в матрице данных, а признаки – столбцами.
4. Объекты можно центрировать, а признаки – нет.

5. Какие типы признаков НЕ используются в многомерном анализе данных?

1. Биномиальные. 2. Количественные. 3. Качественные. 4. Порядковые.

6. Что такое транспонирование матрицы?

1. Замена столбцов на строки и наоборот.
2. Вычисление обратной матрицы.
3. Перенос строк из начала в конец матрицы.
4. Умножение матрицы на единичную.

7. Что такое центрирование и нормирование признаков?

1. Подгонка под нормальное распределение.
2. Деление среднеквадратичного отклонения на среднее.
3. Вычитание среднего и деление на дисперсию.
4. Вычитание среднего и деление на среднеквадратичное отклонение.

8. Как перевести количественный признак в порядковый?

1. Разбить по порядку на группы с одинакового размера и присвоить каждому объекту ранг группы.
2. Никак.
3. Присвоить каждому объекту его ранг.
4. Разбить на группы с одинаковым диапазоном изменений и присвоить каждому объекту среднее группы.

9. Как перевести в количественный качественный признак с двумя градациями?

1. Приписать каждому объекту его номер.
2. Обозначить одну градацию нулем, другую – единицей и приписать эти значения объектам, входящим в соответствующие градации.
3. Приписать каждому объекту случайное число.
4. Никак.

10. Как перевести в набор количественных признаков качественный признак с более чем двумя градациями?

1. Никак.
2. Образовать для каждой градации отдельный признак и заполнить его случайными числами.
3. Взять в качестве значений коды градаций.
4. Образовать для каждой градации отдельный признак со значениями, равными 1, если объекты попадают в эту градацию, и 0 – в противном случае.

11. Можно ли обрабатывать порядковые и качественные признаки по формулам для количественных признаков?

1. Нельзя. 2. Иногда можно. 3. Почти всегда можно. 4. Всегда можно.

12. Что такое скаляр?

1. Набор чисел. 2. Вещественное число. 3. Положительное число. 4. Целое число.

13. Что такое вектор?

1. Набор скаляров.
2. Вещественное число.
3. Несколько наборов чисел.
4. Расстояние между двумя точками.

14. Что такое матрица?

1. Набор чисел.
2. Набор векторов одинаковой длины.
3. Вещественное число.
4. Набор векторов разной длины.

15. Как определяется скалярное произведение векторов?

1. Поэлементное произведение векторов.
2. Произведение попарной суммы координат векторов.
3. Поэлементная сумма векторов.
4. Сумма попарного произведения координат векторов.

16. Как определяется сложение матриц?

1. Сложение каждого элемента первой матрицы с суммой всех элементов второй матрицы.
2. Сложение каждой строки первой матрицы с каждым столбцом второй матрицы.
3. Поэлементная сумма матриц.
4. Сумма всех элементов первой матрицы с суммой всех элементов второй матрицы.

17. Как определяется умножение матриц?

1. Поэлементное произведение матриц.
2. Скалярное произведение каждой строки первой матрицы на каждый столбец второй матрицы.
3. Скалярное произведение каждой строки первой матрицы на каждую строку второй матрицы.
4. Произведение всех элементов первой матрицы на произведение всех элементов второй матрицы.

18. Что такое единичная матрица?

1. Матрица, у которой по главной диагонали стоят нули, а остальные элементы равны единице.
2. Матрица, состоящая из одних единиц.
3. Матрица, у которой скалярное произведение строк (столбцов) самих на себя равно единице, а скалярное произведение строк (столбцов) на другие строки (столбцы) равно нулю.
4. Матрица, у которой по главной диагонали стоят единицы, а остальные элементы равны нулю.

19. Что такое диагональная матрица?

1. Матрица, у которой по главной диагонали стоят ненулевые числа, а остальные элементы равны нулю.
2. Матрица, у которой по главной диагонали стоят любые числа, а остальные элементы равны нулю.
3. Матрица, у которой по главной диагонали стоят единицы, а остальные элементы равны нулю.
4. Матрица, у которой по главной диагонали стоят нули, а остальные элементы равны единице.

20. Что такое ортогональная матрица?

1. Матрица, при умножении которой саму на себя получается единичная матрица.
2. Матрица, у которой скалярное произведение строк (столбцов) самих на себя

равно единице, а скалярное произведение строк (столбцов) на другие строки (столбцы) равно нулю.
3. Матрица, при умножении на которую получается ортогональная матрица.
4. Матрица, при умножении на которую другая матрица не меняется.

21. Какое геометрическое преобразование соответствует центрированию признаков?

1. Растяжение (сжатие) выборки по произвольным направлениям в многомерном пространстве.
2. Поворот совокупности объектов в многомерном пространстве.
3. Перенос начала координат в центр тяжести выборки.
4. Растяжение (сжатие) выборки по координатным осям.

22. Какое геометрическое преобразование соответствует нормированию признаков?

1. Поворот совокупности объектов в многомерном пространстве.
2. Перенос начала координат в центр тяжести выборки.
3. Растяжение (сжатие) выборки по произвольным направлениям в многомерном пространстве.
4. Растяжение (сжатие) выборки по координатным осям.

23. Какое геометрическое преобразование соответствует умножению матрицы «объект-признак» на ортогональную матрицу?

1. Перенос начала координат в центр тяжести выборки.
2. Поворот совокупности объектов в многомерном пространстве.
3. Растяжение (сжатие) выборки по координатным осям.
4. Растяжение (сжатие) выборки по произвольным направлениям в многомерном пространстве.

24. Что такое линейная комбинация признаков?

1. Сумма признаков с некоторыми коэффициентами, дающая новый признак.
2. Расположение признаков на числовой оси.
3. Перестановка исходных признаков.
4. Произведение признаков с некоторыми коэффициентами, дающее новый признак.

25. Как линейная комбинация признаков соотносится с произведением матрицы на вектор?

1. Линейная комбинация признаков НЕ является произведением матрицы «объект-признак» на собственный вектор корреляционной матрицы.
2. Линейная комбинация признаков НЕ является произведением матрицы «объект-признак» на вектор.
3. Линейная комбинация признаков является произведением матрицы «объект-признак» на вектор.
4. Линейная комбинация признаков всегда является произведением матрицы

«объект-признак» на собственный вектор корреляционной матрицы.

26. Чему произвольная линейная комбинация признаков соответствует в многомерном пространстве объектов?

1. Направлению с максимальной дисперсией в многомерном пространстве объектов.
2. Направлению с минимальной дисперсией в многомерном пространстве объектов.
3. Направлению в многомерном пространстве объектов.
4. Центру тяжести выборки.

27. Что такое корреляционная матрица?

1. Матрица коэффициентов корреляции между признаками.
2. Центрированная и нормированная матрица «объект-признак», умноженная сама на себя.
3. Матрица коэффициентов корреляции между объектами.
4. Центрированная и нормированная матрица «объект-признак».

28. Что такое собственный вектор корреляционной матрицы?

1. Вектор, для которого умножение на корреляционную матрицу эквивалентно умножению на скаляр.
2. Диагональ корреляционной матрицы.
3. Строка корреляционной матрицы.
4. Столбец корреляционной матрицы.

29. Что такое собственные значения корреляционной матрицы?

1. Строка корреляционной матрицы.
2. Скаляры, которые получаются при умножении корреляционной матрицы на ее собственные вектора.
3. Дисперсии исходных признаков.
4. Значения, стоящие по диагонали корреляционной матрицы.

30. Что НЕ является главной компонентой исходной матрицы «объект-признак»?

1. Направление с максимальной дисперсией в многомерном пространстве объектов.
2. Произведение признаков с некоторыми коэффициентами, дающее новый признак.
3. Матрица «объект-признак», умноженная на собственный вектор.
4. Направление с минимальной дисперсией в многомерном пространстве объектов.

31. Как устранить межвыборочную изменчивость в случае нескольких выборок с одинаковой ковариационной матрицей?

1. Нормировать каждую выборку отдельно.

2. Центрировать и нормировать все выборки общими средними и дисперсиями.
3. Центрировать все выборки общими средними.
4. Центрировать каждую выборку отдельно.

32. Как определяется главная дискриминантная ось?

1. Направление, в проекции на которое отношение межвыборочной дисперсии к объединенной внутривыборочной максимально.
2. Направление, в проекции на которое общая дисперсия выборки максимальна.
3. Направление, в проекции на которое межвыборочная дисперсия выборки максимальна.
4. Направление, в проекции на которое разница общей дисперсии и объединенной внутривыборочной максимальна.

33. Как провести дискриминантный анализ при вырожденности исходной матрицы «объект-признак»?

1. Никак.
2. Центрировать каждую выборку отдельно.
3. Преобразовать матрицу «объект-признак» в главные компоненты и отбросить компоненты с малыми дисперсиями.
4. Взять другой статистический пакет.

34. Что такое линейная регрессия в одномерном случае?

1. Линейная зависимость одной переменной от другой, найденная по методу наименьших квадратов.
2. Линейная зависимость одной переменной от другой, полученная с помощью логарифмического преобразования.
3. Линейная зависимость одной переменной от другой, полученная с помощью двойного логарифмического преобразования.
4. Зависимость одной переменной от другой, получающаяся при соединении точек прямыми линиями.

35. Что такое множественная линейная регрессия?

1. Прямая линия, максимально близко проходящая через множество объектов.
2. Множество объектов, соединенное отрезками прямых линий.
3. Линейная зависимость многих переменных от одной, найденная по методу наименьших квадратов.
4. Линейная зависимость одной переменной от других, найденная по методу наименьших квадратов.

36. Как вычислить множественную линейную регрессию при вырожденности исходной матрицы «объект-признак»?

1. Никак.
2. Преобразовать матрицу «объект-признак» в главные компоненты и отбросить компоненты с большими дисперсиями.
3. Вычислить гребневую (ridge) регрессию.
4. Взять другой статистический пакет.

37. Указать пример, позволяющий дать биологическую интерпретацию направления в многомерном пространстве.

1. Направление, в проекции на которое достигается максимальная корреляция между родителями и потомками.
2. Направление, в проекции на которое общая дисперсия выборки родителей и потомков минимальна.
3. Направление, в проекции на которое достигается минимальное различие между родителями и потомками.
4. Направление, в проекции на которое достигается максимальная разница между дисперсиями выборок родителей и потомков.

38. Какие нелинейные зависимости можно аппроксимировать нейронной сетью?

1. Только с одним максимумом. 2. Только монотонные.
3. Любые. 4. Только квадратичные.

39. Что такое обучающая выборка?

1. Выборка, которая используется в целях обучения студентов.
2. Выборка, которая прогоняется через обученную нейронную сеть.
3. Множество преподавателей вуза, обучающее данную группу студентов.
4. Выборка, по которой подгоняются веса нейронной сети.

40. Для чего нужна контрольная выборка?

1. Для проверки качества обученной нейронной сети.
2. Для проверки качества обучения и контроля знаний студентов.
3. Для подгонки весов нейронной сети.
4. Для поиска вида нелинейной зависимости между входными и выходными переменными.

41. Какое минимальное количество нейронов может содержать нейронная сеть?

1. Два. 2. Ни одного. 3. Три. 4. Один.

42. Указать пример евклидовой меры различия между объектами.

1. Квадратный корень из суммы разностей значений нескольких признаков.
2. Коэффициент корреляции.
3. Модуль разности значений некоторого признака.
4. Коэффициент Жаккара-Наумова.

43. Как вычислить вклады признаков в оси многомерного шкалирования?

1. Вычислить модули коэффициентов корреляции осей многомерного шкалирования с исходными признаками.
2. Никак.

3. Вычислить коэффициенты корреляции осей многомерного шкалирования с исходными признаками.
4. Вычислить квадраты коэффициентов корреляции осей многомерного шкалирования с исходными признаками.

44. Как представить отрезки временного ряда точками в многомерном пространстве?

1. Взять в качестве координат разности значений между соседними значениями отрезка временного ряда.
2. Отложить на числовой оси разность между начальным и конечным значениями отрезка временного ряда.
3. Никак.
4. Взять в качестве координат последовательность значений отрезка временного ряда.

45. Сколько переменных достаточно наблюдать для нахождения аттрактора многомерной динамической системы?

1. Две. 2. Одну. 3. Все. 4. Две статистически независимых.

46. Как визуализировать аттрактор динамической системы?

1. Применить авторегрессию.
2. Последовательные отрезки временного ряда, отражающего динамику одной из переменных, представить в виде матрицы и обработать методом главных компонент. Построить двумерный или трехмерный график зависимости одной из компонент от других.
3. Построить двумерный или трехмерный график зависимости одной из переменных динамической системы от других.
4. Применить спектральный анализ.

МНОГОМЕРНЫЙ АНАЛИЗ

БИОЛОГИЧЕСКИХ ДАННЫХ

Учебное пособие

2-е исправленное и дополненное издание

Вадим Михайлович Ефимов

Вера Юрьевна Ковалева

e-mail: **vmefimov@ngs.ru**

Научное издание. RIZO-печать
ИННОВАЦИОННЫЙ ЦЕНТР ЗАЩИТЫ РАСТЕНИЙ (ВИЗР)
Лицензия ПЛД № 69-253. Подписано к печати 28 января 2008 г.