

Министерство науки, высшей школы и технической политики
Российской Федерации

РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ ИНСТИТУТ

П. А. ВАЙНОВСКИЙ, В. Н. МАЛИНИН

МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ОКЕАНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

МНОГОМЕРНЫЙ АНАЛИЗ

САНКТ-ПЕТЕРБУРГ

1992

Вайновский П. А., Малинин В. Н. Методы обработки и анализа океанологической информации. Многомерный анализ. Учебное пособие. СПб., изд. РГГМИ, 1992. — 96 с.

Рассматриваются наиболее известные методы многомерной математической статистики и их использование применительно к обработке и интерпретации океанологических данных.

Предназначается для студентов-океанологов гидromетеорологических вузов и университетов.

Ил. 25. Табл. 5. Библ. 13.

The textbook deals with the best-known methods of multivariate mathematical statistics and their use as applied to the processing and interpretation of oceanological data.

The textbook designed to be used by students of oceanology at hydrometeorological colleges and universities.

Рецензенты: кафедра океанологии Одесского гидromетеорологического института; *И. В. Бовыкин*, канд. техн. наук (Институт озерадения АН СССР).

Научный редактор: проф. *Н. П. Смирнов*.

© Вайновский П. А., Малинин В. Н., 1992.

ПАВЕЛ АНТОНОВИЧ ВАЙНОВСКИЙ

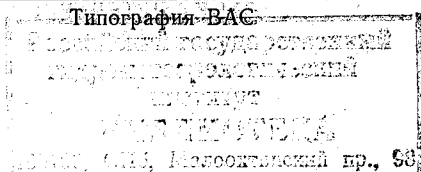
ВАЛЕРИЙ НИКОЛАЕВИЧ МАЛИНИН

МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ОКЕАНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ. МНОГОМЕРНЫЙ АНАЛИЗ

Учебное пособие

Редактор *О. С. Крайнова*

Сдано в набор 29.05.92. Подписано в печать 3.11.92.
Формат 60×90¹/₁₆. Бумага тип. № 2. Лит. гарн. Печать высокая. Зак. 361.
Печ. л. 6,0. Уч.-изд. 6,6. Тир. 400.
РГГМИ. 195196. Санкт-Петербург, Малоохтинский пр., 98. «С» № 5.



356575

ВВЕДЕНИЕ

Многомерный статистический анализ (МСА) широко вторгается в практику гидрометеорологических исследований уже в течение последних 20 лет. На сегодняшний день МСА широко зарекомендовал себя при решении актуальных задач прикладной гидрометеорологии. Методы МСА послужили эффективным инструментом анализа изменчивости ледовых условий в Арктическом бассейне. Они оказались надежным средством автоматизированного диагноза структуры вод океана, выделения водных масс и фронтальных зон, типизации сложных процессов взаимодействия океана и атмосферы. Методы МСА помогают восстанавливать отсутствующие данные при интерполяции в пространстве и во времени. Наконец, принципы многомерной статистики получили распространение при решении задач прогнозирования гидрометеорологических процессов и полей. Сегодня становится все более очевидной не только большая роль методов МСА в диагнозе и прогнозе состояния океана, но и широкая перспектива развития применения этих методов в будущем.

Исторический опыт применения методов МСА в гидрометеорологии связывается с именами выдающихся советских ученых: Н. А. Багрова, Ю. В. Николаева, А. М. Обухова и др., усилиями которых были сформулированы первые идеи многомерного статистического анализа гидрометеорологических процессов и получены первые интересные результаты.

Современный ученый-гидрометеоролог обязан знать методы многомерного статистического анализа и уметь применять их на практике. В данном учебном пособии обобщается опыт применения МСА в океанологических исследованиях, дается обзор основных методов МСА. По каждому из рассматриваемых методов последовательно формулируется постановка задач, рассматриваются основные уравнения и схема вычислений, обсуждаются вопросы программной реализации методов на ЭВМ и демонстрируются примеры применения методов МСА к решению типовых океанологических задач.

Спектр рассматриваемых методов МСА составлен с позиций широкого применения конкретных методик именно в прикладных океанологических задачах. Часть методов МСА, таких, как многомерный спектральный анализ, многомерный дисперсионный анализ и др., здесь не приводится, поскольку учебное пособие не может осветить все современные методы МСА, используемые в океа-

нологии. Интересующие читателя аспекты практического применения МСА или некоторые вопросы теории этих методов могут быть почерпнуты из специальной литературы, список которой приведен в конце учебного пособия.

Проблемы обработки данных на ЭВМ с применением методов МСА рассматриваются с общих позиций формирования расчетных алгоритмов. Вопросы программного обеспечения вычислительного процесса обсуждаются в рекомендуемой литературе, а также в соответствующих разработках учебно-вычислительной лаборатории РГГМИ.

Аппарат МСА основан на представлении и обработке данных в матричном виде, поэтому изложение основ методик будет проводиться в матричных обозначениях. Такое описание вычислительных процессов позволяет существенно упростить структуру записи действий, что выгодно как для общего понимания алгоритмов, так и для эффективной их реализации на ЭВМ. Неподготовленному читателю следует во избежание трудностей обратиться к классическим курсам матричной алгебры. В приложении данного учебного пособия приведена краткая сводка матричных обозначений и описание основных схем матричных вычислений, которые могут помочь при проработке лекционного материала.

ГЛАВА 1. МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ В ОКЕАНОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ: ОБЩИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ

1.1. Понятие о методах МСА

Современное состояние океана характеризуется примерным равновесием основных термодинамических и биогеохимических процессов, формирующих данную многосвязную систему. Это, однако, не означает абсолютной стабильности всех свойств его в пространстве и времени. Совокупность действий внешних и внутренних вынуждающих сил приводит к возникновению изменчивости одного или нескольких параметров состояния океана, что влечет за собой прямое или косвенное реагирование всей системы признаков. Естественно, что выводимый из равновесного состояния океан стремится вернуться в исходное состояние. Но так как внешние факторы (приливообразующие силы, ветер, процессы тепло- и влагообмена и др.) действуют на океан постоянно, то также постоянной является реакция его на эти воздействия. Заметим, что диапазон пространственно-временных масштабов изменчивости океанических процессов очень широк: от сантиметров и секунд до Мирового океана в целом и тысячелетий.

Ограниченность нашего знания океана не позволяет сразу однозначно проследить и интерпретировать наблюдаемые в нем изменения, выявить и описать причинно-следственный механизм колебаний. Поэтому методы статистического анализа выступают в данном случае как инструмент поиска, выявления количественного описания подобных взаимосвязей в пространстве и времени. При этом точность и надежность получаемых выводов во многом определяется количеством и разнообразием исходной информации, привлекаемой к анализу.

Разумеется, что вид исходной информации и методология ее статистической обработки тесно взаимосвязаны между собой. В классических методах одномерного анализа рассматривается единственный вектор наблюдений во времени или в пространстве вида $T(t)$ или $T(z)$. При этом мы имеем возможность исследовать только внутреннюю структуру конкретного вектора. Чтобы изучить взаимосвязь хотя бы двух аналогичных векторов $T_1(t)$ и $T_2(t)$, необходимо обратиться к методам парного корреляционного и спектрального анализа. И только техника многомерного

статистического анализа дает возможность проследить взаимосвязи целого множества векторов наблюдений во всем их многообразии.

Для современного этапа океанологических исследований характерен комплексный подход к океану как сложной многосвязной системе. Экспериментальная океанология ставит своей целью добыть максимум информации о физических, термодинамических, химических, биологических, оптических, акустических и других процессах и явлениях в океане. Теоретическая океанология решает задачи совместного описания разномасштабных процессов, различных по своей природе, но взаимосвязанных в едином глобальном механизме океанской изменчивости.

Исчерпывающая характеристика океана, очевидно, может быть дана на основе тщательного анализа внутренних взаимосвязей и общей изменчивости всех имеющихся материалов наблюдений. Совокупность данных при этом можно наглядно представить в виде «куба данных» (рис. 1.1). Куб данных составляют три рабочих оси: пространство, время, переменные (признаки). Как же расположены данные наблюдений в таком кубе?

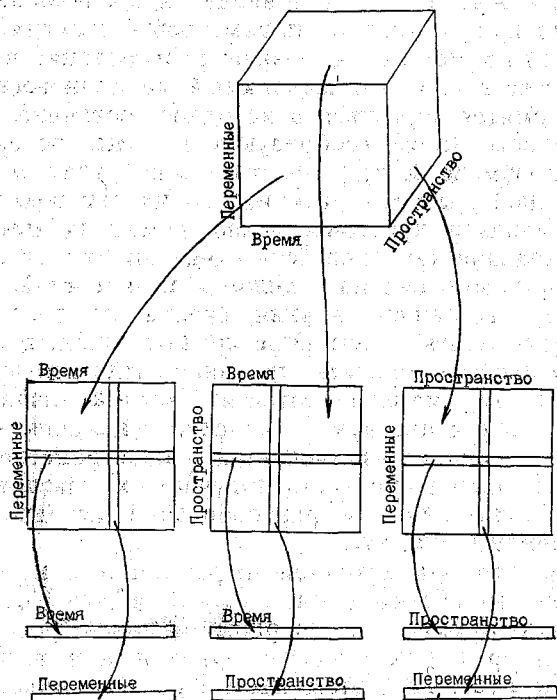


Рис. 1.1. Куб данных океанологической информации, составляющие его оси и плоскости наблюдений

Например, длительная регистрация температуры воды в одной точке в период времени $t_1, t_2, \dots, t_n$ приводит к получению вектора наблюдений $T(t)$, который в кубе данных параллелен оси времени. Единичное наблюдение вертикального распределения температуры воды приводит к получению вектора данных, параллельного оси пространства $T(z)$. Последовательное во времени многократное зондирование вертикального распределения температуры воды в одной точке приведет к получению «плоскости данных», располагающейся в кубе параллельно оси признаков. Если же совместно с температурой многократно наблюдать другие параметры состояния океана (физические, динамические, химические и т. д.), то в результате получается искомый куб данных, заполненный во всех трех выбранных координатных осях $T(z, t, x)$. Таким образом, практически вся информационная база накапливается как бы в трех плоскостях: пространство—время, пространство—признаки, признаки—время, которые в совокупности образуют куб данных. Исследование куба данных с привлечением информации более чем о двух параметрах состояния требует от исследователя перехода к многомерному пространству, изучение которого основывается на специализированном математическом аппарате.

Основная цель подавляющего большинства методов МСА состоит в таком преобразовании исходных данных, которое позволило бы исследователю относительно просто и достаточно полно представить скрытую сущность ведущих взаимосвязей в анализируемом материале. Такое преобразование исходных данных есть последовательность математических операций с количественными и качественными характеристиками, главная задача которой — достоверно описать систематические взаимосвязи между исследуемыми параметрами, исключив при этом из анализа случайную (шумовую) составляющую изменчивости. Разнообразие возможных теоретических подходов к статистическому анализу сложных систем обусловило возникновение и развитие целой совокупности методов МСА, каждый из которых по-своему интерпретирует изменчивость данных, выдвигая на первый план те или иные их характерные особенности.

На первый взгляд, содержание ряда алгоритмов МСА состоит в обобщении на многомерный случай уже изучаемых ранее методик парного корреляционного и регрессионного анализа. Чем же тогда объясняется преимущество методов МСА? Ответ заключается в том утверждении, что изначально выбранная единственная пара исследуемых параметров отнюдь не всегда представляется идеальной для полного описания изучаемого процесса. Привлечение же методов МСА позволяет, включив гораздо больше параметров, во-первых, отобрать наиболее информативные характеристики; во-вторых, описать структуру взаимосвязей характеристик процесса во всем многообразии его проявлений, и, наконец, в-третьих, построить простую и эффективную модель

процесса. В результате получается более полное и глубокое понимание происходящего.

Гидрометеорологические явления обусловлены большим количеством прямо и косвенно влияющих факторов. Полный учет влияния всех факторов в реальных условиях практически невозможен: во-первых, в силу неполноты нашего теоретического знания о происходящих процессах в океане, во-вторых, в виду ограниченных технических возможностей исследования океана. Поэтому при исследовании большинства причинно-следственных связей для анализа используют лишь те факторы, которые по уже имеющимся физическим соображениям следует принять за наиболее важные и функционально определяющие изменчивость изучаемых процессов. Смысл применения методов многомерного статистического анализа в данном случае можно разъяснить двояко.

Во-первых, многомерный статистический анализ выступает здесь как методология представления уже ранее выявленных взаимосвязей. Иными словами, МСА — это инструмент статистического описания известных взаимозависимостей. Примером подобной реализации МСА могут служить широко распространенные в практической океанологии различные эмпирические уравнения состояния морской воды, в которых на основе применения метода множественной регрессии реализованы известные взаимосвязи плотности морской воды с температурой, соленостью, давлением и т. п.

Во-вторых, методология МСА выступает и как фундаментальная основа познания исследуемых процессов в океане. На основе этих методов удастся не только описать сложные системы признаков вообще, но и выполнить наиболее эффективное подобное описание. Это означает, что методами МСА можно отобрать и выявить главные факторы, формулирующие данный процесс, из всей совокупности возможных влияющих факторов и сил. Здесь методы МСА выступают как инструменты формулирования или проверки конкретных физических гипотез. Примером данной функции методов МСА может служить применение процедур классификации на ЭВМ для выделения водных масс в океане на основе известных способов T , S -анализа. Опыт проводимых исследований позволяет в ряде случаев утверждать более высокую объективность и надежность методов МСА при выделении водных масс по сравнению с традиционным подходом.

Данное обстоятельство, на наш взгляд, представляется особенно актуальным, так как широкое внедрение ЭВМ требует скорейшего перевода многих субъективных методов океанологического анализа на рельсы формализованного объективного подхода. Методы МСА при этом выступают не только как возможные базовые методики для подобного перехода, но и как схемы оценивания качества традиционных методов анализа океанологической информации в новых условиях массовой обработки крупных массивов натурных наблюдений.

1.2. Классификация методов МСА

Прежде всего отметим, что МСА относится к числу наиболее интенсивно развивающихся областей математического анализа. Поэтому постоянно появляются все новые и новые методы МСА. В данном пособии мы ограничимся рассмотрением тех методов МСА, которые, с одной стороны, хорошо разработаны в теоретическом отношении, а с другой — уже достаточно широко используются при решении различных океанологических задач. К их числу относятся: регрессионный, дисперсионный, компонентный, факторный, дискриминантный, кластерный и некоторые другие методы. На фундаментальной основе этих методов в настоящее время реализовано более сотни различных модификаций, которые оказываются более эффективными при решении конкретных задач. Многообразие методов МСА указывает на широкие возможности совершенствования данного подхода при решении разнообразных вопросов практической океанологии. При этом задача исследователя состоит в том, чтобы выбрать и использовать наиболее эффективный методический прием.

Эффективность применения МСА на практике определяется и проверяется точностью соответствия построенной статистической модели натурным данным, а также достоверной физической интерпретацией полученных в модели взаимосвязей. Исследователь должен понимать, что низкий уровень отдельных результатов приложения МСА в океанологии говорит не столько о слабости метода, сколько о несоответствии выбранной методики исходным данным и поставленной задаче. Подход с использованием МСА исключает слепое применение стандартных методов анализа к данным. Эффективность МСА достигается только при осознанном, обоснованном и умелом применении этих достаточно сложных методов на практике.

Соответствие методов МСА решаемым задачам представляет здесь особый интерес. В общем виде это соответствие приводится в табл. 1.1. Исходя из целей статистического анализа рассмотрим соответствующие им методы.

Проблема определения подмножеств из всего рассматриваемого множества признаков может быть решена методом регрессионного анализа, где на основании корреляционных параметров представляется возможность выделить из всей совокупности признаков те из них, которые позволяют построить оптимальную модель описания изучаемого процесса. Сходная задача может быть решена и с применением метода кластер-анализа. Однако в этом случае выделение подмножества будет происходить уже как выделение группы объектов, подобных между собой по набору признаков. Кроме того, регрессионный анализ традиционно используется для построения моделей, а также интерполяции и экстраполяции, когда поведение единственной зависимой переменной необходимо описывать оптимальным сочетанием влияющих (взаимонезависимых) переменных.

Классификация методов МСА

Цели	Методы	Разновидности методик
Определение подмножеств	→ Множественный регрессионный анализ	Линейный Нелинейный
Построение моделей и экстраполяция	→ Множественный регрессионный анализ	Кусочный Пошаговый Гребневый
Классификация, типизация и районирование	→ Многомерный дискриминантный анализ → Кластер-анализ	Линейный Нелинейный Непараметрический Последовательный Параллельный Иерархический
Нахождение скрытой структуры данных	→ Компонентный и факторный анализ	Линейный Нелинейный Ортогональный Неортогональный
Экономное описание	→ Многомерный дисперсионный анализ → Канонические корреляции	

Применение многомерного дисперсионного анализа позволяет решить прежде всего задачу моделирования, причем наиболее эффективно — в плане оценивания вкладов различных факторов на формирование общей дисперсии исследуемого параметра.

На более высоком уровне решить задачи моделирования и экстраполяции позволяют методы компонентного и факторного анализа. С помощью этих методов может моделироваться и прогнозироваться изменчивость уже не одной переменной, а целой группы параметров. Для этого взаимосвязанная изменчивость исходных признаков описывается через значения новых статистически независимых переменных — компонент или факторов. Изменчивость компонент (факторов) объединяет в себе наиболее важные характеристики структуры исходных данных.

Поиск оптимального разбиения исследуемых объектов на однородные группы или классы наилучшим образом может осуществляться методами кластерного анализа, в которых учитывается при проведении расчетов совместное влияние факторов или переменных. При использовании процедур кластерного анализа исследователь, как правило, заранее не знает характеристики и количество возможных разбиений, которое определяется в процессе построения классификации. В случае, если классы или груп-

пы уже заранее определены, целесообразно использовать методы дискриминантного анализа. Суть последних состоит именно в оптимальном разбиении по уже известным классам новых объектов исследования.

При анализе пространственно-временной изменчивости признаков, особенно если они имеют разную размерность, наиболее эффективными являются методы компонентного, факторного анализа, которые позволяют решить задачу на основе преобразования матриц корреляций признаков. Метод дисперсионного анализа целесообразен при попытке изучения распределения дисперсии изменчивости наблюдаемых параметров по влияющим факторам. Кластерный анализ может оказаться полезным при оценке взаимосвязей между однородными группами или классами. Особое место при исследовании связности, коррелированности однородных классов занимает канонический корреляционный анализ, смысл которого и состоит в оценке коррелированности не отдельных пар объектов, а целых групп объектов.

Проблема экономного описания может рассматриваться как вершина статистического анализа. При ее решении обобщаются полученные выводы о связанности объектов, признаков в пространстве и времени и представляются в виде минимального числа новых статистически независимых переменных, которые обобщают систематическую изменчивость данных наблюдений.

ГЛАВА 2. МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. Постановка задачи

Несомненно, метод множественной регрессии относится к числу наиболее популярных в статистических исследованиях гидрометеорологических процессов. Это связано как с многофакторностью природных связей, так и с тем, что данный метод признается классическим методом анализа данных. Основы регрессионного анализа были сформулированы Гауссом более 100 лет назад. В течение многих десятилетий он развивался преимущественно в теоретическом направлении. Широкое внедрение его в практику связано, прежде всего, с созданием быстродействующих ЭВМ, способных обрабатывать большие массивы цифровых данных.

В настоящее время множественная регрессия используется практически во всех прикладных науках как один из наиболее универсальных методов анализа данных, причем для решения конкретных задач разработаны десятки различных модификаций метода. Кроме того, довольно часто множественная регрессия является составной частью других методик анализа данных. В океанологических исследованиях множественная регрессия ис-

пользуется при построении моделей и методов прогноза рядов, для интерполяции и восстановления полей, а также при решении целого ряда других задач.

Регрессионный анализ может быть двух видов: линейный и нелинейный. Так как основным считается линейный, то главное внимание будет уделено именно множественной линейной регрессии (МЛР). Постановка задачи МЛР состоит в следующем. Пусть имеется $M+1$ векторов $\vec{Y}, \vec{X}_1, \vec{X}_2, \dots, \vec{X}_M$, которые могут быть представлены в виде следующей матрицы:

$$\begin{array}{ccccccc} y_1 & x_{11} & x_{12} & \dots & x_{1M}, \\ y_2 & x_{21} & x_{22} & \dots & x_{2M}, \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ y_N & x_{N1} & x_{N2} & \dots & x_{NM}. \end{array}$$

Здесь изменчивость переменной $\vec{Y}$ рассматривается как результат влияния переменных $\vec{X}_1, \dots, \vec{X}_M$. Предполагая, что такое влияние можно описать посредством линейных взаимосвязей, основное уравнение МЛР можно записать как

$$y_i = a_0 + \sum_{j=1}^M a_j x_{ij} + \varepsilon_i. \quad (2.1)$$

В уравнении (2.1) y_i представляет зависимую переменную (функцию отклика, предиктант); x_{ij} — независимую переменную (фактор, предиктор); a_0 и a_j — соответственно свободный член и коэффициент уравнения регрессии; ε_i — ошибки описания (остатки) изменчивости переменной y_i с помощью линейной модели по выбранным параметрам $x_{i1}, \dots, x_{iM}$.

Нетрудно видеть, что МЛР представляет обобщение линейной регрессии двух переменных на многомерный случай. Однако если парная регрессия имеет четкую геометрическую интерпретацию, то для МЛР сделать это невозможно, так как не существует аналогичной интерпретации для многомерного пространства. Поэтому M -мерное пространство — это лишь удобный математический прием, позволяющий экстраполировать свойства двумерного пространства на многомерное. Таким образом, уравнение (2.1) представляет некую условную плоскость (гиперплоскость) в $(M+1)$ -мерном пространстве, которая обладает тем свойством, что разброс точек $(y_i, x_{i1}, \dots, x_{iM})$ около нее меньше, чем около любой поверхности.

Основная задача МЛР состоит в нахождении оптимальных значений коэффициентов регрессии, позволяющих наиболее точно описать изменчивость переменной y_i для данного вектора наблюдений. Естественно, чем точнее определяется на основе (2.1) зависимая переменная y_i , тем меньше остатки ε_i . В идеальном случае, когда остатки равны нулю, мы имеем функциональную ли-

$$X^T Y = \begin{bmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \sum y_i x_{i2} \\ \dots \\ \sum y_i x_{iM} \end{bmatrix}, \quad A = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \dots \\ a_M \end{bmatrix}$$

Заметим, что поскольку исходные данные обычно нормируются (стандартизируются)

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}, \quad \sigma_{x_j} = \frac{1}{N} \sum (x_{ij} - \bar{x}_j)^2,$$

то в этом случае матрица $X^T X$ приобретает вид

$$X^T X = \begin{bmatrix} N & Nr_{12} & \dots & Nr_{1M} \\ Nr_{21} & N & \dots & Nr_{2M} \\ \dots & \dots & \dots & \dots \\ Nr_{M1} & Nr_{M2} & \dots & N \end{bmatrix} \quad (2.4)$$

Диагональные элементы этой матрицы оказываются равными N , так как

$$\sum_{i=1}^N x_{ij}^2 = \frac{N}{N} \sum \frac{(x_{ij} - \bar{x}_j)^2}{\sigma_{x_j}^2} = N.$$

Делением на N матрица (2.4) превращается в корреляционную матрицу, т. е.

$$\Phi = \frac{1}{N} X^T X. \quad (2.5)$$

Матрица Φ называется иногда информационной матрицей Фишера и отражает степень линейной взаимосвязи всех факторов между собой.

Матрица $X^T Y$ при нормировании исходных данных превращается в вектор — столбец коэффициентов корреляции функции отклика со всеми факторами, т. е.

$$X^T Y = \begin{bmatrix} r_{y1} \\ r_{y2} \\ \dots \\ r_{yM} \end{bmatrix}$$

Для решения системы нормальных уравнений в матричной форме (2.3) следует умножить ее на матрицу, обратную матрице системы нормальных уравнений:

$$(X^T X)^{-1} (X^T X) A = (X^T X)^{-1} (X^T Y). \quad (2.6)$$

Но, поскольку произведение матриц в левой части выражения (2.6) представляет единичную матрицу I

$$(X^T X)^{-1}(X^T X) = I,$$

то решение системы нормальных уравнений в матричной форме запишется следующим образом:

$$A = (X^T X)^{-1} (X^T Y). \quad (2.7)$$

В поэлементном виде система нормальных уравнений может быть записана как

$$\begin{aligned} r_{y_1} &= \alpha_1 + \alpha_2 r_{12} + \dots + \alpha_M r_{1M}, \\ r_{y_2} &= \alpha_1 r_{21} + \alpha_2 + \dots + \alpha_M r_{2M}, \\ &\vdots \\ r_{y_M} &= \alpha_1 r_{M1} + \alpha_2 r_{M2} + \dots + \alpha_M, \end{aligned} \quad (2.8)$$

где $\alpha_j = (\sigma_j/\sigma_y) a_j$ — параметр, характеризующий относительный вес каждого коэффициента регрессии в уравнении (2.1). Данная система имеет однозначное решение относительно α_j . Например, использование для ее решения метода Крамера в конечном счете сводится к расчету величин α_j по следующей формуле:

$$\alpha_j = \frac{|D_{yj}|}{|D_{vv}|}, \quad (2.9)$$

где $|D_{yj}|$ и $|D_{yy}|$ — миноры определителя $|D_0|$, имеющего вид

$$\det D \equiv |D_0| = \begin{vmatrix} 1 & r_{y1} & r_{y2} & \dots & r_{yM} \\ r_{1y} & 1 & r_{12} & \dots & r_{1M} \\ r_{2y} & r_{21} & 1 & \dots & r_{2M} \\ \dots & \dots & \dots & \dots & \dots \\ r_{My} & r_{M1} & r_{M2} & \dots & r_{MM} \end{vmatrix}$$

Минор $|D_{yj}|$ представляет исходный определитель $|D_0|$, у которого вычеркнута первая строка и j -й столбец, а минор $|D_{y1}|$ означает определитель $|D_0|$, у которого вычеркнута первая строка и первый столбец.

С учетом (2.9) коэффициенты регрессии определяются как

$$a_j = \frac{\sigma_y}{\sigma_j} \frac{|D_{yj}|}{|D_{yy}|}. \quad (2.10)$$

Рассмотрим МЛР в наиболее простом виде: $y = a_0 + a_1 x_1 + a_2 x_2$. В этом случае определитель $|D_0|$ имеет вид

$$|D_0| = \begin{vmatrix} 1 & r_{y1} & r_{y2} \\ r_{1y} & 1 & r_{12} \\ r_{2y} & r_{21} & 1 \end{vmatrix}$$

Миноры этого определителя, необходимые для расчета коэффициентов регрессии, равны

$$|D_{y1}| = \begin{vmatrix} r_{y1} & r_{12} \\ r_{y2} & 1 \end{vmatrix} = r_{y1} - r_{y2}r_{12},$$

$$|D_{y2}| = - \begin{vmatrix} r_{y1} & 1 \\ r_{y2} & r_{12} \end{vmatrix} = r_{y2} - r_{y1}r_{12},$$

$$|D_{yy}| = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix} = 1 - r_{12}^2.$$

Отсюда нетрудно получить

$$a_1 = \frac{\sigma_y}{\sigma_1} \frac{|D_{y1}|}{|D_{yy}|} = \frac{\sigma_y}{\sigma_1} \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2},$$

$$a_2 = \frac{\sigma_y}{\sigma_2} \frac{|D_{y2}|}{|D_{yy}|} = \frac{\sigma_y}{\sigma_1} \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}.$$

Свободный член уравнения регрессии находится как

$$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2.$$

При $M > 2$ формулы для оценки коэффициентов регрессии становятся значительно сложнее. Уже при $N > 20$ и $M > 3$ выполнение расчетов вручную или с помощью микрокалькуляторов является весьма затруднительным. Если же матрица исходных данных весьма велика, то оценка параметров регрессии может быть осуществлена лишь с помощью быстродействующих ЭВМ.

В связи с этим модель МЛР очень часто включается в пакеты прикладных программ стандартного математического обеспечения статистического анализа данных на ЭВМ различных классов. Для ЭВМ класса ЕС схема линейного регрессионного анализа реализована как в советских пакетах прикладных программ СОМИ, ППСА, так и в зарубежных ВМДР и других. Современные цифровые ЭВМ позволяют не только выполнить собственно расчет параметров МЛР, но и провести более глубокое исследование регрессионных зависимостей выбранных переменных. Так, в большинстве перечисленных здесь пакетов программ может быть проведено пошаговое исключение или перебор независимых переменных с целью выбора оптимального количества признаков. Кроме того, появляются дополнительные возможности по предварительному преобразованию исходных признаков с целью их наиболее выгодного представления в модели МЛР (с помощью известных методов линеаризации).

2.3. Коэффициент множественной корреляции и его свойства

В качестве характеристики связи между зависимой переменной y_i и всеми независимыми переменными используется множественный (сводный) коэффициент корреляции R

$$R = \frac{\sum_{i=1}^N (y_i - \bar{y})(\tilde{y} - \bar{y})}{N \sigma_y \sigma_{y(x)}}, \quad (2.11)$$

где $\tilde{y}$ — вычисленные по уравнению регрессии значения y_i . Отсюда следует, что R можно рассматривать как коэффициент парной корреляции между фактическими и вычисленными значениями переменной y_i . Однако в отличие от коэффициента парной корреляции, который изменяется в пределах от -1 до 1 , множественный коэффициент корреляции имеет пределы изменений от 0 до 1 , т. е.

$$0 \leq R \leq 1.$$

При $R=1$ имеем функциональную линейную модель, когда факторы полностью описывают изменчивость функции отклика, в результате чего остатки равны нулю ($\varepsilon_i=0$). При $R=0$, напротив, изменчивость функции отклика полностью определяется остатками ε_i . Это означает также, что все коэффициенты парной корреляции вектора-столбца $X^T Y$ равны нулю.

Кроме того, коэффициент множественной корреляции может служить мерой точности уравнения регрессии. Можно показать, что

$$R^2 = \frac{D_{y(x)}}{D_y} = 1 - \frac{D_\varepsilon}{D_y}, \quad (2.12)$$

т. е. квадрат коэффициента множественной корреляции (коэффициент детерминации) численно равен доле полной дисперсии переменной y_i , которая обусловлена влиянием на нее факторов $x_{i1}, \dots, x_{im}$, и, следовательно, полностью учитывается уравнением регрессии.

Для определения коэффициента множественной корреляции может быть использована также формула вида

$$R = \sqrt{1 - \frac{|D_0|}{|D_{yy}|}}. \quad (2.13)$$

Заметим, что по абсолютной величине R всегда больше максимального коэффициента парной корреляции из вектор-столбца $X^T Y$, т. е. $R > (r_{yij})_{\max}$.

В общем виде коэффициент R связан с вектор-столбцом $X^T Y$ следующим соотношением:

$$R^2 = \frac{1}{N} (\alpha_1 r_{y1} + \alpha_2 r_{y2} + \dots + \alpha_m r_{ym}). \quad (2.14)$$

Нетрудно видеть, что произведение $\alpha_j r_{y_j}$ показывает вклад каждого из факторов x_{ij} в описание изменчивости y_i . Таким образом, влияние факторов x_{ij} на изменчивость y_i зависит не только от коэффициента корреляции, но и от величины коэффициента регрессии и изменчивости самого фактора x_{ij} . Однако, как правило, величина коэффициента корреляции оказывается более существенной по сравнению с величиной параметра α_j .

2.4. Оценивание параметров регрессии

В качестве критерия точности уравнения регрессии (2.1) используется среднее квадратическое отклонение между наблюдаемыми и рассчитанными по уравнению регрессии значениями, т. е.

$$\sigma_{y(x)} = \sqrt{\frac{\sum_{i=1}^N (\tilde{y}_i - y_i)^2}{N}} \quad (2.15)$$

или

$$\sigma_{y(x)} = \sigma_y \sqrt{1 - R^2}.$$

Нетрудно видеть, что этот критерий в точности совпадает с критерием точности уравнения регрессии двух переменных.

Стандартная погрешность множественного коэффициента корреляции определяется как

$$\sigma_R = \frac{1 - R^2}{\sqrt{N - M - 1}}, \quad (2.16)$$

а стандартная погрешность коэффициентов регрессии

$$\sigma_{\alpha_j} = \frac{\sigma_y}{\alpha_j} \sqrt{\frac{1 - R^2}{N - M - 1} \frac{D_{y_j}}{D_{yy}}}. \quad (2.17)$$

Формулы (2.15) — (2.17) применимы в том случае, если все M факторов подчиняются порознь и в совокупности нормальному закону распределения. Заметим, что использование формулы (2.16), строго говоря, правомерно лишь при $R < 0,5$, $N > 30$. При $R > 0,5$ и $N < 30$, как и в случае парной корреляции, необходимо применять преобразование Фишера.

Из формул (2.16) и (2.17) следует одно важное следствие. С увеличением длины рядов и уменьшением их числа точность регрессионной модели увеличивается. Поэтому при выполнении практических расчетов необходима выполнимость условия $N \gg M$.

Доверительные границы для коэффициента множественной корреляции и коэффициентов регрессии устанавливаются с помощью статистики Стьюдента:

$$\begin{aligned} R^* - t_{\alpha} \sigma_R < R < R^* + t_{\alpha} \sigma_R, \\ \alpha^* - t_{\alpha} \sigma_{\alpha_j} < \alpha_j < \alpha_j^* + t_{\alpha} \sigma_{\alpha_j} \end{aligned} \quad (2.18)$$

где t_α — статистика Стьюдента при заданном уровне значимости α и числе степеней свободы $\nu = N - M - 1$. Из (2.18) видно, что истинные оценки параметров отличаются от выборочных оценок на величину $t_\alpha \sigma$. При этом с увеличением длины рядов различия между ними уменьшаются.

Оценка значимости параметров регрессии осуществляется на основе нулевой гипотезы, которая в этом случае выбирается относительно проверки их на равенство нулю, т. е.

$$H_0: R^* = 0,$$

$$H_0: a_j = 0.$$

Затем строится критическая область вида

$$\begin{aligned} |R^*| &> t_\alpha \sigma_R, \\ |a_j^*| &\geq t_{\alpha/2} \sigma_{a_j}. \end{aligned} \quad (2.19)$$

Если данное условие выполняется, то нулевая гипотеза отвергается как несостоятельная и выборочные значения R^* и a_j^* считаются значимыми, т. е. не обусловленными случайными вариациями наблюдений.

Адекватность модели МЛР оценивается таким же образом, как и для модели парной регрессии. Вначале составляется нулевая гипотеза о равенстве дисперсий вычисленных по уравнению (2.1) значений y_i и остатков, т. е. $H_0: D_{y(x)} = D_\varepsilon$. Затем вычисляется дисперсионное отношение (F -критерий) $F^* = D_{y(x)} / D_\varepsilon$, которое сравнивается с табличным значением критерия Фишера при заданном уровне значимости и степенях свободы $\nu_1 = M$, $\nu_2 = N - M - 1$.

Если $F^* > F_{\text{табл}}$, то нулевая гипотеза отвергается и делается предположение, что дисперсия, описываемая МЛР, неслучайным образом отличается от дисперсии шума, т. е. рассматриваемая модель может считаться адекватной.

В некоторых случаях вместо адекватности вводится понятие эффективности модели, которое означает отношение дисперсии фактических значений y_i к дисперсии шума, т. е.

$$\gamma = D_y / D_\varepsilon.$$

Обычно считается, что модель является эффективной при $\gamma > 3$.

2.5. Понятие о частном коэффициенте корреляции

Определение частного коэффициента корреляции непосредственно вытекает из эффекта ложной корреляции. Напомним, что если два ряда переменных $\vec{X}_1$ и $\vec{X}_2$ не содержат в себе какой-либо информации о третьей переменной, то корреляция между $\vec{X}_1$ и $\vec{X}_2$ является истинной. Если же переменные $\vec{X}_1$ и $\vec{X}_2$ связаны функ-

ционально с третьей переменной, то возникает ложная корреляция.

При построении моделей МЛР мы имеем дело обычно с коррелированными факторами, причем оценить эффект ложной корреляции в факторах за исключением очевидных случаев (например, годовой или суточный ход) оказывается чрезвычайно сложно. В связи с этим и возникает задача определения «чистого» влияния каждой независимой переменной на зависимую. Для этого используется частный коэффициент корреляции, под которым понимается мера линейной связи между зависимой переменной $\vec{Y}$ и какой-либо одной из переменных $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_M$ после исключения влияния на нее всех оставшихся переменных.

Сказанное можно пояснить следующим образом. Если исключить из набора $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_M$, например, первую переменную $\vec{X}_1$, то модель МЛР для центрированных данных примет следующий вид:

$$y_i = \sum_{j=1}^N a_{j+1} x_{i(j+1)} + \varepsilon_i.$$

Кроме того, мы можем модель МЛР относительно X_1 представить как

$$x_{i1} = \sum_{j=1}^N a'_{j+1} x_{i(j+1)} + \varepsilon'_i.$$

По существу величина ε'_i и составляет ту часть переменной x_{i1} , которая полностью независима от влияния других переменных. В результате коэффициент парной корреляции между ε_i и ε'_i соответствует частному коэффициенту корреляции

$$r_{0y1} = \frac{\sum_{i=1}^N (\varepsilon_i - \bar{\varepsilon})(\varepsilon'_i - \bar{\varepsilon}')}{N \sigma_{\varepsilon} \sigma_{\varepsilon'}}.$$

Аналогичным образом могут быть получены частные коэффициенты корреляции y_i с другими переменными X_j .

Отсюда следует, что квадрат частного коэффициента корреляции может быть интерпретирован как доля остаточной дисперсии функции отклика y_i , объясненная включением дополнительной переменной в модель МЛР.

В общем случае формула для оценки частных коэффициентов корреляции имеет вид

$$r_{0y_j} = \frac{|D_{yj}|}{\sqrt{|D_{yy}| |D_{jj}|}} \quad (2.20)$$

где D_{jj} — минор определителя $|D_0|$, образованный путем вычеркивания j -й строки и j -го столбца. Значимость и доверительный интервал для коэффициентов частной корреляции определяются

аналогично, как и для коэффициентов парной корреляции, только число степеней свободы оценивается как $\nu = N - M - 3$.

В качестве примера рассмотрим модель МЛР в виде $y = a_0 + a_1x_1 + a_2x_2$. Пусть $r_{y1} = 0,70$, $r_{y2} = 0,40$, $r_{12} = 0,50$. Используя формулу (2.20), для частного коэффициента корреляции между y_i и x_{i1} можно получить

$$r_{0y1} = \frac{r_{y1} - r_{12}r_{y2}}{\sqrt{(1-r_{12}^2)(1-r_{y2}^2)}} = \frac{0,70 - 0,50 \cdot 0,40}{\sqrt{0,75 \cdot 0,84}} = 0,63.$$

При исключении из выборки второго фактора имеем

$$r_{0y2} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{12}^2)(1-r_{y1}^2)}} = \frac{0,40 - 0,50 \cdot 0,70}{\sqrt{0,75 \cdot 0,51}} = 0,09.$$

Итак, результаты расчетов показывают, что истинный вклад обоих факторов в изменчивость переменной y_i меньше коэффициентов парной корреляции r_{y1} и r_{y2} , т. е. корреляция между двумя факторами в определенной степени обусловлена действием какого-то третьего фактора. По разности величин r_{y1} и r_{0y1} нетрудно оценить эффект ложной корреляции. Для первого фактора он составляет $r_{\text{лож}} = 0,07$, а для второго фактора $r_{\text{лож}} = 0,31$.

Частные коэффициенты корреляции обычно используются для ранжирования факторов по значимости их влияния на функцию отклика в рамках моделей пошаговой регрессии.

2/6. Понятие о нелинейной регрессии

Связь между гидрометеорологическими процессами далеко не всегда является линейной. Поэтому построенная модель МЛР может оказаться неудачной именно по данной причине. В этом случае необходимо использовать модели множественной нелинейной регрессии.

Если нет каких-либо физических предпосылок о характере связи между зависимой переменной и аргументами, то обычно используют модель полиномиальной множественной регрессии. Например, двухфакторная модель имеет вид

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_1^3 + a_5x_1x_2 + a_6x_1^3 + a_7x_2^3 + a_8x_1^2x_2 + a_9x_1x_2^2 + \dots$$

Осуществляя замену переменных $x_1^2 = x_3$, $x_2^2 = x_4$, $x_1x_2 = x_5$, ..., нетрудно получить классическое уравнение МЛР, коэффициенты которого определяются методом наименьших квадратов.

Совершенно очевидно, что при достаточно большом числе факторов и, тем более, для достаточно высоких степеней полинома реализация полиномиальной регрессии даже на современных быстродействующих ЭВМ становится практически невозможной.

Поэтому на практике обычно используется малое число факторов (2—5), а сам полином, как правило, принимается не выше второй или третьей степени.

В тех случаях, когда существуют определенные предпосылки относительно формы связи (степенная, экспоненциальная и т. п.), используя некоторые преобразования, приводят эту зависимость к линейному виду и далее методом наименьших квадратов получают коэффициенты регрессии.

Например, если мы имеем зависимость

$$y = a_0 x_1^{a_1} x_2^{a_2} \dots x_M^{a_M},$$

то, логарифмируя ее по основанию e , нетрудно получить

$$\ln y = \ln a_0 + a_1 \ln x_1 + a_2 \ln x_2 + \dots + a_M \ln x_M.$$

Произведем замену переменных: $y' = \ln y$, $a_0' = \ln a_0$, $x_1' = \ln x_1$ и т. д. В результате получим уравнение МЛР, коэффициенты которого находятся с помощью МНК.

Другим примером нелинейной регрессии служат экспоненциальные модели вида

$$y = \exp(a_0 + a_1 x_1 + a_2 x_2 + \dots + a_M x_M),$$

$$y = a_0 \exp a_1 x_1 \exp a_2 x_2 \dots \exp a_M x_M,$$

которые линеаризуются следующим образом:

$$\ln y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_M x_M,$$

$$\ln y = \ln a_0 + a_1 x_1 + a_2 x_2 + \dots + a_M x_M.$$

Мерой точности нелинейной регрессии служит множественное корреляционное отношение, вычисляемое по формуле (2.11) либо по формуле (2.12). Сравнение множественного корреляционного отношения с коэффициентом множественной корреляции дает представление о степени нелинейности многофакторной зависимости, т. е. о ее «кривизне».

Определение стандартной погрешности, построение доверительных границ и оценка значимости для множественного корреляционного отношения осуществляются таким же образом, как и для коэффициента множественной корреляции.

2.7. О выборе оптимальных форм линейной регрессии

Поиск оптимальных форм моделей МЛР относится к числу наиболее актуальных задач, возникающих при анализе многофакторных зависимостей. Это связано с целым рядом причин. Так, большой набор факторов еще не гарантирует полного описания функции отклика. В то же время даже при высокой точности описания функции отклика погрешности параметров регрессии могут оказаться высокими, если разность $N - M$ невелика.

Еще один возможный источник погрешностей — это коррелированность факторов между собой. Вообще говоря, идеальная модель МЛР предполагает отсутствие корреляции между всеми факторами, т. е. их статистическую независимость. Однако на практике добиться этого чрезвычайно сложно, так как обычно факторы скоррелированы. Наличие высокой статистической связи между факторами приводит к тому, что система нормальных уравнений становится плохо обусловленной, вырождающейся и ее детерминант (главный определитель) стремится к нулю. Вследствие этого коэффициенты регрессии становятся неустойчивыми, ошибки их определения могут существенно превышать сами значения коэффициентов. Естественно, что такая модель МЛР является бесполезной с точки зрения анализа данных.

Итак, каковы же пути построения оптимальных моделей МЛР? Прежде всего существует целый ряд требований к исходной информации, без выполнения которых практически невозможно построить оптимальную модель МЛР. К ним относятся:

1. Все факторы и функции отклика должны подчиняться нормальному закону распределения. Хотя, в общем, требования нормальности не следуют из метода наименьших квадратов, тем не менее очевидно, что малые отклонения от математического ожидания должны встречаться значительно чаще, чем крупные отклонения. В этом случае оценки, полученные с помощью МНК, являются устойчивыми. При наличии в исходных рядах грубых промахов или выбросов коэффициенты регрессии могут существенно искажаться. Кроме того, следует иметь в виду, что приведенные выше оценки точности параметров регрессионной модели основаны на предположении о нормальности исходных данных.

2. Переменные должны отвечать условиям стационарности. На практике довольно часто возникает ситуация, когда исходные ряды нестационарны по математическому ожиданию, т. е. содержат в себе некоторую трендовую составляющую. Рассматривая исходные ряды как сумму случайной и трендовой составляющих, можно констатировать, что игнорирование трендов приводит к искажению степени влияния независимых переменных на зависимую переменную, т. е. по существу возникает эффект ложной корреляции. Естественно, этот эффект тем сильнее, чем больше вклад тренда в дисперсию ряда. Особенно опасен случай, когда модель МЛР используется для прогноза. Добиваясь высокой точности описания функции отклика на зависимом материале, при переходе к независимым данным можно получить результаты, далекие от первоначальной точности. Поэтому целесообразно исключить тренды из исходных рядов и самостоятельно проводить анализ и прогноз трендов.

3. Длина выборки должна значительно превосходить число независимых переменных. Однако какого-либо жесткого критерия, устанавливающего связь между величинами N и M , не существует.

4. Статистические связи между функцией отклика и факторами, а также между самими факторами должны быть линейными. Однако, если для модели парной регрессии проверка линейности связи не вызывает каких-либо затруднений и может осуществляться визуально, то для модели МЛР такая проверка не представляется возможной. Поэтому обычно используются косвенные методы проверки, основанные, с одной стороны, на физических соображениях о статистических связях исследуемых процессов, а с другой — на сравнении множественного корреляционного отношения с коэффициентом множественной корреляции.

5. Погрешности зависимой и независимой переменных должны быть примерно одинаковыми. Если, например, функция отклика имеет малые случайные погрешности, а все факторы, напротив, определяются с большими ошибками, то практически невозможно получить адекватную модель МЛР. Приблизительно о степени соответствия погрешностей друг другу можно судить по величинам стандартного отклонения, если все факторы и функция отклика имеют одну размерность, или по коэффициентам вариации, если они имеют различную природу.

6. Факторы должны быть переменными, т. е. влиять на изменчивость функции отклика. Вообще говоря, возможна ситуация, когда физическое воздействие какого-либо фактора на функцию отклика не вызывает сомнений, однако в статистическом плане такая связь может оказаться незначимой. Например, солнечная радиация является важнейшей причиной формирования годовых величин температуры поверхностного слоя океана. Тем не менее межгодовые изменения температуры воды, как правило, статистически очень слабо связаны с межгодовыми колебаниями суммарной радиации, изменчивость которых обычно мала. Переменность факторов проверяется оценкой коэффициентов регрессии на значимость. Незначимые коэффициенты регрессии должны исключаться.

7. Факторы должны быть статистически независимы (некоррелированы) между собой. При наличии высокой корреляции между отдельными факторами целесообразно исключать один из дублирующих факторов. В результате исключения объем информации уменьшается незначительно, но зато система нормальных уравнений будет иметь более устойчивое решение относительно коэффициентов регрессии.

Итак, соблюдение указанных требований к исходной информации является одним из основных этапов построения оптимальной модели регрессии.

Следующий этап — это анализ факторов с целью наиболее эффективного использования их в моделях МЛР. Существует несколько способов анализа факторов. В частности, к ним относятся:

1. Метод всех регрессий.
2. Метод включения переменных.

3. Метод исключения переменных.

Метод всех регрессий заключается в построении путем перебора всех возможных видов МЛР и выборе наилучшего уравнения. Если, например, число факторов равно 3, то общее число уравнений составит 8. При $M=4$ число уравнений возрастает до 16. Следовательно, существует простая зависимость $k=2^M$, где k — число уравнений. Естественно, что при большом числе переменных возникают уже трудности технического характера.

Кроме того, могут появиться трудности иного характера, если для нескольких уравнений МЛР коэффициент детерминации окажется равным или почти равным. Возникает вопрос, какое из этих уравнений выбрать, т. е. какие факторы следует предпочесть для описания функции отклика.

Очевидно, для сравнения важности факторов прежде всего следует использовать формальные критерии: коэффициент множественной корреляции, критерий Фишера, t -критерий Стьюдента, частные коэффициенты корреляции, ошибки коэффициентов регрессии, коэффициенты парной корреляции функции отклика с факторами. Вероятность того, что все эти критерии для сравниваемых факторов окажутся одинаковыми, является ничтожной. Однако следует иметь в виду, что использование лишь формальных критериев еще не служит полной гарантией построения оптимальной модели, поскольку необходимо учитывать и неформальные критерии: качество и стоимость информации, ее доступность, физическую обусловленность связей и др.

Суть метода включения переменных состоит в ранжировании факторов и последовательном включении их в расчёт с анализом получаемой модели регрессии на каждом шаге. Алгоритм, реализующий данный метод, получил название пошаговой регрессии, блок-схема которой представлена на рис. 2.1. В соответствии с этой блок-схемой вначале осуществляется предварительная обработка данных, заключающаяся в проверке требований, предъявляемых к исходной информации. Если число факторов не-

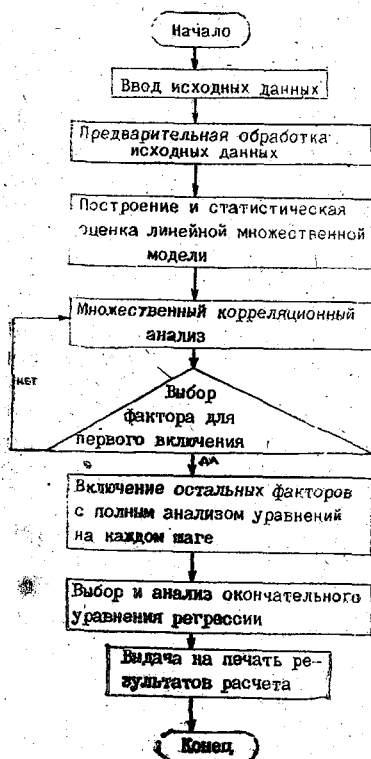


Рис. 2.1. Блок-схема алгоритма пошаговой множественной регрессии

лико и нет ограничений на линейность модели, дополнительно вычисляются нелинейные члены:

$$x_{M+1} = x_1^2, \quad x_{M+2} = x_2^2, \quad x_{M+3} = x_1 x_2 \text{ и т. д.}$$

После этого по полному числу факторов рассчитывается корреляционная матрица, решается система нормальных уравнений, определяются коэффициенты регрессии и коэффициент множественной корреляции. Это нужно, чтобы произвести ранжирование факторов в соответствии с выражением (2.14). В некоторых случаях для ранжирования используется упрощенная процедура, заключающаяся в оценке парных коэффициентов корреляции функции отклика с факторами. Могут приниматься во внимание также и неформальные критерии — стоимость информации, ее доступность и т. п.

В результате ранжирования получаем последовательность факторов по значимости. На первом шаге строится парная регрессия с наиболее значимым фактором. При этом осуществляется полный анализ модели, т. е. рассчитываются дисперсия остатков, коэффициент детерминации, критерий Фишера, критерий Стьюдента, погрешности параметров модели. На втором шаге к первому фактору добавляется второй, строится уже множественная регрессия и осуществляется ее полный анализ, как и в методе всех регрессий. Аналогичным образом включаются в модель другие факторы. В результате получим M моделей, соответствующих числу факторов, если нет дублирующих переменных.

В целях удобства анализа моделей можно формальные критерии представить в графическом виде. Например, на рис. 2.2 приводятся распределение остаточной дисперсии (а), коэффициента множественной корреляции (б), критерия Фишера (в) и критерия Стьюдента (г) для $M=10$. Нетрудно видеть, что довольно резкие изменения указанных характеристик наблюдаются дважды: при переходе от одного фактора к двум и от шести факторов к семи. Но если в первом случае точность описания функций отклика явно недостаточна, то во вто-

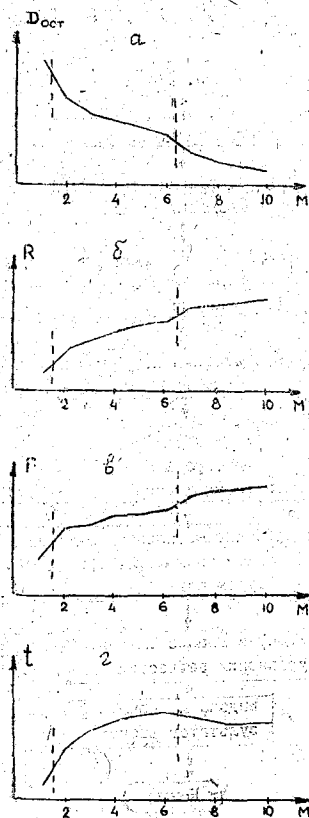


Рис. 2.2. Тестирование модели на основе пошаговой регрессии при числе факторов $M=10$

ром случае уже около 90 % дисперсии зависимой переменной описывается семью факторами.

Кроме того, целесообразно дополнительно провести анализ коэффициентов регрессии. Для этого можно, например, использовать следующее неравенство:

$$a_j / \sigma_{a_j} \geq 2.$$

Если в каком-либо уравнении хотя бы в отношении одного из коэффициентов регрессии данное неравенство не выполняется, то такое уравнение регрессии считается ненадежным. Если данное неравенство оказывается справедливым в отношении нескольких уравнений регрессии, то возможны два варианта выбора лучшего из них. Например, лучшим считается тот, для которого выполняется следующее условие:

$$\frac{1}{M} \sum^M a_j / \sigma_{a_j} \rightarrow \max,$$

в другом варианте для каждого из уравнений определяется минимальное значение $(a_j / \sigma_{a_j})_{\min}$. Лучшим считается то уравнение, для которого выполняется условие $(a_j / \sigma_{a_j})_{\min} \rightarrow \max$. Заметим, что анализ коэффициентов регрессии может не соответствовать результатам, полученным с помощью других критериев. Тем более, что распределение формальных критериев во многих случаях в отличие от рис. 2.2 может приводить к противоречивым выводам. Поэтому в общем задача выбора оптимальной модели регрессии оказывается неоднозначной, причем право окончательного выбора с учетом различных неформальных соображений остается прерогативой исследователя.

Что касается метода исключения переменных, его суть состоит в первоначальном построении полной модели, ранжировании факторов и исключении факторов из полной модели, начиная с наименее значимого. Так как по своей сущности он близок к методу включения переменных и в основном имеет те же самые трудности, то, очевидно, вряд ли целесообразно его здесь рассматривать.

Таким образом, нахождение оптимального вида регрессионной модели не имеет однозначного решения. В конечном счете выбор оптимальной модели определяется самим исследователем.

2.8. Применение МЛР в океанологических исследованиях

Рассмотрим проблемы использования методики МЛР на примере исследования процесса формирования среднегодовой солености Азовского моря [4]. Среди других океанологических факторов соленость воды Азовского моря занимает особое место, поскольку именно ее изменение в значительной степени влияет на всю экологическую систему моря. Для исследования формируется наибо-

лее общая модель, куда в качестве факторов, определяющих среднегодовую соленость S_i , включаются основные параметры процессов, создающих изменение солености: объемы суммарного водного баланса моря в текущий Q_i и предшествующий Q_{i-1} годы, объем атмосферных осадков O_i , объем испарения с акватории моря U_i , наконец, среднегодовая соленость в предшествующий год S_{i-1} . Искомая функциональная схема в упрощенном виде записывается как

$$S_i = S_i(Q_i, Q_{i-1}, O_i, U_i, S_{i-1}).$$

Для вычисления использовались ряды исследуемых переменных, накопленные за период с 1923 по 1977 гг. В предположении линейности взаимосвязей переменных были вычислены коэффициенты регрессии и составлено общее уравнение регрессии

$$S_i = 3,913 + 0,771S_{i-1} + 0,022Q_i - 0,015Q_{i-1} + 0,011O_i + 0,005U_i.$$

Коэффициент множественной корреляции для этой модели составил величину 0,821. Следовательно, на основании полученного уравнения можно описать $0,821^2 = 66\%$ дисперсии среднегодовой солености. Несмотря на адекватность данной модели по критерию Фишера $F^* = 41,95$, что превышает $F_{\text{табл}}$, авторы сочли целесообразным попытаться улучшить статистическую модель путем исключения из уравнения МЛР тех параметров, которые слабо связаны с соленостью и в то же время могут служить источником случайных ошибок.

Анализ коэффициентов парной корреляции среднегодовой солености с влияющими параметрами показал, что взаимосвязь солености с объемом осадков $r = 0,073$ и с объемом испарения $r = 0,322$ незначимо мала. То же можно сказать о коэффициентах регрессии при O_i, U_i . На основании этого анализа параметры O_i и U_i были исключены из списка независимых переменных, участвующих в построении уравнения регрессии вида

$$S_i = S_i(S_{i-1}, Q_i, Q_{i-1}).$$

В результате вычислений были получены коэффициенты регрессии, которые вошли в новое уравнение:

$$S_i = 3,928 + 0,771S_{i-1} + 0,022Q_i + 0,015Q_{i-1}.$$

Данное уравнение уже существенно лучше описывает изменчивость среднегодовой солености Азовского моря. На это указывает высокое значение коэффициента множественной корреляции $R_0 = 0,932$ и соответствующий ему процент описания дисперсии исходной переменной: $0,932^2 = 86\%$. Кроме того, в последней модели МЛР улучшилась по сравнению с предыдущей записью общая адекватность описания солености: вычисленный критерий Фишера составил уже $F^* = 72,840$. Расчеты показывают, что вторая модель процесса формирования среднегодовой солености позволяет не только описать изменчивость S_i в более достоверной форме, но

и сделать это значительно проще с привлечением меньшего числа влияющих параметров.

Рассмотренный пример показывает, что, используя единый подход регрессионного описания данных, автор может получить множество моделей исследуемого процесса в зависимости от формулировки физических предпосылок включения влияющих факторов. В океанологических исследованиях, когда набор влияющих факторов известен недостаточно ясно, применяют метод последовательного перебора независимых переменных при построении оптимального уравнения МЛР. Критерием учета каждой новой переменной принимают: а) увеличение коэффициента множественной корреляции для нового уравнения, б) статистическую значимость коэффициентов корреляции и регрессии данной переменной, в) надежность наблюдения или вычисления данной переменной в практических условиях, ее свободу от случайных ошибок. Последнее условие может быть не столь важным при формировании уравнения регрессии, сколько при его практическом использовании. Ценность регрессионного описания повышается тогда, когда с помощью хорошо известных, легко доступных и точно измеряемых параметров состояния океана удается через регрессионное уравнение оценить характеристику весьма важную, но напрямую не наблюдаемую. Примером такого фундаментального значения уравнения МЛР может служить проблема вычисления плотности морской воды. Кажущаяся простота используемой теоретической функциональной зависимости плотности воды от температуры, солености и давления (глубины) $\rho = \rho(T, S, P)$ на практике реализуется через множество различных эмпирических регрессионных уравнений состояния. Проблема здесь состоит в том, что простейшее линейное описание этой зависимости в виде

$$\rho = \rho_0 + a_1 T + a_2 S + a_3 P$$

оказывается весьма неточным, поскольку плотность является нелинейной функцией от параметров. Стремясь по возможности точнее описать изменчивость плотности в заданном диапазоне изменения T, S, P , различные авторы, используя регрессионный подход, строят функциональную зависимость в виде

$$\rho = \rho_0 + a_1 T + a_2 T^2 + \dots + a_k T^k + a_{k+1} S + a_{k+2} S^2 + \dots \\ \dots + a_{k+m} S^m + \dots$$

Причем эмпирически подбираются как коэффициенты регрессии, так и степени при температуре, солености и давлении. Накопленный опыт в этой области обобщен в книге П. А. Калашникова, где приводятся и обсуждаются основные известные эмпирические уравнения состояния, составленные Кнудсеном, О. И. Мамаевым, Коксом, Фофоновым, Миллером и др.

ГЛАВА 3. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

3.1. Постановка задачи

Метод главных компонент — это целая эпоха в эмпирико-статистических исследованиях гидрометеорологических данных. История применения МГК в гидрометеорологии в нашей стране связывается с именами таких видных известных ученых, как Н. А. Багров, Ю. В. Николаев, А. М. Обухов, вклад которых не только способствовал развитию и распространению МГК в гидрометеорологии, но и вообще открыл дорогу для использования всей совокупности методик ортогонального преобразования данных. В настоящее время в результате более чем тридцатилетнего пути своего развития МГК занял почетное место в гидрометеорологии и признается едва ли не классическим методом анализа данных. Причина столь высокой репутации метода состоит в его весьма высокой эффективности для решения широкого класса гидрометеорологических задач, которая сочетается с простотой вычислительного аппарата.

Прежде всего рассмотрим причины, вызывающие необходимость использования аппарата МГК. В настоящее время информационный куб данных для многих районов океана является уже огромным. Так, например, для Северной Атлантики накоплен архив длительных (более 30 лет) наблюдений за среднемесячной температурой поверхности океана в узлах пятиградусной географической сетки (всего 152 узла). Естественно, что описание этого массива с помощью традиционных методов одномерного анализа оказывается весьма громоздким и малоэффективным. Такие методы практически непригодны для одновременного описания пространственно-временных особенностей в матрице исходных данных.

В связи с этим возникает необходимость:

- существенного сжатия объемов статистической информации без заметных потерь в ее содержании;
- более наглядного представления (визуализации) исходных данных с целью облегчения физической интерпретации исследуемых пространственно-временных связей в новой более информативной системе признаков;
- более лаконичного описания и построения линейных моделей с целью выделения наиболее главных (крупномасштабных) особенностей в распределении новых признаков.

Итак, сущность МГК состоит в расчете неких вторичных признаков или функций, называемых главными компонентами, которые более эффективно описывают структурные закономерности, содержащиеся в матрице исходных данных. При этом вторичные признаки являются линейными комбинациями от исходных переменных.

В общем МГК представляет собой частный случай линейного ортогонального преобразования информации. Поэтому иногда данный метод называется как разложение по естественным функциям. Базисная система функций такого преобразования определяется корреляционной структурой исходной матрицы данных. Этим данный метод принципиально отличается от разложений, где базисная система функций задается априори (полиномы Чебышева, ряды Тейлора, Фурье и др.).

В результате разложения по естественным ортогональным функциям происходит разделение информации на линейно независимые составляющие, причем данный метод обладает максимальной по сравнению с любым другим разложением подобного типа скоростью сходимости. Или, другими словами, метод обеспечивает наилучшую в среднеквадратическом смысле аппроксимацию исходного поля фиксированным числом членов ряда разложения.

Таким образом, задача разложения сводится к тому, чтобы при сохранении общей точности описания исходных данных перейти к новым взаимно независимым (некоррелированным) переменным, которые описывают сходные тенденции в структуре данных и могут быть упорядочены по степени уменьшения их вкладов в общую изменчивость.

Заметим, что вся последовательность указанных задач решается технически просто, а именно, линейным преобразованием исходных данных или, другими словами, путем поэлементного умножения каждого признака на коэффициенты линейной связи данного признака с новой переменной, отвечающей заданным свойствам. При этом основная задача МГК состоит в том, чтобы наилучшим образом определить коэффициенты связи исходных переменных и главных компонент.

Наконец, следует подчеркнуть еще одну важную особенность МГК: он не требует никаких предварительных гипотез к исходным данным, и, в частности, не требует предположений о законе их распределения.

3.2. Математический аппарат МГК

Пусть имеется набор данных из N наблюдений за M переменными. Это равнозначно наличию M векторов наблюдений вида $X^T = \{\vec{X}_1, \vec{X}_2, \vec{X}_3, \dots, \vec{X}_N\}$, которые образуют матрицу данных X из N строк и M столбцов. Матрице данных в соответствие ставится матрица значений главных компонент F , состоящая также из N строк и M столбцов. Каждый столбец в матрице соответственно описывает изменение одной главной компоненты f_j , которая представляется в виде вектора $f_j^T = \{f_{1j}, f_{2j}, \dots, f_{Nj}\}$. Число исходных переменных, равное M , совпадает с числом главных

компонент, а длина каждой компоненты соответствует числу наблюдений. Тогда основное уравнение МГК имеет вид

$$X = FA^T. \quad (3.1)$$

Здесь матрица A представляет собой матрицу коэффициентов связи исходных наблюдений и главных компонент. Матрица A содержит M строк и M столбцов. Каждый столбец матрицы A содержит коэффициенты связи данной компоненты со всеми исходными признаками. Согласно правилам матричной алгебры матрица A в уравнении (3.1) транспонирована A^T (т. е. повернута на 90°).

Раскрывая содержание уравнения (3.1), можно записать формулировку МГК для j -й переменной ($j=1, 2, \dots, M$) в матрице данных:

$$X_j = \sum_{k=1}^M a_{jk} f_k = a_{j1} f_1 + a_{j2} f_2 + \dots + a_{jM} f_M, \quad (3.2)$$

где $f_1, \dots, f_M$ — некоррелированные (ортогональные) безразмерные коэффициенты разложения (главные компоненты); a_{jk} — вес k -й главной компоненты в j -й переменной.

Для нахождения главных компонент прежде всего необходимо определить коэффициенты связи в (3.2). Для этого воспользуемся, например, условием минимизации суммы квадратов ошибок разложения (3.2) при любом фиксированном значении k , т. е.

$$S_k = \sum_i \sum_j |x_{ij} - \sum_{k=1}^M a_{ik} f_{jk}|^2 = \min.$$

Используя традиционный аппарат метода наименьших квадратов и выполнив ряд преобразований, можно получить следующую систему уравнений:

$$\sum_{j=1}^M R_{ij} a_j = \lambda_i a_i. \quad (3.3)$$

Значения λ , при которых система (3.3) имеет решения, отличные от нулевого вектора, называются собственными числами матрицы R ; коэффициенты a_{ik} — собственные векторы матрицы R , соответствующие собственному числу λ_k . Матрица R — симметричная корреляционная (ковариационная) матрица исходных данных размерностью $M \times M$. Напомним, что в ковариационной матрице ее элементами являются ковариации (отклонения от нормы), а диагональными элементами служат дисперсии.

Полагая, что R является корреляционной матрицей, запишем систему (3.3) в виде следующей системы линейных однородных уравнений:

$$\begin{aligned}
 (r_{11} - \lambda) a_1 + r_{12} a_2 + \dots + r_{1M} a_M &= 0, \\
 r_{21} a_1 + (r_{22} - \lambda) a_2 + \dots + r_{2M} a_M &= 0, \\
 \dots & \\
 r_{M1} a_1 + r_{M2} a_2 + \dots + (r_{MM} - \lambda) a_M &= 0
 \end{aligned}
 \tag{3.4}$$

или в матричной записи

$$A \| R - \lambda I \| = 0,$$

где I — единичная матрица одного порядка с A .

Как известно, необходимым и достаточным условием существования нетривиального решения данной системы является равенство нулю ее детерминанта, т. е.

$$|R - \lambda I| = 0. \tag{3.5}$$

Это уравнение получило название характеристического или векового. В поэлементном виде оно записывается как

$$\begin{vmatrix}
 r_{11} - \lambda & r_{12} & \dots & r_{1M} \\
 r_{21} & r_{22} - \lambda & \dots & r_{2M} \\
 \dots & \dots & \dots & \dots \\
 r_{M1} & r_{M2} & \dots & r_{MM} - \lambda
 \end{vmatrix} = 0. \tag{3.6}$$

Раскрывая определитель, получаем уравнение степени M относительно неизвестного λ . В общем случае оно имеет M положительных корней, называемых собственными или характеристическими числами матрицы R , среди которых нет кратных. Расположим их в убывающем порядке ($\lambda_1 > \lambda_2 > \dots > \lambda_M$) и будем подставлять поочередно в систему (3.4). В результате получим M решений

$$\left. \begin{aligned}
 a_{11} \ a_{12} \ \dots \ a_{1M} \quad (\text{для } \lambda_1) \\
 a_{21} \ a_{22} \ \dots \ a_{2M} \quad (\gg \lambda_2) \\
 \dots \\
 a_{M1} \ a_{M2} \ \dots \ a_{MM} \quad (\gg \lambda_M)
 \end{aligned} \right\}.$$

Каждое из этих решений называется собственным вектором матрицы R .

Заметим, что для каждого j -го собственного вектора разложения должно выполняться следующее условие:

$$a_{1j}^2 + a_{2j}^2 + \dots + a_{Mj}^2 = \sum_{k=1}^M a_{kj}^2 = 1.$$

Одновременно с этим сумма собственных чисел разложения должна равняться порядку матрицы, если матрица корреляционная ($\sum_k \lambda_k = M$), или сумме дисперсий исходных рядов, если матрица ковариационная ($\sum_k \lambda_k = \sum_j D_j$).

После нахождения собственных чисел и собственных векторов определение коэффициентов разложения (главных компонент) уже не представляет каких-либо затруднений. Для этой цели используется обратное преобразование уравнения (3.1), т. е.

$$F = AX.$$

Отсюда первая главная компонента определяется как

$$f_{i1} = \sum_{j=1}^M a_{1j} x_{ij} = a_{11}x_{i1} + a_{12}x_{i2} + \dots + a_{1M}x_{iM},$$

где $i=1, 2, \dots, N$. Аналогичным образом рассчитываются оставшиеся $M-1$ главных компонент.

Итак, смысл разложения по естественным составляющим состоит в том, что из большого числа экспериментальных данных прежде всего отбирается первое собственное число и соответствующие ему первый собственный вектор и первая компонента разложения. Именно они описывают наиболее крупномасштабные (главные) особенности рассматриваемого исходного поля, причем их вклад в общую (суммарную) изменчивость поля определяется величиной собственного числа. С увеличением порядкового номера собственных чисел собственные векторы и соответствующие им компоненты разложения будут описывать все менее существенные (более второстепенные) особенности исходных данных.

Как уже указывалось, данный метод наилучшим образом среди подобных преобразований линейного типа исчерпывает дисперсию исходных переменных, т. е. обладает наиболее высокой скоростью сходимости. Это часто позволяет сосредоточивать информацию об исходном поле в первых коэффициентах разложения. Именно поэтому они получили название главных компонент, а сам метод — метод главных компонент.

Скорость сходимости разложения описывается следующим выражением:

$$d = \sum_k^m \lambda_k / \sum_k^M \lambda_k,$$

где M — порядок матрицы, соответствующий общему числу членов разложения; m — часть из них. Следовательно, величина d показывает отношение части дисперсии, описываемой m членами разложения к общей дисперсии рассматриваемого поля. Естественно, что относительный вклад k -той компоненты в общую дисперсию оценивается как

$$d_k = \lambda_k / \sum_{k=1}^M \lambda_k.$$

3.3. Геометрическая интерпретация МГК

Рассмотрим этот вопрос на примере анализа двух векторов наблюдений $\vec{X}_1$ и $\vec{X}_2$. Пусть это будут данные синхронных наблюдений температуры воды в двух ($M=2$) различных регионах океана за период N лет. Если изменчивость температуры в этих районах взаимосвязана, то графически она может быть представлена облаком наблюдений в форме эллипса (рис. 3.1).

Основная изменчивость температуры воды, соответствующая главной диагонали эллипса рассеяния наблюдений может быть оценена по формуле

$$\sigma^2 = \sigma^2(\vec{X}_1) + \sigma^2(\vec{X}_2).$$

Задача вычислительных операций МГК состоит в том, чтобы определить направление главной диагонали эллипса. При этом искомые главные компоненты f_1 и f_2 можно рассматривать как оси новой ортогональной системы координат, ориентация которой в пространстве однозначно задается самой структурой наблюдений, т. е. формой облака.

Процесс поиска главных компонент следует интерпретировать как процедуру поворота координатных осей. Переориентация осей производится по следующим правилам: первая ось (главная компонента f_1) направляется вдоль оси максимального рассеяния наблюдений, которое определяется максимальными коэффициентами корреляции наблюдений. Вторая ось (главная компонента f_2) обязательно перпендикулярна первой и при этом указывает направление максимального остаточного рассеяния наблюдений. Остаточное рассеяние получается в результате вычитания дисперсии

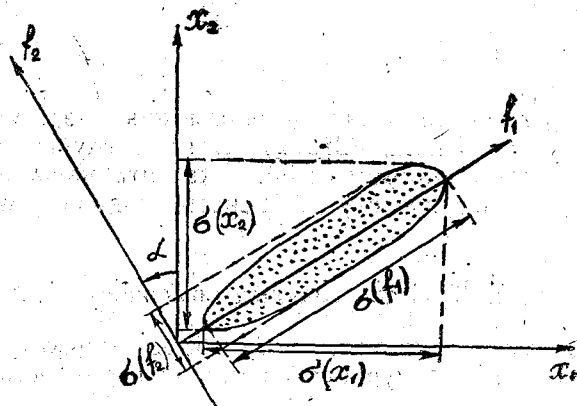


Рис. 3.1. Геометрическая интерпретация применения МГК: x_1, x_2 — исходные признаки; $\sigma(x_1), \sigma(x_2)$ — их стандартные отклонения; f_1, f_2 — главные компоненты; $\sigma(f_1), \sigma(f_2)$ — их стандартные отклонения

наблюдений, описываемых первой главной компонентой $\sigma(f_1)$ из общей дисперсии наблюдений.

Таким образом, переход из исходной системы координат X_1, X_2 в систему координат главных компонент f_1, f_2 можно осуществить простым ортогональным поворотом координатных осей на угол α против часовой стрелки. Дисперсия наблюдений в проекции на ось первой компоненты f_1 при этом достигнет максимума среди дисперсий наблюдений в проекции на другие оси компонент. Форма облака наблюдений и взаиморасположение отдельных точек наблюдений не изменились при переходе в новую систему отсчета. Следовательно, не изменилась и общая дисперсия наблюдений в координатах главных компонент:

$$\sigma^2 = \sigma^2(f_1) + \sigma^2(f_2).$$

При этом всегда выполняется неравенство

$$\sigma^2(f_1) \geq \sigma^2(f_2).$$

Данное неравенство демонстрирует процесс сжатия информации в терминах главных компонент. Дисперсия, более или менее равномерно распределяемая изначально между исходными переменными $\vec{X}_1$ и $\vec{X}_2$, в результате МГК как бы «перекачивается» в первые главные компоненты.

Рассматриваемая геометрическая интерпретация позволяет осмыслить роль коэффициентов a_{ij} в уравнении (3.2). С точки зрения тригонометрии эти коэффициенты перехода от исходных признаков X к главным компонентам f являются косинусами и синусами углов поворота α . Отсюда можно записать матрицу собственных векторов A в виде

$$A = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}.$$

Аналогичная геометрическая интерпретация имеет место и для многомерного пространства ($M > 2$). В этом случае вторая ось ориентируется таким образом, чтобы охватить наибольшую дисперсию из оставшейся после учета первой главной компоненты и так далее.

3.4. Свойства главных компонент

Свойства главных компонент могут быть сформулированы на основе анализа основного уравнения метода и его геометрической интерпретации:

1. ГК между собой линейно независимы (ортогональны), следовательно их попарная корреляция равна 0:

$$r(f_j, f_k) = 0 \quad \text{при } j \neq k.$$

2. Компоненты описывают изменчивость данных таким образом, что первая ГК описывает максимальную дисперсию облака наблюдений, вторая ГК ортогональна первой и описывает максимальную остаточную дисперсию, третья ГК ортогональна двум предыдущим ГК и описывает максимальную остаточную дисперсию и т. д.:

$$\sigma^2(f_1) \geq \sigma^2(f_2) \geq \dots \geq \sigma^2(f_M).$$

3. Применение МГК позволяет оптимальным образом среди всех других линейных преобразований описать дисперсию исследуемых переменных. Другими словами, МГК наилучшим образом «сжимает» информацию, содержащуюся в наблюдениях.

4. Использование ГК позволяет описать исходную информацию с минимальным искажением геометрической структуры облака наблюдений в пространстве ГК.

5. Вклад ГК в описание общей дисперсии переменных пропорционален значению соответствующего ей собственного числа корреляционной матрицы исходных признаков:

$$\sigma^2(f_j) \approx \lambda_j.$$

3.5. Проблемы применения МГК в океанологии

Наибольшую известность МГК завоевал в результате эффективного разложения пространственно-временных матриц данных вида $T(t, x)$. В подобных исследованиях матрица данных раскладывалась на ортогональные составляющие с целью выявления ведущих масштабов колебаний гидрометеорологических полей в пространстве и времени. Наглядным примером такого подхода может служить работа Ю. В. Николаева и Э. И. Саруханяна по исследованию пространственно-временной структуры колебаний полей ледовитости арктических морей СССР. Авторы стремились увязать масштабы типичных пространственных колебаний ледовитости со свойствами временного хода. Необходимость применения МГК возникла еще и потому, что требовалось связать воедино набор данных о локальных колебаниях ледовитости в точках наблюдения с обобщенной характеристикой ледовитости всех арктических морей как единого целого.

Матрица исходных данных составлялась из ежегодных значений ледовитости за август — сентябрь в 7 районах: в юго-западной и северо-восточной частях Карского моря (на рис. 3.2 они обозначены как ЮЗК и СВК соответственно), в западной и восточной частях моря Лаптевых (ЗЛ и ВЛ), восточной и западной частях Восточно-Сибирского моря (ВВС и ЗВС), в Чукотском море (Ч). Количество подобных ежегодных данных составило 36 значений за период с 1935 по 1970 гг.

В процессе вычислений матрица исходных данных размером 36 строк и 7 столбцов была преобразована к матрице корреляций

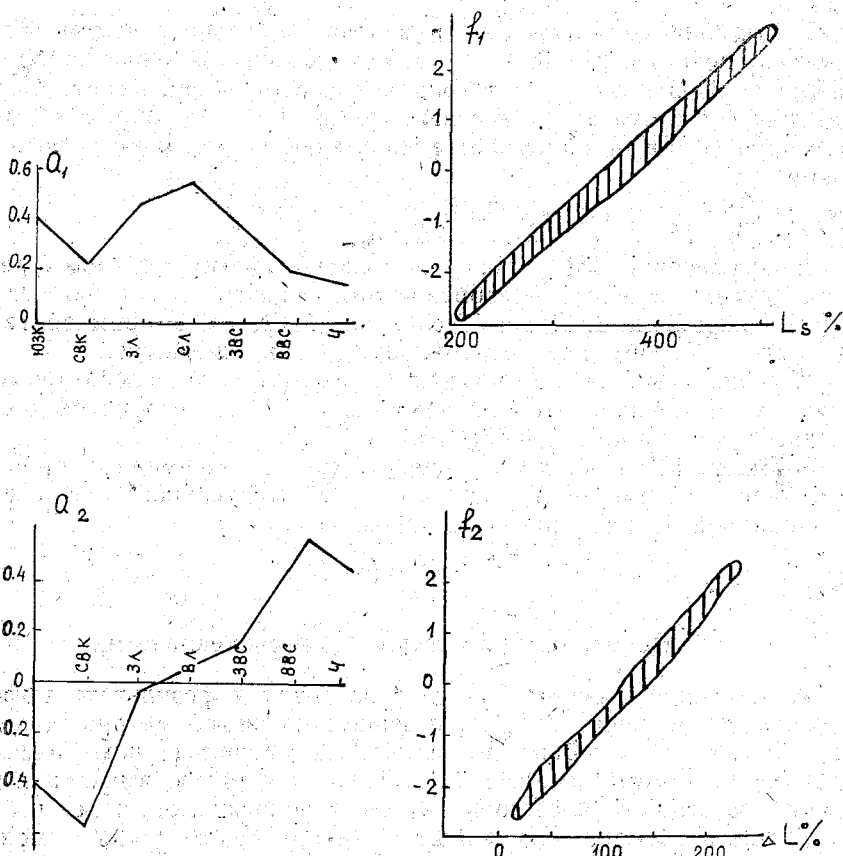


Рис. 3.2. Анализ ледовитости арктических морей методом главных компонент: *а* — распределение первого и второго собственного вектора в пространстве; *б* — взаимосвязь главных компонент с общей ледовитостью L_s и разностью ледовитости восточной и западной Арктики ΔL .

между районами наблюдений размером 7×7 , для которой были найдены 7 собственных чисел $\lambda_1, \lambda_2, \dots, \lambda_m$ и 7 собственных векторов $a_1, a_2, \dots, a_7$. Характеристики разложения корреляционной матрицы имеют следующий вид (значения собственных чисел и соответствующих вкладов округлены):

Характеристики	Номер						
	1	2	3	4	5	6	7
λ_i	2.4	1.8	1.0	0.7	0.6	0.3	0.2
d_i %	35	26	14	9	8	5	3
$\sum d_i$ %	35	61	75	84	92	97	100

Как следует из приводимой таблицы, более 50 % от общей изменчивости ледовых условий описывается первыми двумя собственными векторами и соответствующими им главными компонентами. Каждый из выделенных векторов имеет характерное пространственное распределение. Из рис. 3.2 видно, что первый собственный вектор характеризуется высокой положительной связью с колебаниями ледовитости практически во всех исследуемых районах. Причем наиболее отчетливо подобные колебания прослеживаются в море Лаптевых и в западной части Восточно-Сибирского моря, где значения a_1 максимальны. Поэтому можно полагать, что первая главная компонента f_1 является аналогом колебаний суммарной ледовитости арктических морей в целом, а дисперсия подобных колебаний составляет величину 35 % от общей изменчивости исследуемого процесса.

Координаты второго собственного вектора a_2 отличаются положительными значениями для районов восточной Арктики и отрицательными — для районов западной Арктики. Второй собственный вектор имеет вид волны, что позволяет интерпретировать его как ту долю изменчивости, которая свойственна противофазным колебаниям ледовитости между районами восточной и западной Арктики. Временной ход второй главной компоненты f_2 при этом будет отражать в каждый конкретный момент времени аномалию оппозиции развития ледового покрова между восточной и западной Арктикой.

Для подтверждения выдвинутых физических объяснений природы главных компонент на рис. 3.2 приведены облака рассеяния точек наблюдения в координатах: 1-я ГК — суммарная ледовитость, 2-я ГК — разность ледовитости между востоком и западом. Характерная вытянутая форма полученных облаков рассеяния подтверждает приводимую интерпретацию членов разложения.

Для обобщения особенностей временного хода главных компонент на рис. 3.3 представлены рассчитанные автокорреляционные функции временного хода первой и второй компонент. Вид полученных графиков автокорреляционных функций убеждает в том, что обе ГК имеют выраженные периоды циклических колебаний во времени: 7—8 лет для первой компоненты, 2—3 года для второй.

Суммируя полученные результаты разложения поля ледовитости по главным компонентам, можно сделать следующие выводы:

выявлены общие для всех арктических морей синхронные колебания ледовитости с периодом около 7—8 лет, на долю которых приходится 35 % суммарной изменчивости ледовитого покрова за период с 1935 по 1970 гг.

выявлена оппозиция (противофазность) колебаний ледовитости в западной и восточной Арктике с периодом около 2—3 года, на долю которой приходится 26 % суммарной изменчивости.

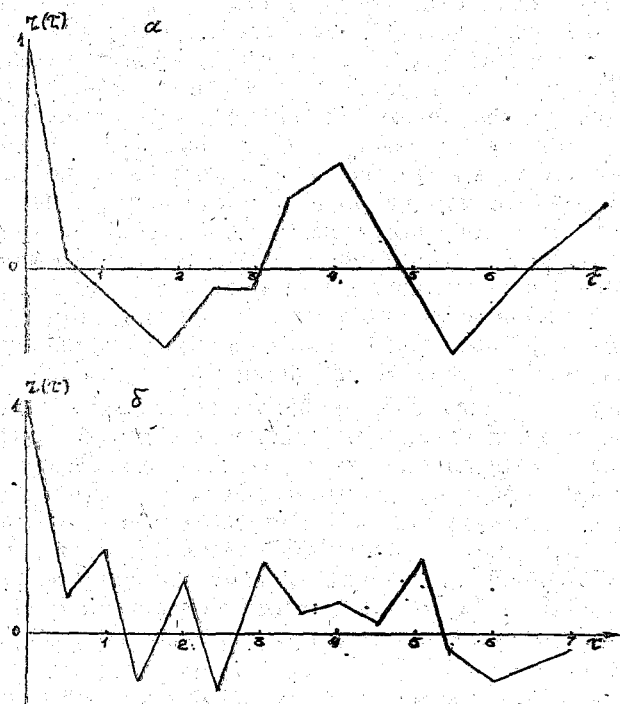


Рис. 3.3. Автокорреляционные функции первой (а) и второй (б) главных компонент разложения ледовитости арктических морей

Приводимый пример наглядно демонстрирует эффективность применения метода главных компонент в исследованиях гидрометеорологических процессов.

Практическое применение МГК в океанологии наталкивается на ряд трудностей. Попробуем обозначить хотя бы основные из них.

1. Проблема обеспечения реальности корреляций исходных признаков между собой.

Суть вопроса состоит в том, что содержательный анализ результатов МГК возможен при обязательном соответствии вычисленных парных коэффициентов корреляции уровню реальных связей между исследуемыми признаками. Для удовлетворения этого условия надо, чтобы а) длина выборки была достаточной для статистически достоверной оценки корреляции, б) наблюдалась однородность (неизменность) взаимосвязи по всей длине выборки, в) выборка должна быть стационарна, г) связь между переменными должна быть линейной. Проверка перечисленных требований должна быть выполнена до начала применения МГК согласно правилам анализа рядов наблюдений.

2. Проблема выбора признаков.

Выбор признаков определяет физическую интерпретацию разложения по главным компонентам. Признаки подбирают, как правило, исходя из априорных физических соображений о структуре исследуемых процессов и о возможной их взаимосвязи. Количество признаков определяется самим исследователем и ограничивается лишь двумя соображениями: необходимостью использования параметров для описания сложного процесса и мощностью ЭВМ. Современные исследования проводятся, как правило, с использованием от 5 до 20 признаков. Мощности ЭВМ позволяют анализировать корреляционные матрицы для 50 и даже 100 признаков. Физическая размерность признаков при проведении вычислений по корреляционной матрице не играет роли. Однако если используется разложение по ненормированной на дисперсию ковариационной матрице, то необходимо подбирать признаки одинаковой размерности.

3. Проблема сходимости разложения.

Под сходимостью понимают скорость описания дисперсий исследуемых переменных на первых главных компонентах. Чем меньше главных компонент необходимо для описания подавляющей части дисперсии (80, 90%), тем выше сходимость разложения, а следовательно, выше эффективность применения метода, надежность и точность результатов вычислений. Эту особенность легко пояснить на геометрическом примере. Существенная вытянутость эллипса рассеяния исходных наблюдений в координатах исходных признаков предопределяет возможность точного поворота осей главных компонент по направлению к максимальной дисперсии. И наоборот, если облако разброса наблюдений близко совпадает по форме со сферой, то практически невозможно найти ось наибольшей дисперсии. Сферичность облака наблюдений предполагает близость к нулю коэффициентов парной корреляции между признаками. Поэтому можно сделать вывод о нецелесообразности привлечения к анализу методом главных компонент признаков, не связанных со всеми остальными параметрами, входящими в матрицу исходных данных.

4. Проблема интерпретации главных компонент.

Эта проблема встает в случае, если решается полная задача поиска главных компонент и делается попытка объяснения природы выявленных пространственно-временных взаимосвязей ортогональных составляющих. Следует помнить, что МГК по своему содержанию является формальным аппаратом поиска ортогональных составляющих, поэтому физическая интерпретация главных компонент возможна не всегда. Чаще всего удается физически адекватно интерпретировать ГК в том случае, когда строго соблюдены условия применения МГК и решены перечисленные выше проблемы. На практике исследователи стремятся интерпретировать не все компоненты, а лишь несколько первых, на которых сосредоточена основная изменчивость исходных данных. Доля

описываемой ими дисперсии составляет, как правило, величину в пределах 70—90 % и определяется соображениями о точности исходных данных и точности аппроксимации исследуемых процессов моделью главных компонент.

ГЛАВА 4. МЕТОД ФАКТОРНОГО АНАЛИЗА

4.1. Постановка задачи и основное уравнение

Метод факторного анализа (МФА) представляется несколько более сложной процедурой по сравнению с ранее рассмотренным МГК, но в то же время и более эффективной при решении задач исследования структуры взаимосвязи признаков. Изучая такую сложную многокомпонентную систему, какой является Мировой океан, исследователь неизбежно встречается с ситуацией, при которой изменчивость большой совокупности параметров определяется существенно меньшим количеством скрытых от непосредственного измерения факторов.

Все содержание МФА направлено как раз на то, чтобы с максимальной точностью и достоверностью проверить гипотезу о наличии малого количества влияющих факторов и добиться статистического описания этих факторов. Принимаемая в факторном анализе схема формирования изменчивости наблюдаемых переменных X_j под влиянием скрытых общих для нескольких переменных факторов f_k и индивидуальных специфических ошибок наблюдения e_j представлена на рис. 4.1. Влияние факторов на наблюдаемые переменные оценивается на основе модельных коэффициентов линейной связи a_{jk} , которые показывают степень связанности k -го общего фактора и j -й переменной. Запишем на основании приведенных рассуждений алгебраическое выражение модели МФА

$$X = FA^* + E. \quad (4.1)$$

Здесь X — матрица исходных данных, состоящая из N наблюдений (строк) за M переменными (столбцов); F — матрица значений общих факторов, состоящая из N значений k факторов; A — матрица коэффициентов связи общих факторов и исходных переменных, составленная из N строк и M столбцов; E — матрица остатков или специфических (характерных) факторов, содержащая N наблюдений за M переменными. Чтобы получить единственное решение представленного уравнения, необходимо ввести некоторые дополнительные условия на вид матриц A , F и E . При формулировке этих условий традиционно предполагается, что из бесконечного количества математических моделей процессов, протекающих в природе, нам необходимо получить максимально простую модель, которая одновременно вписывалась бы в корреляционную структуру матрицы исходных данных.

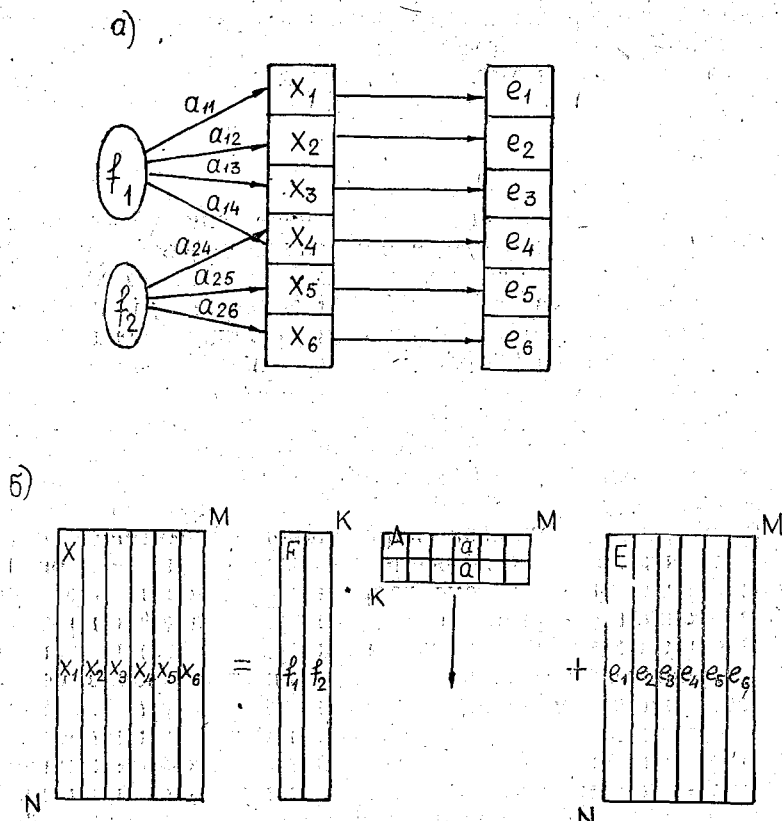


Рис. 4.1. Схема взаимосвязей переменных в модели факторного анализа

Накладываемые условия можно сформулировать следующим образом:

— общие факторы $f_1, \dots, f_k$ должны быть взаимонезависимы (некоррелированы)

$$r(f_i, f_j), i \neq j;$$

— общие факторы $f_1, \dots, f_k$ должны быть нормированы (безразмерны) и приведены к единичной длине $\sigma^2(f_j) = 1$;

— общие факторы не должны быть связаны с ошибками

$$r(f_i, e_j) = 0 \quad \text{при любых } i \neq j;$$

— ошибки не должны быть коррелированы между собой

$$r(e_i, e_j) = 0, \quad i \neq j;$$

— количество общих факторов не должно превышать половину числа наблюдаемых переменных $k \leq M/2$;

— общий фактор должен описывать большую часть изменчивости как минимум двух переменных.

Основное уравнение МФА в поэлементном виде можно записать как

$$x_{ij} = \sum_{p=1}^k f_{ip} a_{pj} + e_{ij}, \quad (4.2)$$

где x_{ij} — значение i -го наблюдения j -й переменной; f_{ip} — значения k общих факторов для i -го наблюдения; a_j — веса (нагрузки) k общих факторов на j -ю переменную; e_{ij} — ошибка наблюдения или специфический фактор для i -го наблюдения j -й переменной. Общие факторы здесь обуславливают корреляцию между переменными X_j и представляют, так же, как и в анализе главных компонент, скрытые (ненаблюдаемые) случайные переменные. Характерные факторы позволяют учитывать оставшуюся дисперсию, связанную либо с ошибками наблюдений, либо с существованием особых (специфических) для каждой переменной X_j факторов.

Одним из ограничений МФА является неоднозначность решения. В результате анализа корреляционной матрицы определяется количество общих факторов и нагрузки на них без привязки к определенной системе координат. Следовательно, любым преобразованием (например, ортогональным вращением или косинусным преобразованием) системы координат полученного факторного пространства можно получить сколько угодно решений, одинаково хорошо описывающих данные. Поэтому МФА строится обычно в виде двух процедур: 1) оценивание размерности факторного пространства и получение оценок факторных нагрузок, адекватно описывающих данные; 2) ортогональное вращение системы факторов с тем, чтобы полученная в итоге система факторов, так же хорошо описывающая данные наблюдений, оказалась более интерпретируемой с точки зрения исследователя.

4.2. Геометрическая интерпретация МФА

Геометрическая интерпретация факторного анализа несколько более сложная по сравнению с МГК. Дело в том, что в МФА происходит процесс проектирования облака исходных наблюдений из M -мерного пространства признаков в k -мерное пространство общих факторов. Учитывая, что $k < M$, мы неизбежно приходим к выводу о появлении при этой процедуре некоторых искажений облака наблюдений. Эти искажения или ошибки содержатся в пространстве специфических или характерных факторов, мерность которого совпадает с количеством переменных. Данная процедура МФА первоначально эквивалентна процедуре вращения исходных координат, которая рассматривалась выше как геометрическая интерпретация МГК. Однако здесь после вращения осей по направлению максимальной изменчивости признаков происходит удаление тех осей, изменчивость наблюдений вдоль которых не вы-

ходит за рамки принятого уровня ошибок. Другими словами, сокращаются те оси рассеяния, в проекции на которые исходное облако наблюдений представляется практически сферой. Так, например, из 20-мерного пространства исходных переменных можно перейти к 3-мерному пространству общих факторов (рис. 4.2, а).

После сокращения осей проводится еще одно дополнительное вращение облака наблюдений уже в пространстве k общих факторов. Цель этого вращения — максимально точно направить оси общих факторов вдоль направлений наибольшего рассеяния наблюдений. Данная операция связана как раз с поиском простой структуры факторного решения, при которой каждый общий фактор описывает максимально для него возможный уровень изменчивости наблюдений (рис. 4.2, б).

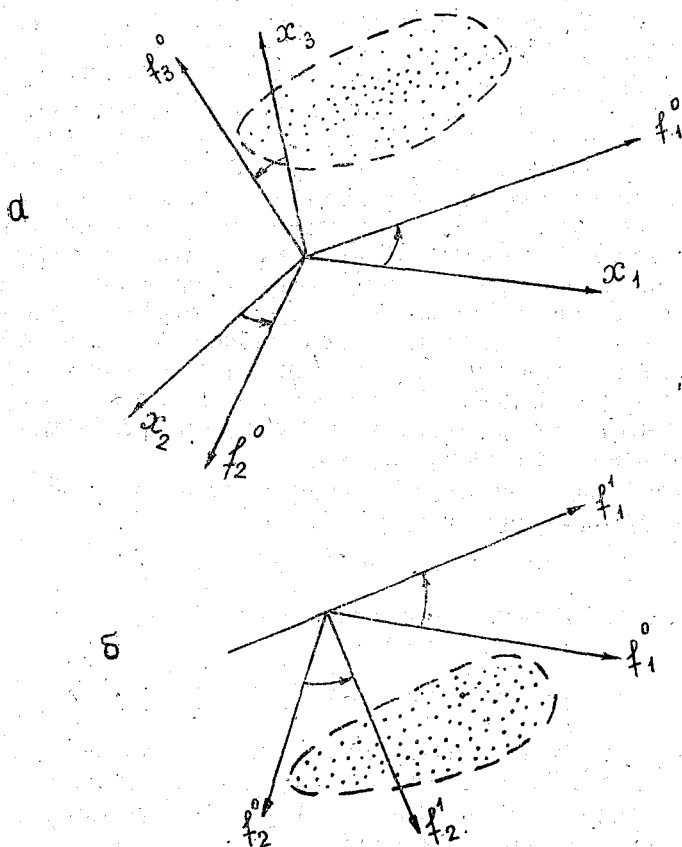


Рис. 4.2. Геометрическая интерпретация МФА: а — переход от осей исходных признаков к осям начальных факторов; б — переход к общим факторам

4.3. Поиск факторного решения

Процесс поиска факторного решения представляет собой совокупность частных процедур матричного анализа. Здесь будет рассмотрен только один из возможных вариантов факторного решения, называемый методом главных факторов, который достаточно прост и нагляден. Заметим также, что практически каждый из рассматриваемых вычислительных этапов может иметь иной способ решения, выбор которого определяется навыком исследователя и общей направленностью исследования.

Итак, процесс решения начинается уже с составления исходной матрицы данных. В матрицу данных включаются те переменные, которые, по мнению исследователя, априорно позволяют решить проблему поиска скрытых факторов. Это могут быть как реально наблюдаемые характеристики состояния океана, так и их различные производные (градиенты по направлению, потоки веществ и т. п.). При составлении матрицы данных важно помнить, что количество наблюдений должно превышать количество переменных хотя бы в 3—5 раз; это необходимо для получения устойчивых оценок факторного решения.

Далее производится вычисление средних и стандартных отклонений переменных, а затем матрица данных стандартизуется. По стандартизованным значениям вычисляются коэффициенты парной корреляции между переменными, из которых составляется матрица корреляций R . Матрица R — симметричная, квадратная матрица, состоящая из M строк и M столбцов. Вся вычислительная процедура МФА строится на основе анализа полученной корреляционной матрицы. Поэтому естественно желать, чтобы полученные оценки коэффициентов корреляции отражали реальные взаимосвязи в исходных наблюдениях.

Для удобства вычислений основное уравнение МФА записывается в корреляционной форме

$$R = AA^T + U^2. \quad (4.3)$$

В поэлементной записи это выражение примет вид

$$r_{ii} = \sum_{p=1}^k a_{ip}a_{ip} + U_{ii}^2 = h_{ii}^2 + U_{ii}^2. \quad (4.4)$$

Здесь h_{ii}^2 называется общностью и характеризует часть дисперсии переменных, связанную с действием общих факторов, т. е. с корреляцией между переменными. Величина U_{ii}^2 называется дисперсией характеристики и описывает часть дисперсии, связанную с действием специфических факторов, т. е. с корреляцией остатков. Заметим, что матрица U^2 является квадратной порядка M , на главной диагонали которой находятся нормированные дисперсии остатков

$$U^2 = \begin{vmatrix} U_{11}^2 & 0 & \dots & 0 \\ 0 & U_{22}^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & U_{MM}^2 \end{vmatrix}.$$

Отсюда следует, что дисперсия переменной r_{ii} содержит систематическую долю дисперсии h_{ii}^2 , которая интерпретируется общими факторами, и случайную долю U_{ii}^2 , которая понимается как остаток и не содержит физического смысла. В связи с этим необходимо изъять из дальнейших вычислений величину U_{ii}^2 , которая ухудшает качество моделей. Процедура замены диагональных элементов матрицы корреляций на значения общностей $h_{ii}^2 = r_{ii}^2 - U_{ii}^2$ называется редуцированием корреляционной матрицы, т. е.

$$R^h = R - U^2 = \begin{vmatrix} r_{11} - U_{11}^2 & r_{12} & \dots & r_{1M} \\ r_{21} & r_{22} - U_{22}^2 & \dots & r_{2M} \\ \dots & \dots & \dots & \dots \\ r_{M1} & r_{M2} & \dots & r_{MM} - U_{MM}^2 \end{vmatrix}.$$

Существуют различные варианты оценки величин общностей. В частности, наиболее простой прием заключается в замене диагональных элементов значениями максимальных коэффициентов корреляции в соответствующей строке. Другой способ состоит в вычислении для каждой переменной X_j квадрата коэффициента множественной корреляции, показывающего, насколько точно данная переменная описывается через остальные переменные. Очевидно, в последнем случае, исходя из свойств коэффициента множественной корреляции, значения общностей окажутся несколько выше. Следует также иметь в виду, что чем ближе оценки общностей к единице, тем выше скорость сходимости разложения по редуцированной корреляционной матрице.

Далее производится определение собственных чисел и собственных векторов редуцированной корреляционной матрицы R^h . С этой целью ищется решение характеристического уравнения вида

$$|R^h - \lambda I| = 0,$$

что дает оценку собственных чисел $\lambda^n = \{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_M\}$, решаются системы линейных однородных уравнений вида

$$A_0 \|R^h - \lambda I\| = 0,$$

что позволяет оценить собственные векторы A_0 .

После вычисления собственных векторов, которые принимаются в качестве начальных приближений соответствующих факторных нагрузок, переходят к оценке количества общих факторов;

достаточных для воспроизведения матрицы корреляций R . Для этого вычисляют матричное произведение

$$A_0 A_0^T = R^+,$$

которое дает оценку матрицы восстановленных корреляций, и вычисляют уровень различий наблюдаемых и восстановленных корреляций

$$R^h - R^+ = D.$$

Две эти операции повторяются многократно, начиная с одного собственного вектора в матрице $A_0^{(1)}$, затем с каждым разом увеличивая количество векторов на единицу: $A_0^{(2)}, A_0^{(3)}, \dots, A_0^{(k)}$. Как только величина остатков D удовлетворяет принятому уровню значимости модели, исследователь прекращает добавление собственных векторов и принимает количество k общих факторов модели достаточным. Полученные собственные векторы являются коэффициентами перехода из M -мерного пространства исходных переменных в k -мерное пространство общих факторов с дисперсией ошибки порядка

$$d = \sum_{i=k+1}^M \lambda_i \cdot \sum_{j=1}^k \lambda_j.$$

Далее необходимо оптимизировать решение МФА уже в сокращенном факторном пространстве. Этот этап называется вращением или поиском простой структуры решения. Суть этапа состоит в том, чтобы провести операцию поворота осей общих факторов и перейти от матрицы A_0 к матрице A так, чтобы по возможности все имеющиеся факторные нагрузки были максимально приближены либо к ± 1 , либо к 0. Такое условие зачастую позволяет избежать половинчатых решений и весьма четко очертить сферу влияния факторов на переменные.

На практике процедура вращения сводится к многократному итеративному умножению вида

$$A = A_0 T,$$

где T — матрица поворота размерностью M строк и M столбцов с взаимонезависимыми и нормированными столбцами. Вращение осей координат приостанавливается, когда достигнут максимум дисперсии факторных нагрузок, приходящихся на каждый (l) признак:

$$S^2 = \frac{1}{k} \sum_{l=1}^k (a_{ll})^2 - \frac{1}{k} \left(\sum_{l=1}^k a_{ll} \right)^2 \rightarrow \max. \quad (4.5)$$

Данное выражение получило название варимакс-критерия. Когда он достигает максимума, то считается, что получена простая структура решения и, следовательно, факторы должны быть наиболее интерпретируемыми.

Полученные значения коэффициентов a_{ij} принимаются в качестве искоемых факторных нагрузок.

Последним этапом факторного анализа традиционно принимается процедура вычисления значений общих факторов и оценивания матрицы остатков. Оценка значений общих факторов f , как правило, осуществляется на основе классической схемы множественной линейной регрессии. Матричное уравнение для этой операции имеет вид

$$F = X(AA^T)^{-1}A^T = X(R^+)^{-1}A^T.$$

Здесь необходимо обратить внимание на оператор $(AA^T)^{-1}$, который является обратной матрицей восстановленных корреляций. Дело в том, что точность вычислений при обращении матрицы R^+ в данном случае определяет надежность оценивания значений общих факторов.

Вычисление остатков проводится в результате оценивания матричной разности

$$E = X - FA^T.$$

Рассмотренная схема МФА получила название метода главных факторов, который довольно широко применяется на практике. К сегодняшнему дню разработан целый ряд процедур оценивания факторных нагрузок и вращения общих факторов, с которыми можно познакомиться в специальной литературе.

4.4. Применение факторного анализа в океанологических исследованиях

МФА не получил в гидрометеорологических расчетах такого широкого распространения, как МГК. При некорректной постановке задачи и невыполнении жестких требований, предъявляемых к исходной информации, вряд ли следует ожидать, что использование МФА принесет какие-либо положительные результаты. Кроме того, МФА имеет ряд различных модификаций, применение которых к одному и тому же набору данных может приводить к расхождению в получаемых оценках.

Одним из основных направлений практического использования МФА является изучение пространственно-временной структуры гидрометеорологических полей. При этом наиболее эффективным данный метод оказывается в анализе таких гидрометеорологических процессов, которые обладают существенно неоднородной структурой.

В качестве примера обратимся к результатам анализа изменчивости полей аномалий среднемесячной температуры поверхности океана (АТПО) в Норвежской энергоактивной зоне. Акватория Норвежского моря была покрыта равномерной сеткой с шагом 60 миль, в узлах которой ($M=26$) снимались среднемесячные АТПО за 1978—1986 гг. В результате была получена матрица

размером 108×26 . В соответствии с МФА данная матрица была преобразована в корреляционную матрицу размером 26×26 , отражающую тесноту линейной статистической связи между отдельными точками акватории.

Значения полученных коэффициентов корреляции оказались в пределах от 0.92 до —0.10. При этом, как и следовало ожидать, наиболее высокая связь наблюдается в основном в близлежащих точках пространства. Исключение составляют лишь районы с резкими градиентами температуры, обусловленные фронтальными разделами водных масс. Однако синхронные колебания температуры охватывают области, пространственный масштаб которых обычно не столь велик и не превышает 400 миль.

Далее корреляционная матрица подвергалась разложению по методу главных факторов с итоговым вращением факторных осей. Операция редуцирования матрицы выполнялась путем замены диагональных элементов r_{ii} значениями максимальных коэффициентов корреляции в соответствующей строке. В результате такой подстановки диагональные элементы r_{ii} превращаются в общности. Максимальное значение общности, равное 0.92, наблюдается на юге рассматриваемой акватории, а минимальное значение общности, равное 0.63, — в центральной части Норвежского моря.

Начальное приближение факторного решения получено в результате расчета собственных чисел (λ_i) и собственных векторов (факторных нагрузок a_{ij}) редуцированной матрицы R^h . Одновременно с этим выполнено разложение исходной матрицы АТПО по схеме МГК. В табл. 4.1 приводятся оценки собственных чисел и собственных векторов, полученные по указанным выше схемам.

Как видно из табл. 4.1, редукция корреляционной матрицы существенно повлияла на получаемые оценки собственных векторов и вкладов факторов уже при их начальном приближении, т. е. еще

Таблица 4.1

Сопоставление результатов разложения исходной матрицы АТПО по схемам МФА и МГК

№ п/п	МГК		МФА		
	λ_i	$d_i\%$	до вращения		после вращения
			λ_i	$d_i\%$	$d_i\%$
1	9.1	4.1	7.2	33	18
2	6.3	2.6	5.6	25	22
3	3.7	13	5.3	24	30
4	1.7	7	2.0	9	14
5	1.1	5	0.3	1	8
Σ	21.9	92	20.4	92	92

до начала вращения. Практически не изменился лишь вклад второго фактора. В то же время произошло значительное перераспределение дисперсии между первым и третьим факторами в сторону увеличения последнего.

Вращение факторов производилось путем умножения матрицы редуцированных факторных нагрузок (собственных векторов) на матрицу поворота T , состоящую из синусов и косинусов углов поворота. Заметим, что столбцы этой матрицы ортогональны (взаимонезависимы), что обеспечивает сохранение ортогональности общих факторов и после вращения. Всего было выполнено 11 итеративных вращений, причем каждое вращение соответствовало повороту факторных осей по часовой стрелке на угол $\alpha = 10^\circ$. В результате вращения достигнут максимум дисперсионного функционала (S^2), что эквивалентно максимизации дисперсии нагрузок каждого фактора.

Как следует из табл. 4.1, после вращения произошло перераспределение дисперсии практически между всеми факторами. Однако особенно значительным оно оказалось для первого фактора, вклад которого уменьшился почти в 2 раза по сравнению с его величиной до вращения и более чем в два раза по сравнению с первым вектором, полученным по схеме МГК. Вследствие этого вклад его оказался даже меньше вкладов второго и третьего факторов, что было бы принципиально невозможно при разложении на главные компоненты. Заметим, что максимальный вклад дает третий фактор, величина которого превышает величину третьего вектора почти в 3 раза.

Карты пространственного распределения первых пяти собственных векторов и пяти факторных нагрузок, полученных после вращения, приводятся на рис. 4.3. Нетрудно видеть, что в результате перераспределения вкладов факторов произошло и существенное изменение величин факторных нагрузок. При этом появляется возможность их физической интерпретации. Этим обстоятельством выгодно отличается анализ факторного разложения от компонентного.

Действительно, как следует из распределения первого собственного вектора изменение величин факторных нагрузок является очень однородным для всей рассматриваемой акватории (рис. 4.3, а), причем вычисленные значения a_{1j} малы и находятся на грани статистической значимости. Отсюда можно сделать вывод, что фоновая изменчивость поля АТПО, описываемая 41% от общей дисперсии, примерно одинакова на всей акватории моря. Распределение нагрузок первого общего фактора совершенно иное (рис. 4.3, б). Максимальные значения их сосредоточены в северо-западной части, причем локализация значимых величин фактически ограничивается зоной влияния Восточно-Исландского течения. Следовательно, первый фактор связан с изменчивостью АТПО в районе действия Восточно-Исландского течения.

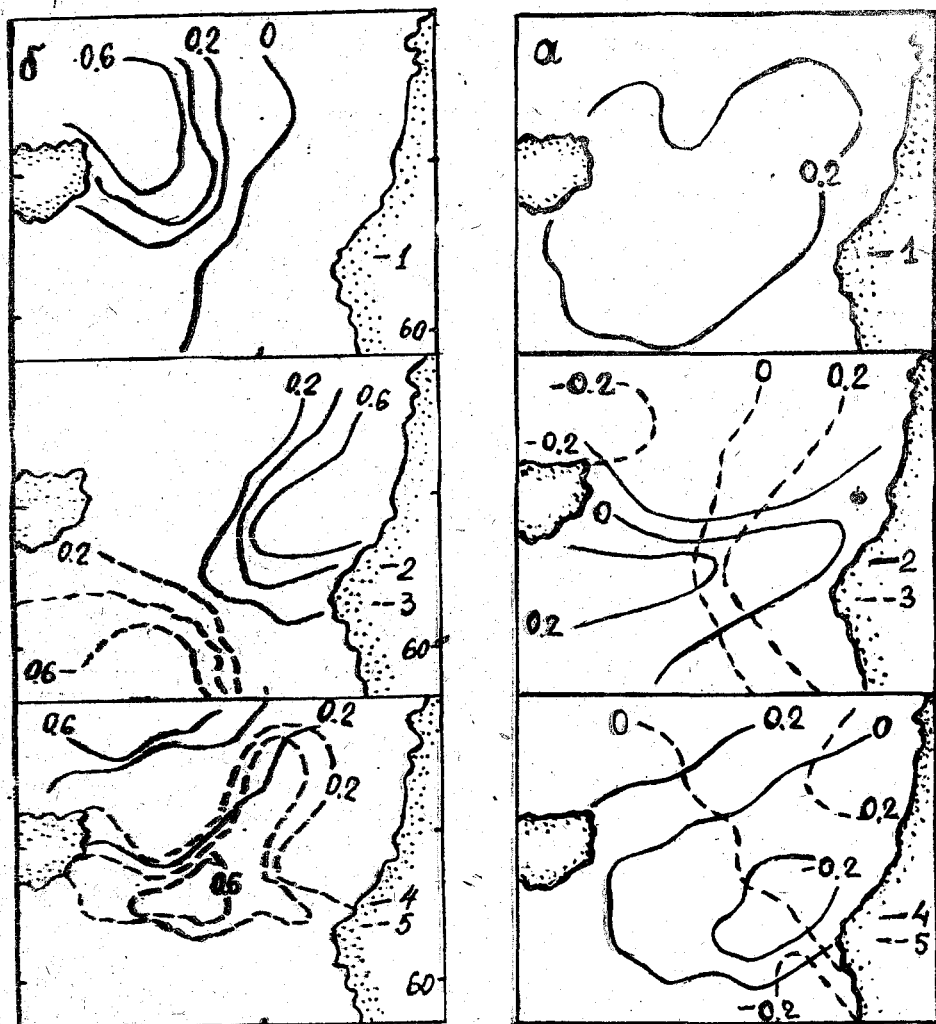


Рис. 4.3. Пространственное распределение первых пяти собственных векторов (а) и пяти факторных нагрузок (б), полученных после вращения поля аномалий температуры воды в Норвежском море

Второй и третий собственные векторы описывают собственно две стоячие волны в поле АТПО — зональную и меридиональную. Пространственный масштаб их примерно в 2 раза меньше, чем фоновая изменчивость, а зоны их влияния взаимопересекаются. Статистическая достоверность отдельных значений векторов f_{2j} и f_{3j} опять, как и в случае МГК, невелика. Данное обстоятельство представляет существенное отличие от распределения нагрузок

второго и третьего факторов (рис. 4.3, б), которые образуют отчетливо выраженные непересекающиеся очаги. Так, наиболее значимый по вкладу в общую изменчивость поля АТПО третий фактор формирует динамику изменений АТПО в северо-восточной части Норвежского моря и связан, очевидно, с деятельностью Норвежского течения. Что касается второго фактора, то он локализуется в районе к юго-востоку от Исландии, где происходит интенсивный перенос вод в системе Северо-Атлантического течения.

Аналогичное соотношение факторов и собственных векторов сохраняется также для четвертого и пятого членов разложения. Четвертый фактор связан, очевидно, с распространением холодных вод в системе Восточно-Гренландского течения, а пятый фактор локализован в центральной части Норвежского моря, где находится полярная фронтальная зона и происходит интенсивное смешение атлантических и полярных вод.

Итак, результаты факторного разложения показывают, что мы имеем дело в Норвежском море с достаточно сложной мозаикой колебаний термического режима, который формируется из пяти основных ядер изменчивости, причем каждое из них связано с действием крупномасштабных океанских течений.

Интересно, что временные колебания АТПО для каждого из факторов различаются друг от друга. При этом в зоне влияния каждого фактора выявляются свойственные только ей циклические колебания. Так, временной ход АТПО первого и четвертого факторов имеет характерный период колебаний в 5—6 лет с выраженным фазовым сдвигом в пространстве, определяющим различие зон Восточно-Исландского и Восточно-Гренландского течений. Для второго и третьего факторов характерны колебания с периодом 2—3 года, которые считаются типичным для водных масс Атлантического происхождения. Временная изменчивость пятого фактора имеет значительно более сложную структуру, в которой выражены колебания с периодами 8—10, 13—15 мес., а также 2—3 года и 5—6 лет.

Следует отметить еще одну важную особенность различий МФА и МГК. Для восстановления поля АТПО по схеме МГК, очевидно, необходимо использовать все 5 главных компонент, т. е.

$$\tilde{x}_{ij} = f_{i1}b_{1j} + f_{i2}b_{2j} + \dots + f_{i5}b_{5j}.$$

где $\tilde{x}_{ij}$ — восстановленное (рассчитанное) значение АТПО в j -й той точке акватории в i -й момент времени; f_{ij} — значения главных компонент; b — регрессионные коэффициенты, определяемые с помощью МНК. При использовании МФА для восстановления поля АТПО может оказаться достаточным учет лишь 2—3-х факторов, имеющих наиболее высокие факторные нагрузки в заданной точке сеточной области.

В заключение отметим, что в настоящее время известны далеко не все возможности практического использования факторного

анализа. Теоретический аппарат МФА, и особенно его адаптация применительно к решению конкретных задач, продолжают интенсивно развиваться и нет сомнений в том, что данный метод станет таким же классическим способом анализа данных, как метод главных компонент.

ГЛАВА 5. МЕТОД КАНОНИЧЕСКОГО КОРРЕЛЯЦИОННОГО АНАЛИЗА

5.1. Постановка задачи

Довольно часто в океанологической практике встречается ситуация, когда необходимо получить количественную оценку связи не между двумя переменными (векторами), а между двумя группами переменных (векторов). Это имеет место, например, при оценке степени сходства термического или халинного режима двух акваторий или когда исследуется взаимодействие атмосферы и океана в нескольких узлах сеточной области одновременно. Использовать для подобной оценки традиционные парные корреляции нецелесообразно, а множественный коэффициент корреляции и сама схема МЛР позволяют найти связь группы переменных относительно только одной зависимой переменной.

Метод канонического корреляционного анализа (МККА) представляется как раз тем средством анализа, который позволяет получить оценку линейно взаимосвязи двух или нескольких групп переменных между собой.

Пусть сходные данные представлены как две одинаковые прямоугольные матрицы X и Y , составленные из N наблюдений за M переменными каждая. Если вычислить коэффициенты корреляции между признаками внутри каждой исходной матрицы наблюдений, то можно сформировать матрицы внутригрупповой корреляции признаков R_{11} и R_{22} . В квадратной симметричной матрице R_{11} размером M строк и M столбцов будут содержаться коэффициенты корреляции между переменными в наборе данных X . Аналогично в матрице R_{22} размерностью $M \times M$ объединятся коэффициенты корреляции переменных набора Y .

Если вычислить попарно все коэффициенты корреляции между переменными набора X и набора Y , то можно составить матрицы межгрупповой корреляции R_{12} и R_{21} . В этих матрицах размерностью $M \times M$ сосредоточены коэффициенты корреляции, которые показывают связь отдельных переменных двух исходных наборов.

Цель применения МККА состоит в том, чтобы отыскать некие канонические (фундаментальные) переменные, коррелированность которых была бы достоверной оценкой взаимосвязи исследуемых групп наблюдений между собой.

Для получения эффективного решения необходимо учесть в процессе вычислений наличие внутригрупповой неоднородности наблюдений, а также наличие шумовых компонент в наблюдениях, которые требуется по возможности нейтрализовать.

Для решения подобных вопросов, как уже указывалось ранее, может быть использован метод главных компонент. При этом заложенная в нем идеология с помощью МККА обобщается на случай исследования обобщенной матрицы данных Z , которая имеет размерность N строк на $2 \times M$ столбцов и объединяет все исходные наблюдения (рис. 5.1). Соответственно, имеющиеся матрицы внутригрупповой и межгрупповой корреляции можно объединить в блочную обобщенную корреляционную матрицу R , которая имеет $2 \times M$ строк и $2 \times M$ столбцов.

Будем искать канонические переменные как линейные комбинации исходных групп наблюдений. Обозначим канонические переменные для группы X как набор векторов U , а канонические переменные для группы Y — как векторы V . По аналогии с МГК запишем

$$U_i = A_i \cdot X^T,$$

$$V_i = B_i \cdot Y^T,$$

где A и B представляют собой некие матрицы коэффициентов перехода, связывающих исходные наблюдения в группе с данной канонической переменной. Для получения единственного решения зададим условия:

— максимальной коррелированности канонических переменных с одинаковыми номерами

$$\rho_i = r(U_i, V_i) \rightarrow \max$$

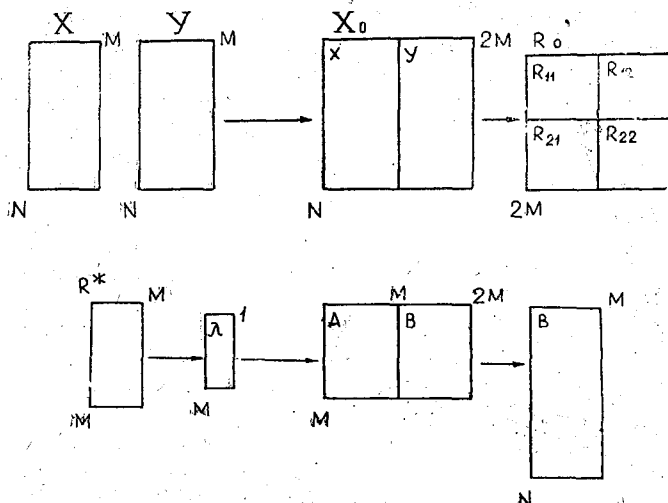


Рис. 5.1. Структурная схема метода канонического корреляционного анализа

среди всех других линейных комбинаций наблюдений;

— некоррелированности канонических переменных с разными номерами

$$r(U_i, V_j) = r(U_i, U_j) = 0, \quad i \neq j.$$

— нормированности дисперсий канонических переменных

$$\sigma^2(U_i) = \sigma^2(V_i) = 1.$$

Количество канонических переменных равно количеству исходных наблюдаемых переменных. Длина вектора каждой канонической переменной равна количеству исходных наблюдений в соответствующей матрице наблюдений. Полная матрица канонических переменных U имеет, следовательно, размерность N строк и M столбцов. Полная матрица коэффициентов перехода также имеет размерность N строк и M столбцов. То же — для матриц V и B из второго набора наблюдений.

Важным достоинством МККА является эффективное сжатие исходной информации уже на первых канонических переменных, которые описывают большую часть дисперсии наблюдений. Соответствующие первые коэффициенты канонической корреляции r_i описывают взаимосвязь наиболее общих колебаний в исследуемых группах наблюдений. Поэтому на практике редко решается полная задача поиска M канонических переменных. Вычислительные процедуры ограничиваются, как правило, первыми 1—3-мя векторами канонических переменных.

5.2. Геометрическая интерпретация МККА

Приближенное геометрическое толкование методики преобразования наблюдений при поиске канонических переменных представлено на рис. 5.2. Пусть используются наблюдения за двумя переменными ($M=2$). Тогда набор X из N наблюдений за двумя

переменными и набор Y аналогичного размера можно представить в координатах (x_1, x_2) как два эллипсовидных облака.

Первая пара канонических переменных U_1 и V_1 представляет собой новые оси координат, вдоль которых одновременно максимизируется дисперсия наблюдений по каждому облаку и

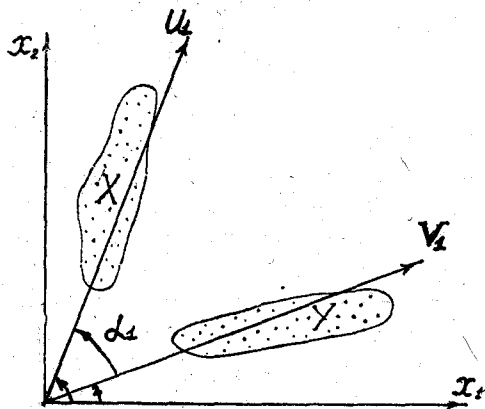


Рис. 5.2. Геометрическая интерпретация МККА: x_1, x_2 — исходные переменные; X, Y — наборы наблюдений; U_1, V_1 — первая пара канонических переменных

максимизируется взаимная корреляция между этими каноническими переменными. Как видно из рис. 5.2, новые оси действительно проходят вдоль направления максимального разброса наблюдений: U_1 — для группы наблюдений X ; V_1 — для группы Y . При этом угол α_1 пересечения векторов U_1 и V_1 должен быть минимальным, что соответствует максимуму коррелированности канонических переменных.

После того как найдена первая пара канонических переменных, описывающих наиболее высоко связанную долю дисперсии в исследуемых выборках, переходят к оценке второй пары переменных U_2 и V_2 для получения α_2 . При этом соблюдается обязательное условие максимального объяснения оставшейся межгрупповой корреляции. Одновременно необходимо, чтобы U_2 и U_1 были ортогональны, аналогичное условие должно выполняться для V_2 и V_1 . Этот процесс продолжается до тех пор, пока не будет исчерпана интересующая исследователя дисперсия наблюдений.

5.3. Вычисление канонических корреляций

Рассматривая общую вычислительную схему, сохраним для простоты принятые параметры исходных матриц наблюдения X и Y . Поэтапная запись алгоритма расчетов может быть представлена в следующем виде.

1. Рассчитываются коэффициенты парной корреляции между всеми переменными в исходных наблюдениях и формируется обобщенная корреляционная матрица R

$$R = \left\| \begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array} \right\|.$$

блоками которой являются матрицы внутригрупповой (R_{11} , R_{22}) и межгрупповой (R_{12} , R_{21}) корреляции.

2. Вычисляются коэффициенты канонической корреляции, для чего решается характеристическое уравнение определителя

$$|R^* - \lambda I| = 0.$$

Здесь R^* — квадратная несимметричная матрица корреляций, элементы которой вобрали в себя и межгрупповую и внутригрупповую корреляцию переменных. Матрица R^* вычисляется по формуле

$$R^* = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21},$$

размерность ее $M \times M$. Следовательно, при решении характеристического уравнения может быть получено M корней

$$\lambda^* = \{\lambda_1, \lambda_2, \dots, \lambda_M\}.$$

Для расчета коэффициентов канонической корреляции используется простое соотношение

$$\rho_i = \sqrt{\lambda_i},$$

где i — номер соответствующей пары канонических переменных.

По величине λ_i нетрудно оценить относительный вклад i -й пары канонических переменных в описание общей изменчивости исходных наблюдений

$$d_i = \lambda_i / \sum_{j=1}^M \lambda_j.$$

3. Расчет коэффициентов (весов) i -й пары канонических переменных A_i и B_i проводится на основе решения системы однородных уравнений вида

$$A_i(R^* - \lambda_i I) = 0,$$

где A_i — искомый вектор, состоящий из M коэффициентов. Для оценки канонических коэффициентов другой выборки наблюдений используется формула

$$B_i = \frac{1}{\lambda_i} R_{22}^{-1} R_{21} A_i.$$

4. Оценивание значений i -й пары канонических переменных осуществляется по уравнениям

$$U_i = A_i X^T,$$

$$V_i = B_i Y^T,$$

где U_i и V_i — векторы длиной N .

Метод МККА пока еще не получил широкого применения в практике океанологических исследований, чтобы можно было в настоящее время раскрыть все его достоинства и недостатки. Поэтому остановимся лишь на тех из них, которые известны уже достаточно хорошо.

1. Величина канонической корреляции в классическом МККА является по определению всегда положительной, поскольку оценивается из выражения $\rho_i = \sqrt{\lambda_i}$. Очевидно, что реальный знак взаимосвязей наборов исходных переменных отнюдь не всегда положителен. Поэтому на практике исследователь вынужден, оценивая знак канонической корреляции, привлекать дополнительные оценки коэффициентов парной корреляции и общие физические соображения.

2. Следует помнить, что хорошая коррелированность канонических переменных еще не означает такой же коррелированности исследуемых групп переменных. Это действительно так, если каждая из канонических переменных при этом описывает подавляющую долю дисперсии признаков внутри каждой из выборок. Если

же вклад первых канонических переменных относительно низок (менее 50—60 %), целесообразно либо изменить исследуемую выборку, либо отказаться от данной методики.

3. Наличие большого количества высоких коэффициентов парной корреляции в обобщенной матрице данных R приводит к затруднениям обращения этой матрицы при решении проблемы собственных чисел. В связи с этим ухудшается точность оценивания канонических корреляций и самих канонических переменных.

4. Оценки канонических переменных оказываются весьма чувствительными к выполнению условий стационарности исходных исследуемых рядов. Поэтому перед проведением МККА целесообразно убедиться в выполнении гипотезы стационарности и провести при необходимости соответствующие преобразования исходных данных.

5.4. Примеры применения МККА

Рассмотрим задачу выявления связи между температурой поверхности океана в Северной Атлантике и температурой воздуха над Европейской территорией России (ЕТР), для решения которой О. В. Батыревой применялся МККА. В качестве исходных данных использовались ряды среднемесячных значений температуры воды на пяти судах погоды в Северной Атлантике и значений температуры воздуха на пяти станциях ЕТР. Канонический корреляционный анализ проводился для обобщения корреляционной матрицы, составленной из 10 строк и 10 столбцов. Анализ парных коэффициентов корреляции показывает, что коррелированность рядов температуры воздуха весьма мала и не превышает 0.3. Коррелированность между выборками также низка: максимальная величина парной корреляции температуры воды и воздуха составляет 0.46, а средняя — 0.13. Наибольшая коррелированность наблюдается в рядах температуры воды — 0.6. В процессе применения МККА получены следующие результаты:

Номер канонической переменной	Величина канонической корреляции	Суммарный вклад канонической переменной
1	0.84	43.7
2	0.52	71.1
3	0.32	88.1
4	0.14	95.4
5	0.03	100

Как видно из приведенной таблицы, первая каноническая переменная указывает на высокую значимую взаимосвязь полей температуры воды и воздуха. Вторая каноническая переменная также соответствует высокой коррелированности. Остальные корреляции незначимы.

Однако на деле оказывается, что эффективной можно признать лишь первую каноническую переменную, поскольку она описывает связь для 44,8 % дисперсии поля температуры воды и 31,2 % дисперсии поля температуры воздуха. Медленная сходимость разложений полей температуры снижает эффективность канонических корреляций.

Анализ пространственной структуры вектора 1-й канонической переменной демонстрирует наличие двух очагов корреляции, один из которых расположен над океаном в зоне Исландской депрессии, а другой — в центральной части ЕТР. Совпадение знака вектора в этих очагах указывает на реальность положительного знака канонической корреляции. На основании проведенных расчетов можно сделать вывод о синхронности крупномасштабных колебаний термического режима центра ЕТР и Исландской депрессии.

ГЛАВА 6. МЕТОД КЛАСТЕРНОГО АНАЛИЗА

6.1. Общие понятия. Постановка задачи

Кластерный анализ, или, другими словами, метод автоматической классификации, относится к числу наиболее быстро развивающихся направлений многомерного анализа. Очевидно, прежде чем рассматривать этот метод, необходимо дать определение понятию «классификация».

Классификация — это разбиение множества объектов (явлений, процессов и т. п.) на ряд подмножеств (классов, таксонов) по признаку их сходства. Объекты, отнесенные к одному и тому же классу, должны быть сходны между собой и отличаться от объектов других классов. Поэтому можно считать, что классификация — это средство выделения наиболее важных особенностей классифицируемых объектов и обобщения наших представлений о них.

Одним из основных требований, предъявляемых классификации, является ее объективность, а именно: выделенные классы должны соответствовать реально существующим классам и не зависеть от точки зрения лица, выполняющего классификацию.

Классификация может быть специальной (целевой) и естественной (универсальной). При целевой классификации признаки задаются заранее, а число их обычно невелико (например, классификация поверхности океана по температуре воды или классификация вертикальной структуры океана по температуре). Универсальная классификация — это классификация по комплексу различных признаков, причем при изменении набора признаков (частичном или даже полном) она остается неизменной (например, периодическая система химических элементов Д. И. Менделеева). Естественно, что построение универсальной классификации

представляет чрезвычайно сложную задачу. В гидрометеорологии построение такой классификации может быть связано, например, с проблемой описания климата или с выделением водных масс, так как эти понятия носят преимущественно качественный характер и не фиксируются каким-либо заранее заданным набором признаков. Очевидно, целевая классификация — это частный случай универсальной классификации. Так, выделение водных масс по температуре и солёности — это уже целевая классификация.

Исходя из определения классификации, можно, очевидно, рассматривать объекты как некоторые таксономические единицы, а признаки, их характеризующие, — многомерное пространство (гиперпространство). Практически именно это и составляет сущность геометрической интерпретации кластерного анализа (КА).

В качестве примера рассмотрим вертикальное распределение температуры, солёности и содержания кремния на гидрологической станции (рис. 6.1). Реальные наблюдения представлены в признаковом пространстве, состоящем из трех взаимно ортогональных осей координат как точки, которые образуют неоднородное по плотности облако. Естественно, что каждая водная масса должна характеризоваться близкими значениями наблюдаемых параметров. Поэтому водная масса может быть представлена

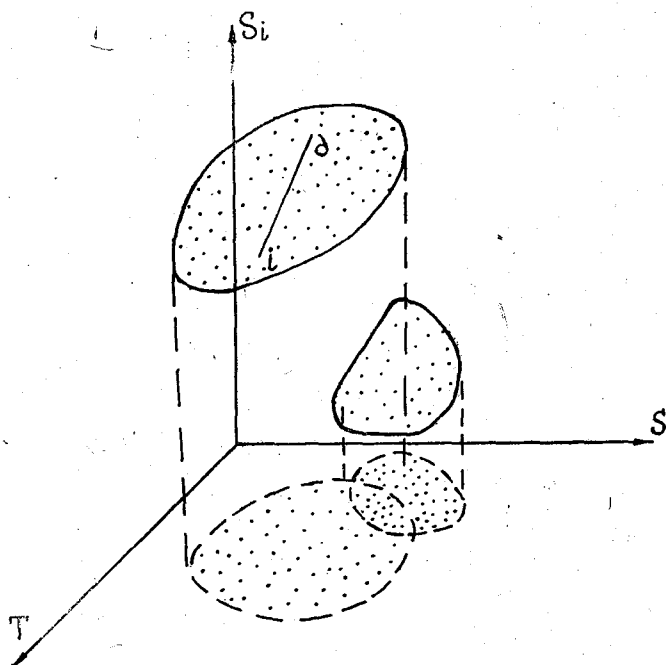


Рис. 6.1. Графическое разделение водных масс в координатах: температура T , солёность S , содержание кремния Si

в этом пространстве признаков как некоторый замкнутый объем, внутри которого точки наблюдений образуют достаточно плотную компактную группу. Данный объем отделяется от соседнего объема, составляющего другую водную массу, достаточно широкой разделительной зоной.

Итак, в общей постановке проблема классификации объектов заключается в том, чтобы всю анализируемую совокупность объектов разбить на сравнительно небольшое, обычно заранее не известное, число квазиоднородных групп или классов. При этом понятие однородности основывается на предположении, что геометрическая близость двух или нескольких объектов означает близость их физических состояний, их сходство. Считается также, что варьированием характеристик внутри классов можно пренебречь как несущественным. Иногда, исходя из анализа дисперсий, говорят, что внутригрупповая дисперсия должна быть значительно меньше межгрупповой дисперсии.

Пусть выборка из M исходных наблюдений представляет собой множество Ω в M -мерном пространстве признаков. Требуется статистически обоснованно выделить ряд подмножеств $\omega_1, \omega_2, \dots, \omega_k$, на которые разбивается исходное множество

$$\Omega = \omega_1 \cup \omega_2 \cup \dots \cup \omega_k,$$

где k количество таких подмножеств, которые традиционно называются классами. Устойчивое выделение подмножеств возможно при выполнении следующих условий:

- 1) получаемые подмножества не будут пересекаться:

$$\omega_i \cap \omega_j = \emptyset \quad \text{при любых } i \neq j.$$

Следуя этому требованию, каждый вектор наблюдений из M переменных $X_i^T = \{X_{i1}, X_{i2}, \dots, X_{iM}\}$ может одновременно принадлежать только одному из выделенных подмножеств ω_i . Из других подмножеств этот вектор наблюдений должен быть исключен;

- 2) каждое подмножество не должно быть пустым

$$\omega_i \neq \emptyset, \quad i = 1, 2, \dots, k.$$

Другими словами, в каждом выделенном классе должен находиться хотя бы один вектор наблюдений. Следовательно, возможное количество классов колеблется в пределах от $k=1$ (все наблюдения относятся к одному классу) до $k=N$ (каждое наблюдение образует отдельный класс);

- 3) неоднородное распределение объектов в пространстве признаков, т. е. наличие компактных групп точек. Действительно, в противном случае, когда имеет место сравнительно однородное распределение объектов, всякая классификация теряет смысл.

В общем виде задачу классификации можно разделить на ряд последовательных этапов:

- а) выбор исходной системы признаков;
- б) снижение размерности признакового пространства;

- в) выбор меры сходства (близости);
 - г) выбор алгоритма классификации и оценка их качества.
- Рассмотрим отдельно каждый из этих этапов.

6.2. Выбор исходной системы признаков

Вопрос выбора исходных признаков в значительной степени является субъективным и зависит от цели классификации, наличия априорной информации, ее качества и других факторов. В случае целевой классификации исходные признаки, как уже указывалось выше, задаются заранее. При построении естественной классификации число исходных признаков обычно неизвестно. В общем случае следует ожидать, что увеличение количества исходных признаков должно приводить к улучшению качества классификации. Но априори нельзя судить о достаточном числе признаков, описывающем систему с заданной точностью. Зачастую неизвестны взаимосвязи признаков, их возможные группировки, зашумленность и т. п. Поэтому возникает проблема оптимального выбора исходных признаков, обеспечивающих наилучшее распознавание объектов.

Естественно, весьма трудно дать конкретные рекомендации по подбору исходных признаков. В принципе желательно использовать те из них, которые некоррелированы между собой. Наилучшими здесь будут признаки, у которых разность средних значений велика по сравнению со стандартными отклонениями. К сожалению, на практике соблюсти это требование зачастую оказывается очень сложно. Кроме того, следует иметь в виду, что выбор системы исходных признаков тесно связан с предварительной обработкой данных наблюдений.

Один из конкретных аспектов проблемы выбора признаков наглядно следует из рис. 6.1. Предположим, что данные наблюдений за содержанием кремния отсутствуют. В этом случае все облако точек будет спроектировано в плоскость T , S -координат. Однако разделение водных масс по температуре и солености практически невозможно. Нетрудно видеть, что в плоскости T , S -координат классы являются пересекающимися: заштрихованная область включает в себя наблюдения, одновременно принадлежащие двум классам. Дополнительное привлечение кремния в качестве классификационного признака принципиально меняет картину. При одновременном учете всех трех признаков уже становится возможным достоверное выделение водных масс, поскольку именно в 3-мерном пространстве обе группы точек не пересекаются.

6.3. Снижение размерности признакового пространства

Данный вопрос является составной частью задачи классификации. Особенно актуальным он оказывается в том случае, если

матрица исходных данных имеет большой размер. Главная цель подобного преобразования состоит в максимальном сжатии информации об объектах при минимизации потерь и искажений. Заметим, что сжатие информации означает по существу переход из первичного пространства исходных признаков во вторичное пространство меньшей размерности. Эффективными способами сжатия информации являются, например, МГК и МФА.

МГК позволяет разделить информацию об исходном поле признаков на линейно независимые составляющие с существенно различной дисперсией. При этом, отбрасывая составляющие с малой дисперсией, мы, тем самым, отфильтровываем мелкомасштабные шумы, что имеет большое значение для успешной классификации. Достоинством МГК является также то, что он одновременно производит «взвешивание» признаков, придавая больший вес тем из них, которые обладают большей дисперсией. Наконец, важно и то, что проецирование исходной совокупности объектов в пространство меньшей размерности, натянутое на k' первых главных компонент, наименее искажает ее геометрическую конфигурацию.

6.4. Выбор меры близости

Выбор меры близости является узловым моментом классификации, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при заданном алгоритме классификации. Мерой близости объектов может служить любой коэффициент связи, позволяющий численно оценить степень сходства или различия данной пары объектов по набору признаков. Наиболее часто встречаются метрические показатели сходства, характеризующие расстояние между объектами в геометрическом пространстве признаков.

Если расстояние между двумя объектами X_j и X_{j+1} обозначить как $d(X_j, X_{j+1})$, то для него должны выполняться следующие условия:

- а) $d(X_j, X_{j+1}) \geq 0$;
- б) $d(X_j, X_{j+1}) = 0$ при $X_j = X_{j+1}$;
- в) $d(X_j, X_{j+1}) = d(X_{j+1}, X_j)$;
- г) $d(X_j, X_{j+1}) \leq d(X_j, X_c) + d(X_{j+1}, X_c)$.

Последнее условие означает, что расстояние между объектами X_j и X_{j+1} всегда является наикратчайшим по сравнению с любым другим расстоянием между этими объектами, проходящим через произвольную точку X_c .

Наиболее универсальным метрическим показателем является метрика Минковского

$$d(X_j, X_{j+1}) = \left[\sum_{i=1}^n p_i (x_{ij} - x_{i, j+1})^e \right]^{1/e}, \quad (6.1)$$

где n — число исходных признаков; p_i — вес i -го признака; ε — показатель степени. Показатель степени в (6.1) может быть любым положительным числом. На практике наиболее часто принимают $\varepsilon=2$. Тогда имеем взвешенную евклидову метрику

$$d(X_j, X_{j+1}) = \sqrt{\sum_{i=1}^n p_i (x_{ij} - x_{i,j+1})^2}.$$

Однако далеко не всегда мы можем взвесить исходные признаки, т. е. придать им тот или иной вес. Поэтому, чтобы не отдавать предпочтение каким-либо признакам, обычно принимают $p_i=1$. В результате получим евклидову метрику

$$d(X_j, X_{j+1}) = \sqrt{\sum_{i=1}^n (x_{ij} - x_{i,j+1})^2}. \quad (6.2)$$

В некоторых случаях для оценки расстояния по формуле (6.1) используется $\varepsilon=1$. Тогда имеем так называемое манхэттенское расстояние

$$d(X_j, X_{j+1}) = \sum_{i=1}^n (x_{ij} - x_{i,j+1}).$$

Могут быть и другие варианты применения формулы (6.1). Подчеркнем, что окончательный выбор вида метрики всегда является прерогативой исследователя.

Существуют и другие способы задания расстояния между объектами помимо метрики (6.1). В частности, достаточно широко используется так называемое расстояние (метрика) Махаланобиса, имеющее следующий вид:

$$d_{ij} = (X_i - X_j)^T S^{-1} (X_i - X_j),$$

где $(X_i - X_j)$ и $(X_i - X_j)^T$ — соответственно вектор-строка и вектор-столбец разностей значений M признаков на i -м и j -м наблюдениях; S — матрица ковариации переменных. По существу данная метрика означает расстояние от объекта до центра тяжести всей совокупности. Наиболее эффективно оно характеризует сходство объектов в случае коррелированности их между собой.

При построении классификации возможна ситуация, когда мы имеем дело только с качественными признаками, задаваемыми как наличие или отсутствие данного явления или объекта. В этом случае информация о признаках задается в виде набора нулей и единиц. Мерой сходства таких объектов может служить коэффициент ассоциативности

$$d(X_j, X_{j+1}) = \frac{\sum_i p_i \varphi_i}{\sum_i p_i}, \quad (6.3)$$

где p_i — вес, соответствующий тому или иному признаку; φ_i — степень сходства между объектами X_j, X_{j+1} ($0 \leq \varphi_i \leq 1$). Принимая равенство весов в (6.3), имеем расстояние (метрику) Хемминга

$$d(X_j, X_{j+1}) = \sum_{i=1}^n \varphi_i,$$

$$\text{где } \varphi_i = \begin{cases} 0, & \text{если } X_j = X_{j+1} \\ 1, & \text{если } X_j \neq X_{j+1}. \end{cases}$$

Наконец, еще один способ задания меры близости — это оценка коэффициента корреляции или других подобных ему коэффициентов, характеризующих степень различия между объектами.

В результате последовательного расчета степени близости между m объектами получим следующую матрицу:

$$d = \begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1m} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ d_{m1} & d_{m2} & d_{m3} & \dots & d_{mm} \end{pmatrix}, \quad (6.4)$$

которая является симметричной и квадратной, причем элементы, находящиеся на главной диагонали, будут равны нулю ($d_{11} = d_{22} = \dots = d_{mm} = 0$). Матрица в виде (6.4) является основой для выполнения непосредственно самой процедуры классификации.

6.5. Алгоритмы классификации и методы оценки их качества

К настоящему времени накоплено большое число различных алгоритмов автоматической классификации, которые можно разделить на три большие группы: вероятностные, эвристические, иерархические. Наибольшее распространение в гидрометеорологии получили эвристические алгоритмы, основанные на выделении в многомерном пространстве в виде некоторых геометрических образов компактных групп точек. Как правило, эвристические алгоритмы достаточно просты, легко реализуются на ЭВМ и обладают сравнительно высокой эффективностью.

Наибольшую известность в океанологических исследованиях заслуженно завоевал метод кратчайшей связывающей сети (кратчайший незамкнутый путь). Этот метод относится к числу параллельных процедур, т. е. процедур, в которых выборка вся сразу поступает на классификацию. При последовательных процедурах, наоборот, элементы выборки поступают на классификацию по одному или небольшими частями.

Суть метода кратчайшей связывающей сети (КСС) состоит в построении сети отрезков, соединяющих исследуемые объекты

в пространстве признаков в единую сеть таким образом, чтобы ее суммарная длина была бы минимальной. Отсюда следует, что должно выполняться условие

$$\sum_{j=1}^{m-1} d(X_j, X_{j+1}) \rightarrow \min,$$

где m — число точек (объектов).

Порядок работы алгоритма КСС состоит в следующем. Прежде всего из матрицы расстояний (6.4) выбирается минимальное расстояние, соответствующее паре наиболее близких точек в пространстве исходных или вторичных признаков. Если таких пар несколько, то берется любая из них. После этого из матрицы расстояний выбирается точка, наиболее близкая к первому звену, и соединяется с ним вторым звеном. Далее ищется точка, ближайшая к уже построенной сети. Эта процедура повторяется до тех пор, пока все точки не окажутся связанными друг с другом кратчайшей сетью.

В результате построения КСС мы получаем некоторую последовательность звеньев с различной длиной. «Разрубая» звено с максимальной длиной, получим разбиение исходного множества на 2 класса. Если «разрубим» следующее звено, имеем уже 3 класса и т. д. Возникает вопрос, какое из разбиений будет наилучшим. Другими словами, при какой величине $d_{кр}$, представляющей собой некоторое критическое (пороговое) значение длины звена КСС, классификация оказывается наиболее эффективной (оптимальной).

Решение данного вопроса возможно, как минимум, двумя способами. Во-первых, величина $d_{кр}$ может априорно оцениваться исходя из неформальных (физических) соображений. Например, при выделении водных масс дополнительное использование процедуры классического T, S -анализа оказывается весьма полезным для принятия правильного решения о числе выделенных классов. Так как в каждом конкретном случае неформальные соображения носят индивидуальный характер, то какие-либо рекомендации по этому поводу невозможны.

Второй способ является формальным и состоит в итеративном поиске оптимального разбиения на классы путем сравнительного анализа различных вариантов классификации. С этой целью обычно производится построение тех или иных функционалов качества (целевых функций, решающих правил) классификации.

При использовании метрических показателей сходства весьма удобным оказывается, например, такой критерий качества, который достигает максимума при среднем минимальном внутриклассовом расстоянии и среднем максимальном межклассовом рас-

стоянии. Если соответственно среднее внутриклассовое и среднее межклассовое расстояние выразить как

$$J_{\text{вк}} = \frac{1}{m-k-1} \sum_{j=1}^{m-k-1} d(X_j, X_{j+1}),$$

$$J_{\text{мк}} = \frac{1}{k-1} \sum_{j=1}^{k-1} d(X_j, X_{j+1}),$$

где k — число классов, то критерий качества классификации, позволяющий оценить оптимальное число классов, имеет вид

$$J_{\text{опт}} = J_{\text{мк}} - J_{\text{вк}} \rightarrow \max. \quad (6.5)$$

Заметим, что кроме критерия (6.5) существуют десятки других функционалов качества, однако практически все они основываются на поиске каких-либо соотношений между внутригрупповыми и межгрупповыми расстояниями, дисперсиями или иными коэффициентами связи. Следует иметь ввиду, что чисто статистические критерии играют лишь вспомогательную роль в классификационных процедурах. С одной стороны, они позволяют отбраковать заведомо неудачные группировки, но с другой — оптимальные по этим критериям классификации могут и не иметь содержательной ценности. Поэтому неформальный анализ результатов следует считать более важным по сравнению со статистическими критериями.

Как уже указывалось, алгоритм КСС легко реализуется на ЭВМ. Однако в тех случаях, когда число исходных признаков мало ($n \leq 3$), классификационная процедура существенно облегчается и может быть выполнена без применения ЭВМ. Это значительно упрощает интерпретацию и содержательный анализ полученных результатов.

В качестве примера рассмотрим классификацию 9 гидрологических станций в некотором районе океана по вертикальному профилю температуры воды. В результате наблюдений была составлена матрица размером 9×26 , разложение которой по естественным составляющим показало, что первое собственное число описывает 50 % дисперсии, а второе — 31%, т. е. оставшиеся 7 собственных чисел дают в сумме лишь 19% дисперсии. Очевидно, для классификации этих гидрологических станций достаточно ограничиться первыми двумя собственными векторами X_1 и X_2 .

Матрица расстояний между точками в пространстве X_1 и X_2 (рис. 6.2), определенная на основе евклидовой метрики, приводится в табл. 6.1.

Первую строку этой матрицы занесем в первую строку так называемой решающей таблицы (табл. 6.2). Определим точку, ближайшую к первой. Очевидно, это будет точка под номером 6, так как расстояние 1—6 является минимальным в данной строке. Эти точки образуют первое звено и заносятся в итоговую таблицу

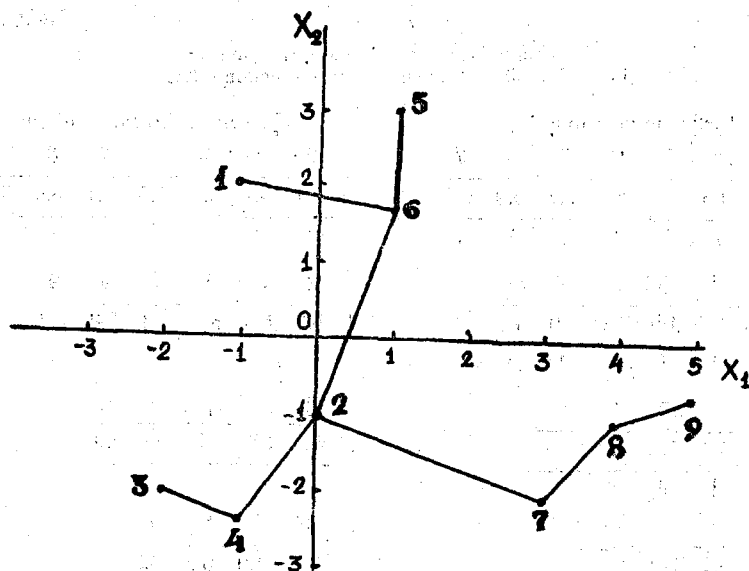


Рис. 6.2. Кратчайшая связывающая сеть гидрологических станций в пространстве X_1 и X_2

Таблица 6.1

Матрица расстояний между точками
в пространстве собственных векторов X_1 и X_2

	1	2	3	4	5	6	7	8	9
1	—	3.1	4.2	4.5	2.3	2.2	5.7	5.9	6.5
2	3.1	—	2.1	1.3	4.2	2.7	3.1	4.0	5.0
3	4.2	2.1	—	1.2	5.8	4.6	5.0	6.1	7.1
4	4.5	1.3	1.2	—	5.8	4.5	4.0	5.2	6.3
5	2.3	4.2	5.8	5.8	—	1.5	5.4	5.0	5.4
6	2.2	2.7	4.6	4.5	1.5	—	4.1	3.9	4.6
7	5.7	3.1	5.0	4.0	5.4	4.1	—	1.4	2.6
8	5.9	4.0	6.1	5.2	5.0	3.9	1.4	—	1.2
9	6.5	5.0	7.1	6.3	5.4	4.6	2.6	1.2	—

с указанием его длины. Запомним расстояния от выбранной точки до всех остальных, записав их в первую строчку вспомогательной таблицы.

После этого начинаем формирование второй строки решающей таблицы, в которую заносятся минимальные звенья, соединяющие 1 и 6 точки со всеми остальными. Как видно из этой строки, 3 и 4 точки оказались ближе к первой, а 2, 5, 8 и 9 точки ближе к 6. В результате ближайшей точкой к звену 1—6 оказывается точка 5,

Таблица 6.2

**Процедура составления таблиц
при ручной классификации на основе КСС**

Решающая таблица									Вспомогательная таблица								
2	3	4	5	6	7	8	9		2	3	4	5	7	8	9		
3.1	4.2	4.5	2.3	2.2	5.7	5.9	6.5		6	2.7	4.7	4.5	1.5	4.1	3.9	4.6	
1	1	1	1	1	1	1	1										
2	3	4	5	7	8	9			2	3	4	7	8	9			
2.7	4.2	4.5	1.5	4.1	3.9	4.6			5	4.2	5.3	5.8	5.4	5.0	5.6		
6	1	1	1	1	1	1											
2	3	4	7	8	9				3	4	7	8	9				
2.7	4.2	4.5	4.1	3.9	4.6				2	2.1	1.8	3.1	4.0	5.0			
6	1	1	6	6	6												
3	4	7	8	9					3	7	8	9					
2.1	1.3	3.1	3.9	4.6					4	1.2	4.0	5.2	6.3				
2	2	2	6	6													
3	7	8	9						7	8	9						
1.2	3.1	3.9	4.6						3	5.0	6.1	7.1					
4	2	6	6														
7	8	9							8	9							
3.1	3.9	4.6							7	1.4	2.6						
2	6	6															
8	9								9								
1.4	2.6								8	1.2							
7	7																
9																	
1.2																	
9																	

Итоговая таблица			
Звено		Длина	
1—6		2.2	
6—5		1.5	
6—2		2.7	
2—4		1.8	
4—3		1.2	
2—7		3.1	
7—8		1.4	
8—9		1.2	

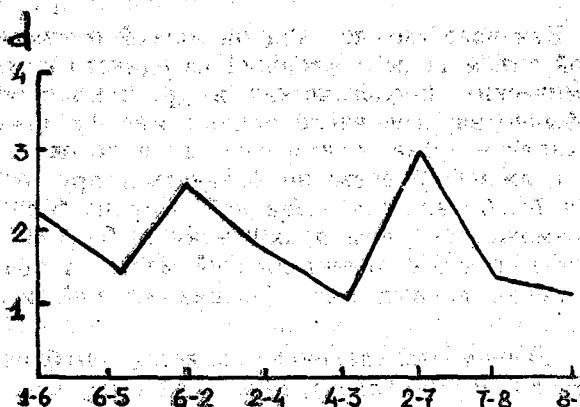


Рис. 6.3. График длин звеньев кратчайшей связывающей сети.

расстояние до которой минимальное. Заносим это звено в итоговую таблицу с указанием его длины. Далее заполним для 5 точки соответствующую строку вспомогательной таблицы и продолжим процесс составления звеньев до тех пор, пока все точки не окажутся соединенными между собой (рис. 6.2).

Если данные итоговой таблицы нанести на график (рис. 6.3), то они образуют некоторую кривую с хорошо выраженными пиками. Очевидно, каждый пик может служить границей классов. Поскольку задать величину $d_{кр}$ априори исходя из неформальных соображений не представилось возможным, то для оценки оптимальности полученной классификации воспользуемся критерием (6.5). Оказалось, что оптимальным с точки зрения данного критерия является число классов, равное трем. Заметим, что критерий (6.5) является достаточно грубым и дает четкое разделение лишь в очевидных случаях. При неэффективности критерия (6.5) используются более тонкие методы.

В заключение подчеркнем, что даже автоматическая классификация, которая по праву может считаться объективной, имеет целый ряд субъективных моментов. К их числу, в частности, относится выбор системы исходных признаков, который полностью является прерогативой исследователя. При снижении размерности признакового пространства определенные затруднения вызывает вопрос отсеивания мелкомасштабных составляющих. Наконец, выбор конкретной меры сходства, алгоритма классификации и критерий ее качества также зависит от самого исследователя. Все это необходимо учитывать при решении классификационных задач.

6.6. Использование кластерного анализа

К числу наиболее типичных задач, где использование МКА является весьма эффективным, относится выделение и анализ

водных масс. Как известно, под водной массой понимается достаточно большой объем воды, состоящий из единого комплекса физических, химических, биологических и др. характеристик. Учитывая, что обычно заранее число водных масс, а также определяющих их характеристик неизвестно, но в то же время ядра водных масс должны существенно отличаться друг от друга, то эффективность МКА является очевидной. Именно большая объективность и возможность учета практически любого набора исходных признаков выгодно отличает данный метод от традиционных методов выделения водных масс (изопикнический анализ, T , S -анализ и др.).

Первой в данном направлении является опубликованная в 1976 г. работа Ю. В. Николаева, в которой выполнена классификация водных масс арктических морей.

Возможны различные варианты конкретного применения МКА, однако в общем можно выделить три типичных случая:

- 1) вертикальное распределение водных масс в заданной точке (например, на гидрологической станции);
- 2) пространственное распределение водных масс на заданном горизонте (например, на поверхности океана);
- 3) пространственно-вертикальное распределение водных масс (например, через гидрологический разрез).

Рассмотрим конкретный пример выделения водных масс на меридиональном разрезе, расположенном в восточной части Северной Атлантики вдоль 10° з.д. от 33° с.ш. до 36° с.ш., выполненный в XI рейсе УЭС «Профессор Сергей Дорофеев» (ноябрь 1988 г.). Для анализа привлечены 7 параметров: температура воды (T), соленость (S), условная удельная плотность (ρ), содержание растворенного кислорода (O_2) в мл/л и в %, концентрации общего фосфора (P) и кремния (Si). В поле каждого из параметров наблюдается ряд особенностей, которые не позволяют однозначно выделить водные массы и зоны их смешения. Так, в поле температуры (рис. 6.4) отчетливо выделяется верхний однородный слой толщиной 40—60 м с $T \approx 20^\circ\text{C}$, а также слабо выраженный максимум температуры ($T > 11^\circ$) в слое 700—1100 м. В распределении солености наиболее отчетливо просматриваются ядра минимальной ($S = 35,6\text{‰}$ в слое 300—600 м) и максимальной ($S = 36,4\text{‰}$ в слое 1100—1300 м) солености воды. В распределении плотности выявляется слой скачка на горизонте 80 м, а в остальном отмечается плавный рост плотности с глубиной с практически постоянным вертикальным градиентом (рис. 6.5).

Построенные T , S -диаграммы выявляют наличие 4-х основных водных масс, ядра которых помечены на этом же рисунке цифрами 1—4. В соответствии с существующей типизацией вод Северной Атлантики полученные массы могут быть классифицированы следующим образом. Три водные массы имеют атлантическое

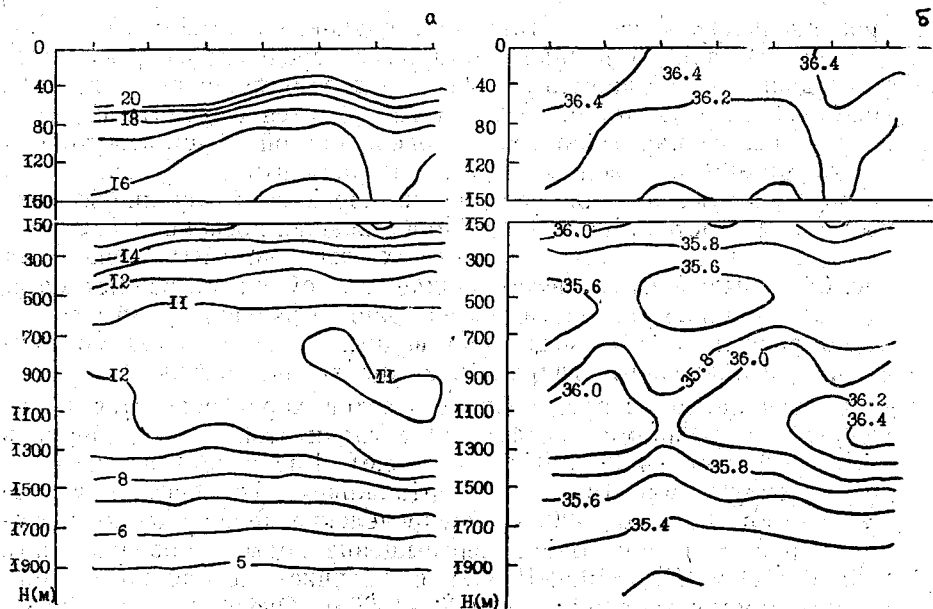


Рис. 6.4. Вертикальное распределение на разрезе температуры (а) и солености (б) воды

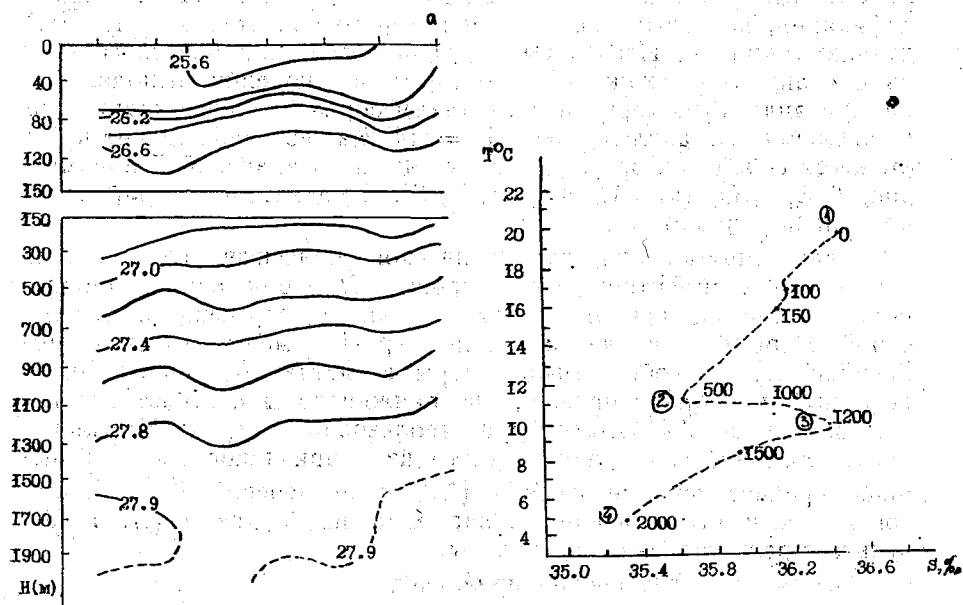


Рис. 6.5. Вертикальное распределение на разрезе условной удельной плотности воды (а) и осредненная T, S-диаграмма (б)

происхождение: поверхностная (1), подповерхностная (2) и глубинная (4), а одна водная масса имеет средиземноморское происхождение — это промежуточная средиземноморская вода (3). Наиболее точно методом T, S -анализа удастся выделить индексы ядер и положение границы подповерхностной и промежуточной средиземноморской водных масс. Разграничить поверхностную и промежуточную, а также средиземноморскую и глубинную водные массы удастся лишь весьма приближенно, поскольку T, S -кривая для них не имеет отчетливых изгибов. Штриховкой на рис. 6.8, а выделены соответствующие слои смешения водных масс, где потенциально располагается их граница раздела. Чтобы достоверно решить задачу выделения водных масс, необходимо привлечь дополнительные гидрохимические характеристики.

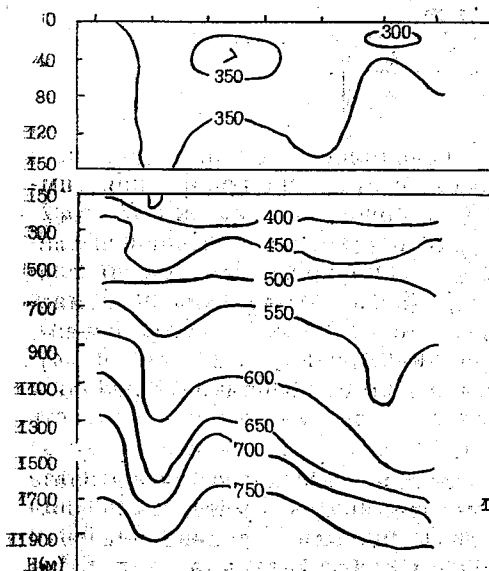
Вертикальное распределение фосфатов характеризуется повышенными значениями вертикальных градиентов концентрации в глубинном слое 1400—1600 м (рис. 6.6). Распределение кремния демонстрирует максимальный вертикальный градиент в подповерхностном слое 100—200 м. Распределение кислорода по вертикали (рис. 6.7) показывает пересыщение воды в поверхностном слое 0—100 м ($O_2 = 100 \div 105\%$) и наличие кислородного минимума в промежуточном слое 600—1300 м. Очевидно, что только совместный анализ всех имеющихся полей океанологических характеристик может позволить достоверно выделить все водные массы. Для этого и был привлечен метод автоматической классификации, исходные данные для которого подготавливались следующим образом. Исследуемый океанологический разрез состоял из 6 станций, на каждой из которых наблюдения выполнялись на 19 стандартных горизонтах, и таким образом, общее количество наблюдений составило: $N = 6 \times 19 = 114$. Каждое наблюдение представляет собой вектор (X), составленный из значений 7 параметров: T, S, ρ, O_2 (мл/л), O_2 (%), P, Si . Следовательно, количество признаков равно $M = 7$.

Задача применения автоматической классификации состоит в том, чтобы объективно сгруппировать N наблюдений в k однородных классов, причем классы не должны пересекаться между собой. Первый шаг классификации — расчет меры сходства наблюдений между собой. Выбор меры сходства обусловлен видом исходных признаков: признаки не ранжированы и имеют различную размерность. Поэтому была использована метрика Махалобиса. Предварительно для каждого из 7 признаков были вычислены среднее арифметическое ($\bar{X}_i$) и дисперсия (S_{ii}^2). Кроме того, рассчитаны значения ковариации признаков (S_{ij}). В результате расчетов были составлены

вектор средних значений признаков

$$\bar{X}^T = \{\bar{X}_1, \bar{X}_2, \dots, \bar{X}_7\}$$

а



б

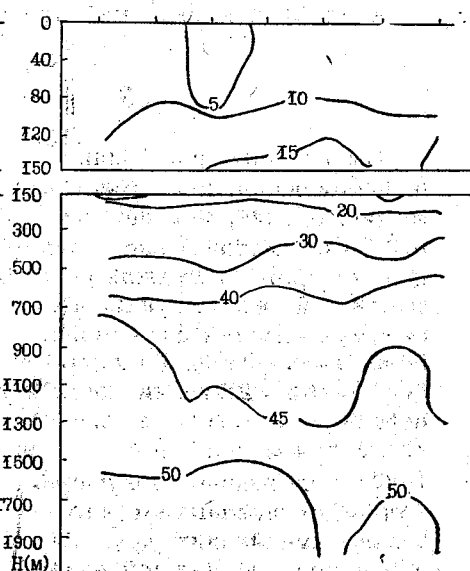
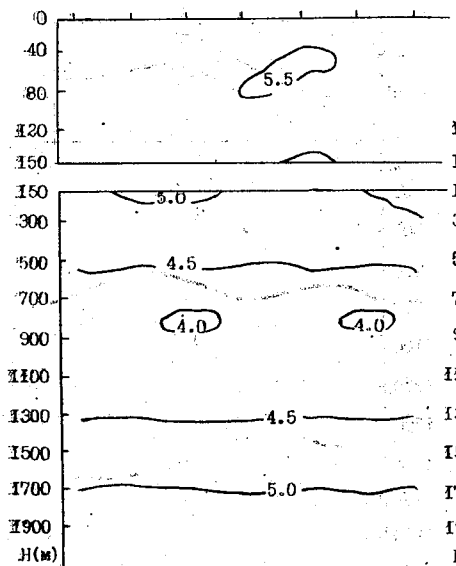


Рис. 6.6. Вертикальное распределение на разрезе концентраций фосфатов (а) и кремния (б)

а



б

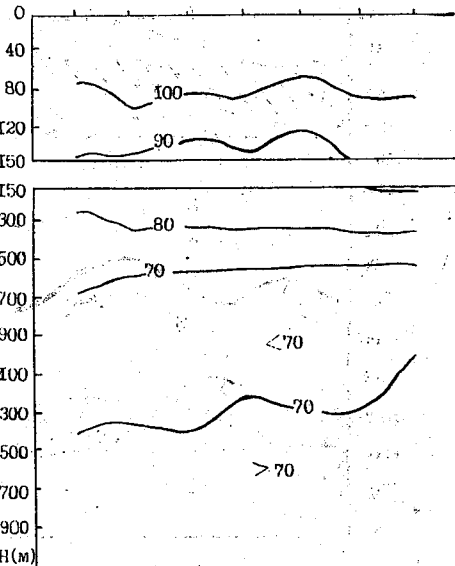


Рис. 6.7. Вертикальное распределение на разрезе концентраций кислорода, мл/л (а) и насыщения воды кислородом, % (б)

и матрица ковариаций признаков

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{17} \\ S_{21} & S_{22} & \dots & S_{27} \\ \vdots & \vdots & \ddots & \vdots \\ S_{71} & S_{72} & \dots & S_{77} \end{pmatrix}.$$

Вычисление расстояний d_{ij} проводилось попарно для всех N наблюдений, и в результате была составлена квадратная симметричная матрица расстояний D размерностью N строк и N столбцов, которая послужила основой для построения кратчайшей связывающей сети. Группировка наблюдений производилась по максимуму значений функционала $J_{\text{опт}}$ (6.5). Оптимальное разбиение было получено для случая выделения четырех классов ($k=4$). При таком разбиении достигается минимум внутриклассовой дисперсии признаков, а следовательно, наибольшая однородность классов.

Расположение полученных классов в плоскости разреза (рис. 6.8) показывает наличие четырех водных масс, последовательно сменяющих друг друга по вертикали. Средние значения признаков для каждой водной массы представлены в табл. 6.3 и наглядно показывают степень различия между ними.

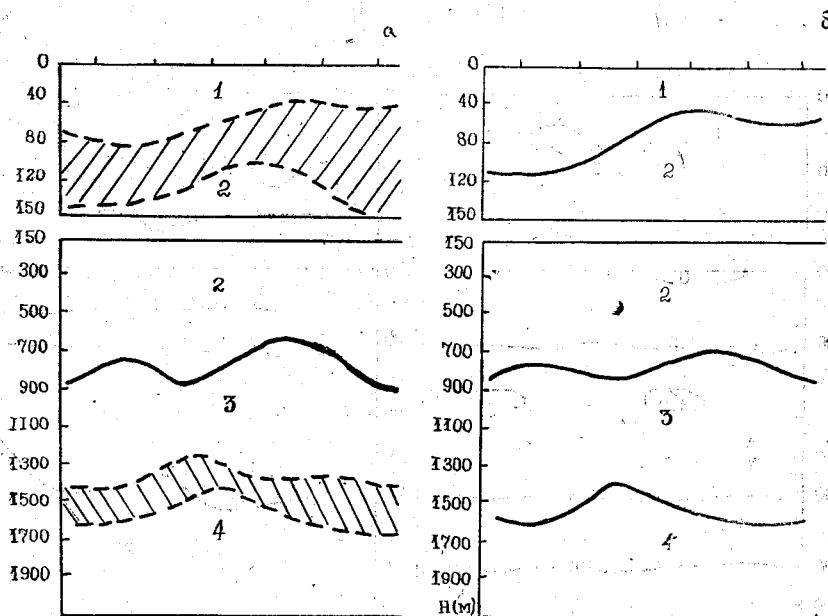


Рис 6.8. Распределение водных масс по результатам T, S -анализа (а) и автоматической классификации (б). 1—4 — водные массы

Таблица 6.3

Типичные значения характеристик выделенных водных масс

Водная масса	Среднее значение признаков							Тип водной массы
	$T^{\circ}\text{C}$	$S\text{‰}$	ρ	$\text{O}_2\text{мл/л}$	$\text{O}_2\%$	Pмкг/л	Si мкг/л	
1	19.3	36.4	25.7	5.6	102	339	7.2	Поверхностная
2	14.0	35.8	27.1	4.7	88	423	24.3	Подповерхностная
3	10.6	36.2	27.7	4.3	69	581	42.6	Средиземная морская
4	5.9	35.7	27.9	5.1	71	766	51.9	Глубинная

Вертикальные распределения водных масс, полученные в результате T, S -анализа и автоматической классификации в целом имеют много общего. Однако привлечение гидрохимических характеристик в процессе автоматической классификации позволило существенно уточнить положение границ поверхностной и глубинной водных масс. Таким образом, применение одного из методов многомерной статистики дает возможность повысить объективность и достоверность результатов выделения водных масс.

Другое, весьма перспективное направление использования МКА состоит в физико-географическом районировании океана. Рассмотрим этот вопрос на примере районирования Мирового океана по характеру межширотной и внутригодовой изменчивости испарения.

Районирование — это территориальная классификация, направленная на выделение некоторых однородных (естественных) районов по признаку их сходства. Таким образом, процедура районирования состоит в перенесении результатов классификации в пространственные географические координаты. При этом число районов может быть больше числа выделенных классов. Отсюда следует, что районирование можно рассматривать как дополнительный этап процесса классификации.

Итак, поскольку нашей задачей является построение целевой классификации, то исходные признаки определяются заранее. Это средние месячные значения испарения в десятиградусных зонах Мирового океана. В результате имеем исходную матрицу размером 12×14 , где 14 — число десятиградусных зон от 70°с.ш. до 70°ю.ш.

Разложение данной матрицы с помощью МФА показало, что первые два фактора описывают 90 % дисперсии. При этом первый фактор хорошо интерпретируется как центр тяжести внутригодовых колебаний (математическое ожидание), т. е. он характеризует межширотную (меридиональную) изменчивость исходных данных. Действительно, коэффициент корреляции между f_1 и средним го-

довым испарением E составляет $r(f_1, \bar{X}_E) = 0.90$. Второй фактор интерпретируется как изменчивость внутригодовых колебаний $r(f_1, \sigma_E) = 0.86$. Это по существу означает, что классификация может выполняться непосредственно в пространстве $\bar{X}_E$ и σ_E без предварительного применения методов сжатия информации.

Для классификации испарения использовалась обычная евклидова метрика и алгоритм КСС с последующей оптимизацией внутриклассовых и межклассовых расстояний. График КСС приводится на рис. 6.9. Оптимальным в смысле критерия (6.5) является разбиение всей совокупности точек на три класса. При этом роль обоих признаков в формировании классов примерно одинакова. Если разделение первого и второго классов обусловлено межширотной изменчивостью испарения, то разделение второго и третьего классов уже связано с σ_E .

При переносе результатов классификации на карту получим четыре района, так как второй класс не имеет общих границ и поэтому распадается на два района. Самый обширный квазиоднородный район составляют низкие широты в пределах от 40°N до 40°S . Здесь испарение достигает максимальных значений, но внутригодовая изменчивость его невелика. К этому району примыкают два района ($60^\circ - 40^\circ\text{N}$ и $60^\circ - 40^\circ\text{S}$), составляющие один класс и, следовательно, характеризующиеся примерно одинаковыми межширотной и внутригодовой изменчивостью испарения. Наконец, зона $70^\circ - 60^\circ\text{N}$ характеризуется максимальной внутригодовой изменчивостью испарения.

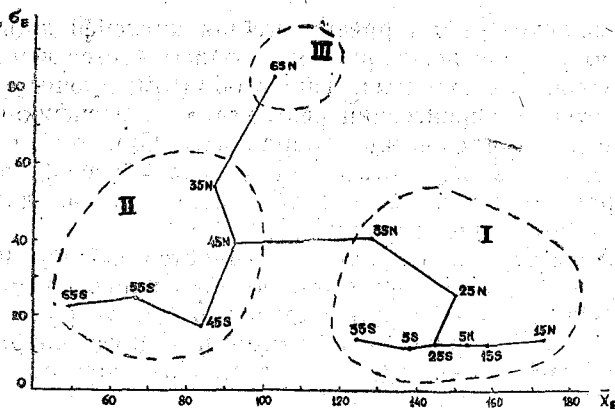


Рис. 6.9. Классификация зонально-осредненных величин испарения с акватории океана в пространстве $\bar{X}_E$ и σ_E

ГЛАВА 7. МЕТОД ДИСКРИМИНАНТНОГО АНАЛИЗА

7.1. Постановка задачи

Процесс накопления новых знаний об океане неизбежно связан с соотнесением новой информации к уже накопленным и обработанным материалам. Обработанная информация об океане, как правило, представляет собой структуру классов или типов явлений и процессов, которые выявлены в результате исследования всего имеющегося архива данных. Полученные типы представляют собой наиболее вероятные, устойчивые состояния исследуемых объектов, для которых сформулированы физические законы возникновения, существования и взаимоперехода.

Собственно, объективное соотнесение новой информации с уже классифицированными данными и является основной задачей применения метода дискриминантного анализа (МДА). Сама методика при этом выступает как формальное математическое правило, согласно которому путем статистического сравнения характеристик исследуемого наблюдения и выделенных ранее классов дается оценка принадлежности данного вектора наблюдения к тому или иному классу. Применение МДА позволяет свести к минимуму возможные ошибки неправильного определения принадлежности наблюдения к выделенным классам, что обеспечивает эффективность практического применения методики.

Область знания, описываемая здесь с позиции МДА, называется «распознаванием образов». Статистические характеристики выделенных ранее классов представляют собой образы, на соответствие которым анализируется образ нового вектора наблюдения, составленный из аналогичных статистик.

Примером использования теории распознавания образов в океанологических исследованиях может служить проблема типизации вертикального распределения плотности воды в Мировом океане. Акватория океана уже достаточно подробно районирована по типам вертикального распределения плотности воды. Для каждого региона определены статистические характеристики распределения плотности (среднее, дисперсия, асимметрия). Кроме того, увязан с распределением плотности и комплекс других параметров состояния океана: термохалинный режим, динамика вод, оптические и акустические характеристики, распределение биологических и химических показателей. И если удастся объективно определить принадлежность вновь наблюдаемого вертикального профиля плотности к одному из выделенных классов, значит удастся получить весь комплекс сопутствующей данному классу информации и разобраться в физической стороне протекающих процессов.

Перейдем непосредственно к формулировке задач метода дискриминантного анализа. В процессе применения МДА требуется решить две основные задачи:

— сформулировать математическое правило отнесения объектов к одному из выделенных классов;

— определить принадлежность нового вектора наблюдений к одному из выделенных классов.

В качестве исходных данных используется так называемая обучающая выборка X , состоящая из N наблюдений за M переменными. Наблюдения заранее разнесены по K классам. Для каждого класса уже вычислены первичные статистические характеристики: среднее для M параметров $\bar{x}_i^T = \{\bar{x}_{i1}, \bar{x}_{i2}, \dots, \bar{x}_{iM}\}$, дисперсии для M параметров $D_i^T = \{D_{i1}, D_{i2}, \dots, D_{iM}\}$, построены эмпирические кривые распределения $P_i(x_1, x_2, \dots, x_M)$. Обучающая выборка используется для нахождения решающего правила. Под правилом понимается математическое выражение — дискриминантная функция (Z), по величине которой можно сделать однозначный вывод о принадлежности вектора наблюдений к конкретному классу.

В рамках теории линейных статистических моделей принимается, что дискриминантная функция линейно зависит от исходных переменных. Тогда, рассматривая дискриминантную функцию $Z(X)$ как зависимую переменную в уравнении множественной линейной регрессии от набора переменных $\{x_{i1}, x_{i2}, \dots, x_{iM}\}$, можно записать

$$Z(x_i) = Z_0 + \sum_{j=1}^M b_j x_{ij} = z_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_M x_{iM}.$$

В этой записи основного уравнения МДА использованы следующие обозначения: $Z(x_i)$ — величина дискриминантной функции для i -го наблюдения; Z_0 — свободный член дискриминантной функции, характеризующий границу классов; $b_1, \dots, b_M$ — значения M коэффициентов дискриминантной функции; x_{ij} — соответствует i -му наблюдению за j -й исходной переменной.

7.2. Геометрическая интерпретация МДА

Рассмотрим геометрическую интерпретацию МДА на простом примере, когда обучающая выборка, состоящая из N наблюдений, проецируется в виде облака рассеяния в плоскости координат $x_1 \circ x_2$ (см. рис. 7.1). При этом имеются два отчетливо выделяемых облака данных, которые априорно объединены в два класса: № 1 и № 2. Для каждого класса вычислены значения величин дисперсии и математического ожидания по каждому из двух признаков. Кроме того, для каждого из признаков построены обобщенные кривые распределения облаков рассеяния $P(x_1)$ и $P(x_2)$. Нетрудно видеть, что часть наблюдаемых значений попадает в область пересечения кривых, внутри которой определить принадлежность наблюдений конкретным классам не представляется возможным.

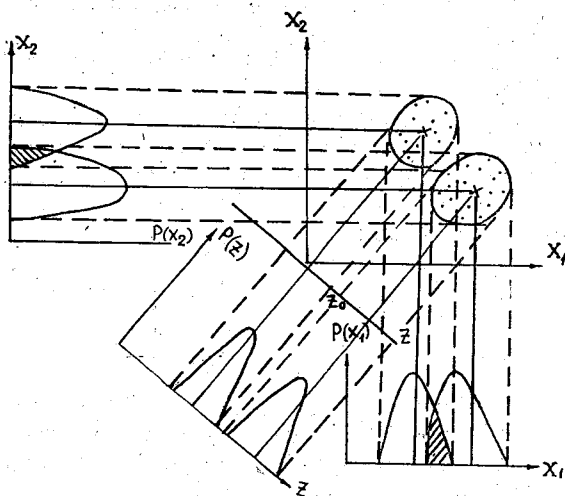


Рис. 7.1. Геометрическая интерпретация дискриминантного анализа

В процессе вычислений исследователь подбирает как бы новую ось измерения наблюдений, в проекции на которую классы будут разделены наиболее отчетливо. Эта ось или координата и называется линейной дискриминантной функцией $Z(X)$. В проекции на $Z(X)$ каждое наблюдение x_i получает дискриминантную метку $Z(x_i)$, вычисленную с помощью основного уравнения МДА. Аналогичные метки получают центры классов $\bar{Z}(X_I)$, $\bar{Z}(X_{II})$ и граница разделения классов Z_0 . Теперь, чтобы определить принадлежность любого наблюдения x_i к одному из имеющихся классов, достаточно вычислить его дискриминантную метку $Z(x_i)$ и сравнить ее с индексом Z_0 : при $Z(x_i) > Z_0$ наблюдение относится к классу 1, при $Z(x_i) < Z_0$ — к классу 2.

Важнейшей характеристикой класса является его центр, который рассматривается как среднее значение всех наблюдений в классе. В проекции на дискриминантную функцию центры классов оказываются максимально удаленными друг от друга $|Z(X_I) - \bar{Z}(X_{II})| \rightarrow \max$, хотя компактность распределения наблюдений внутри классов при этом остается неизменной и геометрическая форма классов не меняется. Действительно, как видно из рис. 7.1, центры классов, спроецированные на дискриминантную ось, дают наибольшее расстояние по сравнению с другими осями. Кроме того, по аналогии с МГК, $Z(x)$ можно трактовать как ось первой главной компоненты для ковариационной матрицы различий классов.

7.3. Вычислительная процедура МДА

Рассмотрим одну из наиболее распространенных вычислительных схем дискриминантного анализа, в основу которой положен хорошо известный алгоритм расчета собственных векторов ковариационной матрицы признаков. При поиске значений коэффициентов дискриминантной функции $B^T = \{b_1, b_2, \dots, b_M\}$ в этом алгоритме реализуется условие максимизации разности средних значений между классами:

$$(\bar{X}_I - \bar{X}_{II}) = \left\| \begin{array}{c} \bar{x}_{11} - \bar{x}_{12} \\ \bar{x}_{21} - \bar{x}_{22} \\ \vdots \\ \bar{x}_{M1} - \bar{x}_{M2} \end{array} \right\| \rightarrow \max,$$

где $\bar{x}_{ij}$ — среднее значение i -го признака в j -м классе.

Также максимизируется отношение межклассовой (C) и внутриклассовой (V) дисперсии наблюдений:

$$\frac{C}{V} \rightarrow \max.$$

В процессе вычислений можно выделить несколько этапов. При этом для простоты записи рассмотрим построение дискриминантной функции на примере двух классов с N_I и N_{II} наблюдениями в каждом.

1. Для каждого класса рассчитывается внутриклассовая ковариация исходных переменных c_{ij} . Для I-го класса эта величина находится из формулы

$$c_{ij} = \frac{1}{N_I} \sum_{p=1}^{N_I} (x_{pi} - \bar{x}_i)(x_{pj} - \bar{x}_j),$$

где $\bar{x}_i, \bar{x}_j$ — средние значения i -го и j -го признаков для наблюдений в I-м классе, x_{pi}, x_{pj} — значения i -го и j -го признака в p -м наблюдении I-го класса. Из полученных значений c_{ij} составляется матрица ковариации признаков I-го класса C_I :

$$C_I = \left\| \begin{array}{cccc} c_{11} & c_{12} & \dots & c_{1M} \\ c_{21} & c_{22} & \dots & c_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \dots & c_{MM} \end{array} \right\|,$$

которая состоит из M строк и столбцов. Аналогично составляется матрица внутриклассовой ковариации II-го класса — C_{II} .

2. Оценивается обобщенная для всей выборки X из N наблюдений матрица внутриклассовой ковариации C

$$C = (C_I N_I + C_{II} N_{II}) / N.$$

Очевидно, что при условии взаимонезависимости исходных переменных внедиагональные элементы матрицы становятся нулями, а на диагонали выстраиваются дисперсии переменных.

3. Вычисляется межклассовая дисперсия v_{ij} по формуле

$$v_{ij} = (\bar{x}_{iI} - \bar{x}_{iII})(\bar{x}_{jI} - \bar{x}_{jII}),$$

где $\bar{x}_{iI}$, $\bar{x}_{jI}$ — средние значения i -го и j -го признаков в I-м классе, $\bar{x}_{iII}$, $\bar{x}_{jII}$ — то же во II-м классе. Затем формируется матрица V , содержащая M строк и столбцов:

$$V = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1M} \\ v_{21} & v_{22} & \dots & v_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{M1} & v_{M2} & \dots & v_{MM} \end{bmatrix}.$$

4. Составляется квадратная матрица вида $(V - \Lambda C)$, определитель которой приравняется нулю:

$$|V - \Lambda C| = 0.$$

Решение получаемого характеристического уравнения относительно неизвестных Λ позволяет найти M собственных чисел $\lambda_1, \lambda_2, \dots, \lambda_M$.

5. Составляется система уравнений вида

$$B_1 \| V - \Lambda C \| = 0,$$

которая решается относительно неизвестного первого собственного вектора $b_{11}, b_{12}, b_{13}, \dots, b_{1M}$ путем подстановки найденного первого собственного числа λ_1 . Вычисляемые значения собственного вектора и являются искомыми коэффициентами дискриминантной функции $Z(X)$.

Очевидно, что полученные таким образом коэффициенты $b_{11}, \dots, b_{1M}$ определяют положение $Z(X)$ как новой координатной оси, в проекции на которую в данном признаковом пространстве будет максимизирована дисперсия расстояний между центрами классов.

6. Вычисление дискриминантного индекса границы классов Z_0 проводится по формуле

$$Z_0 = \frac{\bar{Z}(x_I)N_I + \bar{Z}(x_{II})N_{II}}{N},$$

где $\bar{Z}$ — среднее значение дискриминантной функции наблюдений I-го и II-го классов соответственно.

7.4. Применение дискриминантного анализа

Наиболее широкое применение в океанологии техника МДА нашла для решения задач прогнозирования гидрометеорологиче-

ских явлений, которые разбиваются на несколько классов (градаций). Использование дискриминантных функций позволяет отобрать наилучшие прогностические переменные, в терминах которых прогнозируемое явление по каждой градации предсказывается с наибольшей достоверностью.

Так, В. В. Оганесян показал, что применение МДА позволяет надежно прогнозировать ледовые условия в Аральском море. Сроки появления льда (L) исторически сгруппированы в 3 градации: 1-й класс — раннее появление, 2-й класс — нормальное и 3-й класс — позднее появление льда. Сложность процессов взаимодействия океана и атмосферы столь высока, что пока не удалось однозначно статистически описать взаимосвязи термодинамических характеристик в процессе формирования ледовых условий. Поэтому задача применения МДА заключалась в оптимизации многочисленных известных прогностических связей температуры воздуха (T_a), атмосферного давления (P) и классов появления льда. Многочисленные эксперименты показали, что наилучшее разбиение ряда L на три непересекающихся класса производится в пространстве следующих переменных: температура воздуха в г. Аральске за октябрь месяц ($\bar{T}_a$) и две первые главные компоненты (P_1 и P_2) разложения сентябрьского поля атмосферного давления над исследуемой акваторией. Полученные дискриминантные функции приближенно имеют вид

$$Z_{I-II} = 2.54 + 0.10P_1 - 0.07P_2 - 0.36\bar{T}_a$$

— (для разделения классов раннего и нормального ледообразования) и

$$Z_{II-III} = 3.45 + 0.02P_1 - 0.04P_2 - 0.32\bar{T}_a$$

— (для разделения классов нормального и позднего ледообразования). Проекция имеющихся наблюдений из выбранного 4-мерного пространства признаков на полученные дискриминантные функции обеспечивает максимальное взаимное разделение индексов центров классов ($\bar{Z}(x_I) - \bar{Z}(x_{II})$, $\bar{Z}(x_{II}) - \bar{Z}(x_{III})$), что указывает на минимальность ошибки неверной классификации. Чтобы оценить будущие сроки появления льда, достаточно рассчитать обе дискриминантные функции по фактическим значениям $\bar{T}_a$ и P_1 , P_2 . А затем воспользоваться общими правилами: если $Z_{I-II} \geq 0$, то прогнозируется раннее ледообразование, если $Z_{I-II} < 0$ и $Z_{II-III} \geq 0$, то ожидаются нормальные сроки появления льда, если $Z_{I-II} < 0$ и $Z_{II-III} < 0$, то ледообразование будет поздним.

Рассматривая проблему применения МДА в океанологии, необходимо отметить, что она еще далеко не полностью раскрыта. Из теоретических соображений ясно, что обсуждаемый выше линейный алгоритм дискриминантного анализа при всех его достоинствах не лишен недостатков, общих для линейных методов преобразования данных (см. разд. 2.3, 3.5, 4.4). Поэтому при наличии

сложной структуры наблюдений (значительного пересечения классов, нелинейной формы классов в пространстве признаков и т. п.) применение классического линейного МДА может не дать эффективных результатов.

Известны алгоритмы нелинейного (квадратичного) решения подобных вопросов. Однако сложность вычислительных процедур сдерживает применение последних в практической океанологии.

ЛИТЕРАТУРА

1. Айвазян С. А. и др. Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
2. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. — М.: Финансы и статистика, 1986. — 232 с.
3. Белкин И. М. Морфолого-статистический анализ стратификации океана. — Л.: Гидрометеониздат, 1991. — 133 с.
4. Бронфман А. М. и др. Статистическая структура океанологических и биологических параметров экосистемы Азовского моря. — Ростов-на-Дону, изд. РГУ, 1979. — 160 с.
5. Девис Д. С. Статистический анализ данных в геологии. Т. 2. — М.: Недра, 1990. — 427 с.
6. Енюков И. С. Методы, алгоритмы и программы многомерного статистического анализа. — М., Финансы и статистика, 1986. — 232 с.
7. Иберла К. Факторный анализ. — М., Статистика, 1980. — 398 с.
8. Львовский Е. Н. Статистические методы построения эмпирических формул. — М., Высшая школа, 1982. — 224 с.
9. Николаев Ю. В. Классификация гидрометеорологических процессов с помощью ЭВМ. — Л., Гидрометеониздат, 1976. — 36 с.
10. Петрович М. Л. Регрессионный анализ и его математическое обеспечение на ЕС ЭВМ. — М.: Финансы и статистика, 1989. — 299 с.
11. Смирнов Н. П., Вайновский П. А., Титов Ю. Э. Статистический анализ и прогноз океанологических процессов. — СПб.: Гидрометеониздат, 1992. — 198 с.
12. Справочник по прикладной статистике. Т. 2. — М.: Финансы и статистика, 1990. — 526 с.
13. Шелутко В. А. Численные методы в гидрологии. — Л.: Гидрометеониздат, 1991. — 238 с.

ОСНОВНЫЕ ПОНЯТИЯ

Матрица — массив (таблица) чисел из нескольких строк и столбцов. Каждое число из этого массива рассматривается как элемент матрицы.

Порядок матрицы означает количественное выражение ее размера, которое состоит из количества строк и количества столбцов. Так, обозначение $_{10}X^4$ или $X_{(10 \times 4)}$ соответствует матрице X , имеющей 10 строк и 4 столбца. Элемент матрицы, находящийся в i -й строке и j -м столбце, обозначается как x_{ij} .

Пример

$$_3X^4 = X_{(3 \times 4)} = \begin{bmatrix} 3 & 2 \\ 6 & 7 \\ 9 & 4 \end{bmatrix}, \quad \begin{aligned} x_{11} &= 3, & x_{21} &= 2. \\ x_{21} &= 6, & x_{22} &= 7. \\ x_{31} &= 9, & x_{32} &= 4. \end{aligned}$$

Вектор — это форма вырожденной матрицы, которая имеет либо одну строку — вектор-строку $(x)^T$, либо один столбец — вектор-столбец (x) .

Транспонированная матрица — это исходная матрица, повернутая на 90° так, что строки становятся столбцами, а столбцы — строками. Операция транспонирования обозначается значком $(^T)$.

Пример

$$X_{(3 \times 2)} = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}, \quad X_{(2 \times 3)}^T = \begin{bmatrix} 3 & 9 & 6 \\ 7 & 4 & 1 \end{bmatrix}.$$

ЭЛЕМЕНТАРНЫЕ ОПЕРАЦИИ НАД МАТРИЦАМИ

Равенство матриц. Две матрицы считаются равными, если совпадают их порядки и попарно равны соответствующие элементы этих матриц.

Сложение и вычитание матриц осуществляется путем поэлементного сложения (вычитания) соответствующих членов матриц. При этом необходимо выполнение равенства порядков матриц.

Пример

$$X_{(3 \times 2)} = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}, \quad Y_{(3 \times 2)} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 1 \end{bmatrix}.$$

$$X_{(3 \times 2)} + Y_{(3 \times 2)} = Z_{(3 \times 2)} = \begin{bmatrix} 3+2 & 7+1 \\ 9+3 & 4+4 \\ 6+5 & 1+1 \end{bmatrix} = \begin{bmatrix} 5 & 8 \\ 12 & 8 \\ 11 & 2 \end{bmatrix}.$$

$$X_{(3 \times 2)} - Y_{(3 \times 2)} = T_{(3 \times 2)} = \begin{bmatrix} 3-2 & 7-1 \\ 9-3 & 4-4 \\ 6-5 & 1-1 \end{bmatrix} = \begin{bmatrix} 1 & 6 \\ 6 & 0 \\ 1 & 0 \end{bmatrix}.$$

В символьной форме операция сложения записывается как

$$z_{ij} = x_{ij} + y_{ij},$$

$$t_{ij} = x_{ij} - y_{ij}.$$

Очевидно, что порядок результирующей матрицы (Z и T) совпадает с порядком исходных матриц.

Умножение матриц. Операция умножения возможна только при выполнении следующего условия: число столбцов левой матрицы (X) будет равно числу строк правой матрицы (Y), то есть, если даны ${}_N X^M$ и ${}_L Y^k$, то операция ${}_N Z^k = {}_N X^M \cdot {}_L Y^k$ возможна при условии $M=L$. Важность этого условия становится очевидной из символьной записи операции умножения:

$$z_{ij} = \sum_{p=1}^M x_{ip} y_{pj}.$$

Пример

$${}_3X^2 \cdot {}_2Y^3 = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 3 & 5 \\ 1 & 4 & 1 \end{bmatrix} = \begin{bmatrix} (3 \cdot 2 + 7 \cdot 1)(3 \cdot 3 + 7 \cdot 4)(3 \cdot 5 + 7 \cdot 1) \\ (9 \cdot 2 + 4 \cdot 1)(9 \cdot 3 + 4 \cdot 4)(9 \cdot 5 + 4 \cdot 1) \\ (6 \cdot 2 + 1 \cdot 1)(6 \cdot 3 + 4 \cdot 1)(6 \cdot 5 + 1 \cdot 1) \end{bmatrix} =$$

$${}_3Z^3 = \begin{bmatrix} 13 & 37 & 23 \\ 22 & 35 & 49 \\ 13 & 22 & 31 \end{bmatrix}.$$

В графической форме умножение представлено на рис. 1, где отчетливо видно, что количество строк в матрице Y всегда совпадает с количеством строк в левой матрице (X), а число столбцов в матрице Z всегда совпадает с числом столбцов в правой матрице (Y). Поэтому в матричном умножении используется термин « X умножить на Y справа» или « X умножить на Y слева». Результаты этих двух операций различны.

Пример

$${}_3X^2 \cdot {}_2Y^3 = {}_3Z^3 = \begin{bmatrix} 13 & 27 & 23 \\ 22 & 35 & 49 \\ 13 & 22 & 31 \end{bmatrix}.$$

$${}_2Y^3 \cdot {}_3X^2 = \begin{bmatrix} 2 & 3 & 5 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix} = \begin{bmatrix} (2 \cdot 3 + 3 \cdot 9 + 5 \cdot 6)(2 \cdot 7 + 3 \cdot 4 + 5 \cdot 1) \\ (1 \cdot 3 + 4 \cdot 9 + 1 \cdot 6)(1 \cdot 7 + 4 \cdot 4 + 1 \cdot 1) \end{bmatrix} =$$

$${}_2T^2 \cdot {}_3X^2 = {}_2T^2 = \begin{bmatrix} 63 & 31 \\ 45 & 24 \end{bmatrix}.$$

Определитель — это скаляр (число), который характеризует квадратную матрицу. Определитель матрицы второго порядка ${}_2T^2$ вычисляется по формуле

$$|T| = t_{11}t_{22} - t_{21}t_{12}.$$

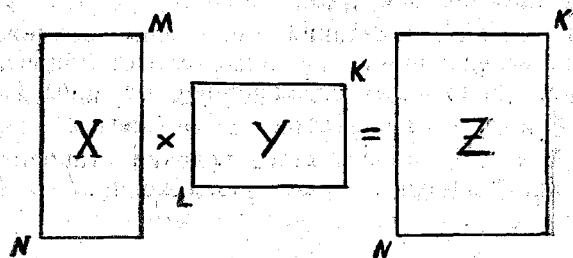


Рис. 1. Схема умножения матриц: X и Y — сомножители, матрица Z — результат умножения

Определитель матрицы третьего порядка вычисляется по формуле:

$$|T| = t_{11}t_{22}t_{33} - t_{11}t_{23}t_{32} - t_{12}t_{21}t_{33} + t_{12}t_{23}t_{31} - t_{13}t_{21}t_{32} - t_{13}t_{22}t_{31}.$$

Минор — это определитель особого рода. Минором T_{ij} из матрицы ${}_M T^M$ называется определитель квадратной матрицы порядка $(M-1)$, полученной из T вычеркиванием i -й строки и j -го столбца, причем знак минора определяется знаком выражения $-1^{(i+j)}$.

Пример

$${}_3 T^3 = \begin{bmatrix} 13 & 27 & 23 \\ 22 & 35 & 49 \\ 13 & 22 & 31 \end{bmatrix}, \quad |T| = -457, \quad T_{12} = - \begin{bmatrix} 22 & 49 \\ 13 & 31 \end{bmatrix} = -45.$$

Обращение матриц. Операции матричного деления не существует. Ей соответствует процедура обращения. Обратная матрица обозначается знаком $(^{-1})$: $T^{-1} = C$.

Обратная матрица удовлетворяет обязательному условию $TC = I$, где I — единичная квадратная матрица. Элементы обратной матрицы (c_{ij}) определяют с помощью соответствующих миноров:

$$c_{ij} = \frac{T_{ji}}{|T|}.$$

Из приводимой формулы следует, что при равенстве определителя нулю обратной матрицы не существует.

Пример

$${}_2 T^2 = \begin{bmatrix} 1 & -2 \\ 4 & 2 \end{bmatrix}, \quad |T| = 1 \cdot 2 - 4(-2) = 10,$$

$$C_{11} = T_{11}/|T| = 2/10 = 0.2, \quad C_{22} = T_{22}/|T| = 1/10 = 0.2,$$

$$C_{21} = T_{12}/|T| = -4/10 = -0.4, \quad C_{12} = T_{21}/|T| = 2/10 = 0.2,$$

$$TT^{-1} = C = \begin{bmatrix} 1 & -2 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} 0.2 & 0.2 \\ -0.4 & 0.1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I.$$

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ В МАТРИЧНЫХ ОБОЗНАЧЕНИЯХ

Последующие рассуждения и примеры будут проводиться для прямоугольной матрицы ${}_N X^M$, состоящей из N наблюдений (строк) за M переменными (столбцами).

Среднее арифметическое — это вектор $\bar{X}$, состоящий из M средних значений. Элементы вектора вычисляются по формуле

$$\bar{X}_j = \frac{1}{N} \sum_{i=1}^N x_{ij}.$$

В матричных обозначениях процедура вычисления вектора средних имеет вид

$$\bar{X}^T = I^T X / I I.$$

Вариация — это величина разности (y_{ij}) между конкретным наблюдением (x_{ij}) и средним значением данной переменной ($\bar{x}_j$):

$$y_{ij} = x_{ij} - \bar{X}_j.$$

Взяв из матрицы X строку наблюдений за M переменными X_i^T , можно получить вектор вариаций для i -го наблюдения:

$$y_i^T = \bar{x}_i^T - \bar{X}^T.$$

Последовательный перебор N наблюдений позволит получить матрицу вариаций ${}_N Y^M$

$$Y = X - \bar{X}^T I^T.$$

Пример

$${}_N X^M = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}, \quad \bar{X}^T = [1 \ 1] \begin{bmatrix} 3 & 7 \\ 9 & 7 \\ 6 & 1 \end{bmatrix} / [11] \begin{bmatrix} 1 \\ 1 \end{bmatrix} = [6 \ 4],$$

$$Y = \begin{bmatrix} -3 & 3 \\ 3 & 0 \\ 0 & 3 \end{bmatrix}.$$

Дисперсия может быть вычислена как средний квадрат вариаций:

$$D_j = \frac{1}{N} \sum_{i=1}^N y_{ij}^2.$$

что в векторном виде будет записано как

$$D_j = y_j^T y_j / I^T I.$$

Ковариация определяется как среднее произведение вариаций исследуемых переменных j и k :

$$D_{jk} = \frac{1}{N} \sum_{i=1}^N y_{ji} y_{ki}.$$

Соответственно ковариационная матрица будет вычисляться по формуле

$${}_M D^M = Y^T Y / I^T I.$$

Матрица D квадратная, симметричная, порядка M , на ее диагонали расположены дисперсии переменных.

Стандартизованные значения исходных данных (z_{ij}) получаются переводом их в вариации и последующим делением на стандартные отклонения:

$$z_{ij} = (x_{ij} - \bar{X}_j) / D_j.$$

Матрица стандартизованных данных ${}_N Z^M$ получается из формулы

$$Z = Y D_0^{-1/2},$$

где D_0 — диагональная матрица, составленная из стандартных отклонений переменных:

$$D_0 = \begin{bmatrix} \sqrt{D_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{D_2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sqrt{D_M} \end{bmatrix}.$$

Пример

$$X = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} -3 & 3 \\ 3 & 0 \\ 0 & -3 \end{bmatrix},$$

$$D_1' = [-3 \ 3 \ 0] \begin{bmatrix} -3 \\ 3 \\ 0 \end{bmatrix} / [1 \ 1 \ 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = (9 + 9 + 0) / 3 = 6,$$

$$D_2 = [3 \ 0 \ -3] \begin{bmatrix} 3 \\ 0 \\ -3 \end{bmatrix} / [1 \ 1 \ 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = (9 + 0 + 9) / 3 = 6,$$

$$D_0 = \begin{bmatrix} \sqrt{6} & 0 \\ 0 & \sqrt{6} \end{bmatrix} = \begin{bmatrix} 2.45 & 0 \\ 0 & 2.45 \end{bmatrix}, \quad |D_0| = 6.$$

$$Z = \begin{bmatrix} -3 & 3 \\ 3 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2.45/6 & 0 \\ 0 & 2.45/6 \end{bmatrix} = \begin{bmatrix} -1.23 & 1.23 \\ 1.23 & 0 \\ 0 & 1.23 \end{bmatrix}.$$

Матрица корреляции — это квадратная симметричная матрица R порядка M . Она составлена из коэффициентов корреляции r_{ij} , которые вычисляются как частное от деления ковариации на произведение стандартных отклонений:

$$r_{ij} = s_{ij} / (s_i \cdot s_j).$$

В матричном виде данная формула запишется следующим образом:

$$R = Z^T Z / N.$$

Пример.

$$Z = \begin{bmatrix} -1.23 & 1.23 \\ 1.23 & 0 \\ 0 & 1.23 \end{bmatrix}, \quad R = \begin{bmatrix} -1.2 & 1.2 & 0 \\ 1.2 & 0 & 1.2 \\ 0 & 1.2 & 0 \end{bmatrix} \cdot \begin{bmatrix} -1.2 & 1.2 \\ 1.2 & 0 \\ 0 & 1.2 \end{bmatrix} / N.$$

$$R = \begin{bmatrix} (-1.2^2 + 1.2^2 + 0) & (-1.2^2 + 0 + 0) \\ (-1.2 \cdot 1.2 + 0 + 0) & (1.2^2 + 0 + 1.2^2) \end{bmatrix} / 3,$$

$$R = \begin{bmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}.$$

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ МАТРИЦ

Значения элементов матрицы можно интерпретировать как координаты положения точек (или концов соответствующих векторов) в пространстве переменных или пространстве объектов.

Пространство переменных — это геометрическое пространство с ортогональными (взаимоперпендикулярными) осями, каждая из которых определяет значения только одной соответствующей ей переменной. Это пространство используется для интерпретации строк матриц. Элементы данной i -й строки определяют координаты конца i -го вектора наблюдений $\vec{X}_i$, проведенного из начала координат в пространстве, образованном осями M переменных.

Пространство объектов (наблюдений) — это геометрическое пространство с ортогональными осями, каждая из которых соответствует одному из объектов (наблюдений) в матрице данных. Переменная с номером j отображается в пространстве объектов как вектор $\vec{X}_j$ с координатами конца, соответствующими значениям j -го столбца матрицы данных (рис. 2).

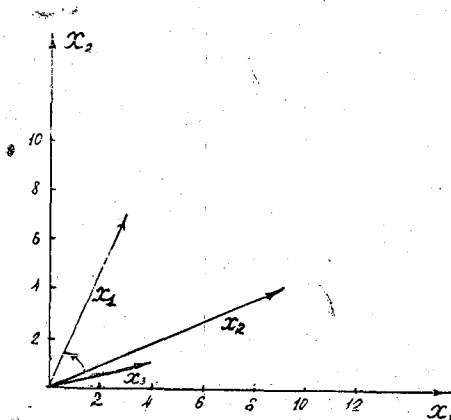


Рис. 2. Векторы наблюдений в пространстве переменных: исходная

матрица $X = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}$. Векторы наблюдений $\vec{x}_1^T = [3 \ 7]$, $\vec{x}_2^T = [9 \ 4]$, $\vec{x}_3^T = [6 \ 1]$

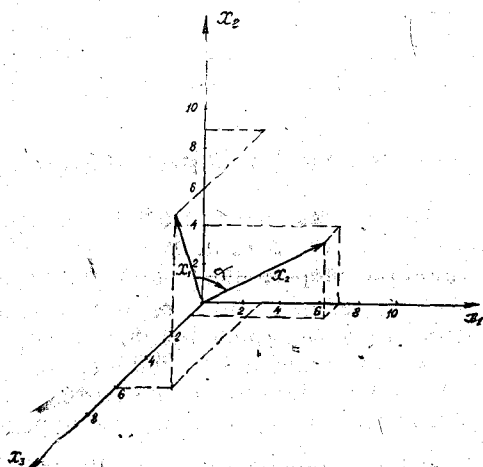


Рис. 3. Векторы переменных в пространстве объектов: исходная матрица $X = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}$. Векторы наблюдаемых переменных $x_1^T = [3 \ 9 \ 6]$, $x_2^T = [7 \ 4 \ 1]$.

Интерпретация корреляционной матрицы. Коэффициент корреляции r_{ij} характеризует линейную связь двух переменных i и j . Отсутствие связи приводит к нулевой корреляции. Линейная связь соответствует единичной корреляции (± 1). Геометрическая интерпретация коэффициента парной корреляции строится на основе принятия эквивалентности величины r_{ij} и косинуса угла α между векторами переменных в пространстве объектов (рис. 3).

Пример.

$$R = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}, \quad r_{12} = 0.5, \quad \alpha \cong 60^\circ.$$

Интерпретация матричного умножения. В ряде случаев умножение двух матриц (одна из которых квадратная) можно интерпретировать как вращение исходных векторов переменных или объектов относительно осей координат. В процессе вращения связность векторов не нарушается (то есть корреляции между векторами остаются неизменными), но изменяются углы наклона векторов к осям координат. Запишем матричное выражение вращения:

$$XT = Y,$$

где X — исходная матрица из N наблюдений за M переменными; T — квадратная симметричная матрица поворота порядка $M \times M$, Y — результирующая матрица наблюдений за M переменными после вращения.

В простейшем случае ортогонального вращения необходимо, чтобы столбцы матрицы поворота T были взаимортогональны и нормированы на дисперсию. В этом случае при вращении система координат останется ортогональной и масштабы ее осей не будут изменены. Данному свойству удовлетворяет, например, матрица вида

$$T = \begin{bmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{bmatrix}.$$

Геометрия подобного вращения приведена на рис. 4. Алгебра вращения получается после раскрытия операции умножения матриц:

$$Y = \begin{bmatrix} x_{11} \cos \alpha - x_{12} \sin \alpha & x_{11} \sin \alpha + x_{12} \cos \alpha \\ x_{21} \cos \alpha - x_{22} \sin \alpha & x_{21} \sin \alpha + x_{22} \cos \alpha \\ x_{31} \cos \alpha - x_{32} \sin \alpha & x_{31} \sin \alpha + x_{32} \cos \alpha \end{bmatrix}.$$

В случае, если столбцы матрицы поворота взаимозависимы, то происходит вращение векторов с итоговой проекцией их в систему новых координат, где оси координат будут взаимозависимы (коррелированы).

Если матрица поворота не нормирована на дисперсию, то после вращения окажутся измененными масштабы осей координат.

Пример

$$X = \begin{bmatrix} 3 & 7 \\ 9 & 4 \\ 6 & 1 \end{bmatrix}, \quad \alpha = 60^\circ, \quad T = \begin{bmatrix} \cos 60^\circ & \sin 60^\circ \\ -\sin 60^\circ & \cos 60^\circ \end{bmatrix} = \begin{bmatrix} 0.5 & 0.87 \\ -0.87 & 0.5 \end{bmatrix}.$$

$$XT = \begin{bmatrix} 7.6 & -0.9 \\ 8.0 & -5.8 \\ 3.9 & -4.7 \end{bmatrix} = Y.$$

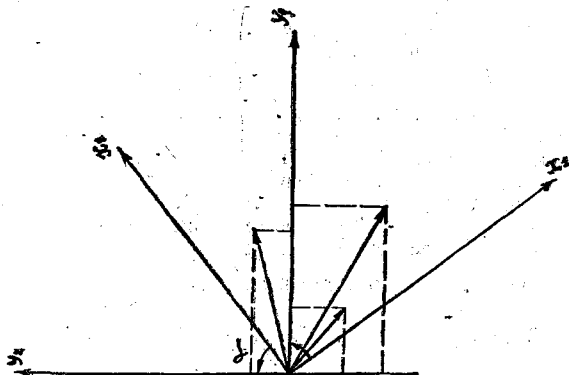


Рис. 4. Вращение векторов $\vec{X}_1$, $\vec{X}_2$, $\vec{X}_3$ в пространстве координат на угол $\alpha = 60^\circ$

ОГЛАВЛЕНИЕ

Введение	3
Глава 1. Многомерный статистический анализ (МСА) в океанологических исследованиях: общие проблемы применения	5
Multivariate statistical analyses (MSA) in oceanological researches: general problems of use	5
1.1. Понятие о методах МСА	5
1.2. Классификация методов МСА	9
Глава 2. Множественный регрессионный анализ	11
Multiple regression	11
2.1. Постановка задач	11
2.2. Математический аппарат МЛР	13
2.3. Коэффициент множественной корреляции и его свойства	17
2.4. Оценивание параметров регрессии	18
2.5. Понятие о частном коэффициенте корреляции	19
2.6. Понятие о нелинейной регрессии	21
2.7. О выборе оптимальных форм линейной регрессии	22
2.8. Применение МЛР в океанологических исследованиях	27
Глава 3. Метод главных компонент	30
Method of principal components	30
3.1. Постановка задачи	30
3.2. Математический аппарат МГК	31
3.3. Геометрическая интерпретация МГК	35
3.4. Свойства главных компонент	36
3.5. Проблемы применения МГК в океанологии	37
Глава 4. Метод факторного анализа	42
Factor analysis	42
4.1. Постановка задачи и основное уравнение	42
4.2. Геометрическая интерпретация МФА	44
4.3. Поиск факторного решения	46
4.4. Применение факторного анализа в океанологических исследованиях	49
Глава 5. Метод канонического корреляционного анализа	54
Canonical correlations	54
5.1. Постановка задачи	54
5.2. Геометрическая интерпретация МККА	56
5.3. Вычисление канонических корреляций	57
5.4. Примеры применения МККА	59
Глава 6. Метод кластерного анализа	60
Cluster analysis	60
6.1. Общие понятия. Постановка задачи	60
6.2. Выбор исходной системы признаков	63
6.3. Снижение размерности признакового пространства	63
6.4. Выбор меры близости	64
6.5. Алгоритмы классификации и методы оценки их качества	66
6.6. Использование кластерного анализа	71
Глава 7. Метод дискриминантного анализа	79
Discriminant analysis	79
7.1. Постановка задачи	79
7.2. Геометрическая интерпретация МДА	80
7.3. Вычислительная процедура МДА	82
7.4. Применение дискриминантного анализа	83
Литература	85
Приложение	86