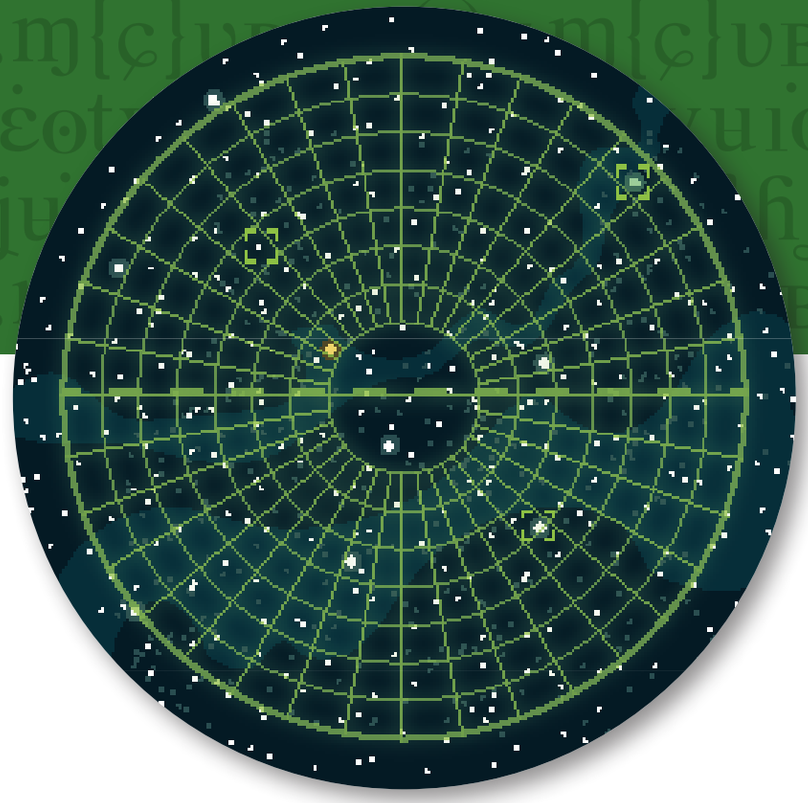
Изложен подход к использованию вычислительного вероятностного анализа для решения задач с неопределенными входными данными. Основное внимание уделено процессу обработки, представления, моделирования и анализа информации для разных типов неопределенности. Рассмотрены различные математические модели и численные методы их обработки, вопросы надежности результатов численного моделирования для разнообразных задач в условиях ограниченного и большого объемов информации. Даны примеры применения рассматриваемого подхода для практических задач цифровой экономики, надежности технических систем и оборудования. Разработанные алгоритмы могут быть использованы для исследования сложных систем с входными данными, обусловленными различными типами неопределенности.

ISBN 978-5-7638-4232-6

Б. С. Добронец, О. А. Попова

ВЫЧИСЛИТЕЛЬНЫЙ ВЕРОЯТНОСТНЫЙ АНАЛИЗ: МОДЕЛИ И МЕТОДЫ

СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
SIBERIAN FEDERAL UNIVERSITY



Б. С. Добронец
О. А. Попова

ВЫЧИСЛИТЕЛЬНЫЙ ВЕРОЯТНОСТНЫЙ АНАЛИЗ: МОДЕЛИ И МЕТОДЫ

Министерство науки и высшего образования Российской Федерации
Сибирский федеральный университет

Б. С. Добронец, О. А. Попова

Вычислительный вероятностный анализ: модели и методы

Монография

Красноярск
СФУ
2020

УДК 519.676
ББК 22.192.3
Д564

Р е ц е н з е н т ы:
К. В. Сафонов, доктор физико-математических наук, профессор, заведующий кафедрой прикладной математики СибГУ им. М. Ф. Решетнёва;
Г. А. Доррер, доктор технических наук, профессор, профессор кафедры системотехники СибГУ им. М. Ф. Решетнёва

Добронец, Б. С.
Д564 Вычислительный вероятностный анализ: модели и методы : монография / Б. С. Добронец, О. А. Попова. – Красноярск : Сиб. федер. ун-т, 2020. – 236 с.

ISBN 978-5-7638-4232-6

Изложен подход к использованию вычислительного вероятностного анализа для решения задач с неопределенными входными данными. Основное внимание уделено процессу обработки, представления, моделирования и анализа информации для разных типов неопределенности. Рассмотрены различные математические модели и численные методы их обработки, вопросы надежности результатов численного моделирования для разнообразных задач в условиях ограниченного и большого объемов информации. Даны примеры применения рассматриваемого подхода для практических задач цифровой экономики, надежности технических систем и оборудования. Разработанные алгоритмы могут быть использованы для исследования сложных систем с входными данными, обусловленными различными типами неопределенности.

Предназначена для магистрантов, аспирантов и специалистов, занимающихся научными исследованиями и работающих в области решения задач с неточными входными данными.

Электронный вариант издания см.:
http://catalog.sfu-kras.ru

УДК 519.676
ББК 22.192.3

ISBN 978-5-7638-4232-6

Оглавление

Введение

Монография посвящена вопросам исследования сложных систем на основе применения современных математических методов представления, численного моделирования и анализа в условиях различных видов неопределенности данных. Большинство компьютерных моделей для инженерных приложений разрабатываются для того, чтобы помочь оценить проектные или нормативные требования. В рамках этой задачи критически важна способность количественно оценить влияние изменчивости и неопределенности в контексте принимаемого решения. Вычислительная стоимость инженерных имитационных моделей довольно дорога: для моделирования с конечными элементами высокой точности может потребоваться несколько часов или дней, десятки процессоров. Таким образом, понимание того, как работают методы снижения уровня неопределенности и их относительные преимущества и затраты, очень важно.

В работе обсуждаются и находят дальнейшее развитие идеи, представленные в монографии [17], рассматриваются новые, активно развивающиеся направления анализа данных, такие как вероятностный анализ (probabilistic analysis), функциональный (functional analysis) и символьный анализ (symbolic analysis). Изучаются новые аспекты повышения точности и организации вычислительного процесса обработки и анализа данных, связанные с разработкой технологии быстрых и надежных вычислений.

Предлагаются новые методы и алгоритмы, учитывающие такие виды информационной неопределенности, как элиторная (aleatory uncertainty) и эпистемическая (epistemic uncertainty). Теория вероятностей предназначена для моделирования, оценки и оперирования именно элитерными неопределенностями. Элиторная неопределенность характеризует присущую случайность в поведении системы или в стадии ее изучения. Она включает в себя: изменчивость, стохастическую неопределенность. Примерами случайной неопределенности являются отказы компонентов си-

стемы, полученные в результате статистически значимых испытаний в условиях, относящихся к применению. Элиторные неопределенности характеризуются частотными распределениями.

В свою очередь, неопределённость самих вероятностных оценок называют эпистемической. Эпистемическая неопределённость прямо связана с объёмом и достоверностью информации, на основании которой получаются эти оценки [68].

Эпистемические неопределенности могут быть устранены путем более глубокого понимания (исследования), на основе увеличения объема данных или с помощью более новых достоверных предположений.

Проблема надежных вычислений сегодня выходит на передний план среди проблем вычислительной математики. Следует отметить, что значительную часть производимых сегодня в мире вычислений нельзя назвать надежными, поскольку методы обеспечения надежности еще не получили должного распространения, а после выполнения обычных вычислений пользователи не всегда могут получить убедительные аргументы относительно важнейших свойств полученного решения, в том числе и его точности.

Надежные вычисления (reliable computing) достигаются с учетом многих факторов, прежде всего оценками погрешности вычислительных алгоритмов и учетом неопределенностей входных данных. В этой связи важное значение приобретают апостериорные оценки погрешностей результатов численного моделирования [35]. Для практической реализации идеи повышения надежности вычислений важную роль сыграли достижения интервальной математики. Корректные интервальные вычисления гарантируют выполнение важнейших свойств численного решения и прежде всего — его локализацию.

В настоящее время актуализировалась проблема применения и разработки вычислительных технологий, реализующих технику быстрых и надежных вычислений для решения разнообразных практических задач, имеющих отношение к исследованию состояний и процессов функционирования сложных систем. Например, использование систем искусственного интеллекта в технике и других областях неизбежно приводит к необходимости обработки огромных массивов информации, поступающих в устройства. В этой связи специалисты по созданию интеллектуальных систем столкнулись с проблемой обработки данных объемов (big data). Отметим также задачи, которые решаются в рамках бизнес-аналитики, дистанционного мониторинга распределенных систем, робо-

тотехники, гидро- и атомной энергетики, при анализе отказов технических систем ответственного назначения, оценки и прогрнозирования техногенных, экологических, экономических и других видов рисков и т. д. Информация, которая составляет основу подобных задач, характеризуется имеющимся объемом данных, неоднородностью, динамичностью, уровнем и различными видами неопределенности.

Специфика сложности исследования таких систем обусловливается как объективными, так и субъективными аспектами. К объективным аспектам можно отнести следующие три группы факторов. Первая группа обусловливается внутренней сложностью системы как таковой. Вторая — внешней сложностью, непредсказуемостью, неопределенностью явлений и процессов, влияющих на систему и взаимодействующих с ней. Третья группа факторов связана с особенностями имеющейся у исследователя эмпирической информации и возможностями для ее обработки и анализа. Субъективный аспект связан прежде всего с тем, что практикам необходимо иметь определенный уровень доверия к применяемым математическим моделям и методам. Для них важно иметь убедительный ответ на вопрос, суть которого заключается в возможности получить достоверные, обоснованные результаты исследований, позволяющие установить с помощью численных расчетов достаточно полезную и реалистичную картину последствий принимаемых управленческих решений, несмотря на тот факт, что информация, на основе которой принимается решение, носит существенно неопределенный характер.

Обеспечение необходимой надежности и сложность исследования таких систем требует привлечения большого объема материальных, финансовых, интеллектуальных, временных, информационных и других ресурсов. При этом практика показывает, что привлекаемые ресурсы и вложения их в исследования не всегда пропорциональны требуемому уровню надежности и качеству функционирования систем в условиях различных видов неопределенности. Поэтому изучение способов и разработка новых моделей и методов представления информационной неопределенности в данных, обоснованное применение известных методов моделирования и разработка новых, реализующих перечисленные выше аспекты, представляет собой актуальную задачу.

Существующая неопределенность информации отражается в данных. Можно выделить три типа «неопределенных» данных: случайные, нечеткие и интервальные. Случайные числа задаются некоторыми вероятностными распределениями их возможных значений, нечеткие данные зада-

ются лингвистически сформулированными распределениями их возможных значений, интервальные данные задаются интервалами их возможных значений без указания какого-либо распределения внутри заданного интервала [101, 10, 50]. Изучение интервальной неопределенности способствовало созданию интервального анализа. Для случайной неопределенности знание законов распределения случайных величин позволяет оценивать параметры стохастических систем, используя метод Монте-Карло. Теория нечетких множеств широко используется для моделирования систем и принятия решений. В настоящее время для ряда задач в условиях стохастической неопределенности используется вычислительный вероятностный анализ [17, 31, 33, 71, 76].

В ряде случаев он успешно заменяет метод Монте-Карло [27, 39, 51], обладая значительно более высокой скоростью сходимости. В отличие от метода Монте-Карло он направлен на непосредственное построение распределений вероятности выходных переменных. Это существенно повышает качество полученных численных решений.

Для оценки качества решений важное значение имеет надежность полученных результатов. Любое измерение и методы его обработки содержат неточности. Рассмотрим последовательно этапы «эпохи» развития надежных вычислений. До «эпохи» надежных вычислений использовали «сырые данные» без предварительной обработки. Первый этап надежных вычислений заключался в статистической обработке и приближенном вычислении различных статистических характеристик. Ошибки численных методов приближенно оценивались с помощью двусторонних методов, например, правило Рунге, машинные арифметики не учитывали ошибки округления на компьютерах [21].

Второй этап — эра интервального анализа (ИА) началась с 50-х годов прошлого века. На этом этапе неопределенные данные представлялись в виде интервальных данных. Машинные арифметики, используемые в ИА, уже учитывали ошибки округления, а ошибки численных методов оценивались с помощью интервалов. ИА дает полностью гарантированные оценки, при этом значительно увеличивая время работы алгоритмов. К недостаткам интервального анализа можно отнести значительную ширину интервальных оценок по сравнению с оптимальными. ИА не использует информацию о возможных распределениях входных данных и соответственно не дает внутреннего распределения результатов вычислений, которые часто оказывались сосредоточенными только в небольших областях. Интервальные данные можно отнести к эпистимическому

типу неопределенности. Несмотря на указанные недостатки интервальный анализ позволяет эффективно решать многие практические задачи и широко используется при численном моделировании, например [24].

Третий этап — использование распределенных данных, в частности функций плотностей вероятности. Понятие распределенных данных — достаточно новое и появилось в научной литературе совсем недавно. Начало было положено разработкой численных операций над плотностями случайных величин, включая гистограммную арифметику. Одно из интересных представлений распределенных данных — символьные данные. Символьные данные были описаны Edwin и Diday в 1987 году [58]. Символьные переменные позволяют описывать группы индивидов и понятия. Символьные переменные включают списки значений (с весами или без них), интервальные переменные и даже гистограммы. Символьные представления могут включать внутреннюю структуру (иерархии) и логическую зависимость (правила).

Другой подход, при котором данные представляются в агрегированном виде, получил название Granular Computing (см., например, [112]). Информационные гранулы определяются, как группы отдельных наблюдений, которые отражают семантику абстрактных объектов, представляющих интерес. Как правило, с учетом набора данных D, в результате грануляции получается набор гранул, образованных на основе сходства или близости, которая может быть достигнута, например, с помощью алгоритмов кластеризации. Когда данные числовые, гранулы часто принимают форму гиперкубов. Информационные гранулы, описанные в теории нечетких множеств, представляются с помощью функции принадлежности. Распределенные переменные позволяют описывать каждую группу переменных посредством распределений. Распределения не используют статистические данные, такие как среднее, дисперсия, минимум и максимум и т. д. На практике сосредоточиваются на представлениях, которое лучше подходят для решения проблемы. Методы для распределенных данных включают в себя следующие разделы: описательная статистика, регрессия, кластеризация, уменьшение размерности, прогнозирование временных рядов, методы визуализации. Параллельно с символьным анализом развивается вычислительный вероятностный анализ (BBA).

Вычислительный вероятностный анализ разработан как новое направление в вычислительной статистике (Computational Statistics) и предназначен для решения практических задач, связанных с исследованиями сложных систем в условиях различных видов неопределенности и типов

эмпирических данных. Основой ВВА являются численные операции над плотностями случайных величин. В ВВА используются различные типы представления плотностей случайных величин: дискретные, гистограммы, полигоны, кусочно-полиномиальные модели и аналитическое представление. Использование порядковых статистик и случайных интерполяционных полиномов позволяет строить достоверные оценки функций распределения [36].

Одним из основных разделов вычисленного вероятностного анализа являются арифметики над данными, представленными в виде кусочно-полиномиальных функций.

В рамках ВВА реализуется подход, созвучный тезису «распределения — числа будущего» (Distributions are the Numbers of the Future), сформулированному в 1984 году B. Schweizer [126].

Суть данного подхода реализует идею представления эмпирических данных в виде функций распределений на основе применения кусочно-полиномиальных моделей. Разработанные в рамках ВВА численные арифметики и использование нового понятия «вероятностное расширение» позволили авторам разработать методы численного моделирования и анализа распределений, которые можно рассматривать как особый вид переменных, над которыми выполняются соответствующие операции и процедуры.

Отметим, что ВВА оперирует в первую очередь с понятиями функции плотности вероятности (ФПВ) и для изучения свойств изменчивости данных и разработки численных операций над ними использует ФПВ-представление в виде кусочно-полиномиальных моделей.

Применение его методов и процедур позволяет представить выходное распределение вероятностей как функцию входных распределений и использовать методы анализа неопределенностей, чтобы оценить влияние входных неопределённостей, привносимых входными характеристиками, на неопределенность выходных параметров модели.

Для построения распределений используются специальные способы агрегации данных [73, 74], рассматриваются задачи интерполяции [36], задачи оценки надежности.

В рамках ВВА решаются задачи случайной оптимизации [69, 113], повышения точности и оценок погрешности получаемых решений [35, 71].

Как показал анализ литературы, проблема снижения уровня неопределенности в исходных данных и повышения эффективности численных методов представления, обработки, моделирования и анализа в течение

многих десятилетий находится в центре внимания и остается предметом многих научных исследований. Наиболее значимые результаты в данной предметной области были получены учеными: B. Liu [95], S. Ferson [85, 86], A. Neumaier [107, 108], H. Schjaer-Jacobsen [125], D. Dubois [22], О. И. Ужга-Ребровым [42, 43, 44, 45, 46].

Актуализировалась проблема вычисления и анализа неопределённости выходных характеристик системы, индуцированных неопределённостями на её входах [19]. Среди публикаций, посвященных вопросам оценивания, анализа, управления неопределенностями следует указать на работы О. И. Ужга-Реброва. Можно выделить три основных группы методов, направленных на исследование и решение данной проблемы [42]: методы представления и оценивания неопределённости на выходе (выходах) модели, в зависимости от вида и уровня неопределённости на её входах (методы распространения неопределённостей); методы расчёта эффекта изменений на входах на предсказания модели, т. е. анализ чувствительности; методы сравнения важности входных неопределённостей в терминах их относительных вкладов на неопределённость на выходе (выходах), т. е. анализ неопределённостей.

Наиболее общий и распространенный метод включения неопределенности в моделирование состоит в том, чтобы предположить определенные распределения неопределенных входных значений, произвести выборку из этих распределений, запустить модель с выбранными значениями и делать это многократно, чтобы создать распределение выходных данных. Это классическое распространение неопределенности.

Для анализа и распространения элиторных и эпистемических неопределенностей в инженерных моделях используются также подходы: методы с использованием выборки на основе латинского гиперкуба (Latin Hypercube sampling), аналитические методы надежности (Analytic Reliability Methods) и методы разложения по полиномиальному хаосу (Polynomial Chaos Expansions) [78]. Эти методы являются альтернативными методами статистических испытаний, но опираются на аналитические вычисления и требуют от входных данных определенных свойств гладкости и принадлежности к гауссовым распределениям.

Методы построения выборки могут быть различными, например, можно использовать метод статистических испытаний (метод Монте-Карло), включая построение стратифицированной выборки (Latin Hypercube sampling), которая распределяет выборки по пространству, или квазивыборку, построенную методом Монте-Карло, который является способом гене-

рации последовательностей, приближающихся к равномерному распределению.

Отметим, что при решении проблемы снижения уровня выходной неопределенности важное значение имеет способ получения дополнительных оснований (знаний), снижающих уровень неопределенности во входных данных. Чтобы получить необходимые основания для оценки или восстановления неизвестного входного распределения на основе неполной, неточной информации, можно использовать различные процедуры и способы представления неопределенностей. Например, P-boxs [86], облака [108], интервальные гистограммы, гистограммы второго порядка [19].

Как один из подходов к идее «распространения неопределенности» использовался метод вероятностных границ (Probability bounds). Его основная идея в том, что функция неизвестного распределения вероятностей (Cumulative Distribution Function) должна лежать в обрасти — ящике (box), ограниченная нулем и единицей по вертикали и от минимума и максимума горизонтально. Истинная функция распределения, какой бы она ни была, должна находиться в этой области. Облака Неймайера (Neumaier'sclouds) являются еще одним способом представления неопределенности, выступая посредником между понятием нечеткого множества и вероятностным распределением [108].

В рамках основных подходов к распространению неопределенности следует указать также на математическую теорию очевидностей (свидетельств) Демпстера–Шафера [128], основанную на функции доверия (belief functions) и функции правдоподобия (plausible reasoning), которые используются, чтобы скомбинировать отдельные части информации (свидетельства) для вычисления вероятности события. Данная теория позволяет построить необходимые основания в условиях неопределенности путем оценки верхней и нижней границы интервала возможностей.

Среди подходов к распространению неопределенностей следует особенно выделить метод, который опирается на понятие «вероятность второго порядка», и известен как second-order probability. Данный подход представляет собой метод, позволяющий строить вероятностные оценки в случае эпистемистической неопределенности. Концепция вероятностей второго порядка была изложена в 1996 году в работах A. Mosleh и V. M. Bier. Анализ публикаций показал, что, несмотря на то, что данное направление достаточно активно развивается за рубежом, понятие «вероятность второго порядка» еще находится в стадии определения [103].

В монографии [17] предлагается новый способ представления данных в виде гистограмм второго порядка. Такой способ преобразования данных можно рассматривать в контексте проблемы распространения неопределенности и эффективно применять, когда законы распределения вероятностей зависят от неопределенных параметров.

Далее рассмотрим более подробно содержание глав монографии. Отметим, что при составлении содержания монографии и написания ее текста авторы исходили из того, что при работе с эмпирическими данными процесс их исследования представляет собой последовательность взаимосвязанных этапов, включающую методы предобработки, обработки и постобработки данных. Многообразие применяемых на каждом этапе методов, актуализирует проблему оценки точности полученных результатов. Ее решение во многом определяется надежностью тех вычислительных алгоритмов и методов, которые были выбраны для обработки, численного моделирования и анализа данных. Очевидно, что уже на стадии подготовки и преобразования данных необходимо применять процедуры представления данных в зависимости от имеющегося объема информации и типа неопределенности.

В главе 1 приводятся основные сведения из теории вероятностей.

В главе 2 рассматривается непараметрический подход, применяемый в настоящее время для оценки эмпирических функций распределений. Он имеет свои плюсы и минусы. Так, в отличие от параметрических методов не требует предположений о виде закона распределения наблюдаемых величин. Заметим, что во многих случаях достаточно сложно найти убедительные доказательства, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. В работе для решения задач анализа статистической информации развиваются методы ядерного оценивания и повышения их точности.

Главы 3, 4, 7 посвящены вопросам представления различных типов данных в виде математических моделей, приводятся примеры таких данных, рассматриваются вопросы построения и исследования функциональных (глава 3), символьных (глава 4) и временных рядов распределений (глава 7). В монографии большое внимание уделяется представлению и исследованию различных типов данных и выбору соответствующих процедур их обработки и анализа. Например, широко распространенными типами данных, которые в процессе наблюдения за объектом или объектами фиксируются непрерывно в течение определенного про-

межутка времени или периодически в дискретные моменты времени. Высокая внутренняя размерность этих данных создает проблемы как для теории, так и для вычислений, а их исследование требует применения специальных методов и подходов. Эти проблемы зависят во многом от того, как были собраны данные, какова структура и размерность данных, каковы их источники. Ответы на эти вопросы позволяют выявить новые направления исследований и анализа данных с целью разработки моделей и методов, учитывающих их особенности и повышающих надежность численных процедур обработки и анализа данных. Отмечается, что для изучения таких данных можно применять функциональный анализ данных (ФАД) (Functional Data Analysis) [102, 116, 117, 118, 119], который занимается анализом и теорией данных, представленных в виде некоторых функций, изображений или более общих объектов. Одним из основных понятий ФАД является понятие функциональных данных, которые представлены так, что для каждого субъекта в случайной выборке записывается одна или несколько функций.

Важно отметить, что идея представления эмпирических данных на основе применения математических моделей на этапе предобработки данных и последующего их использования в виде входных и выходных факторов для моделирования способствовала появлению особого вида переменных. Например, использование гистограммных моделей данных в виде входных переменных для регрессионного моделирования способствовало появлению нового понятия гистограммно-значные переменные, которые представляют собой особый вид переменных, где каждому такому объекту (признаку) соответствует распределение, которое может быть представлено в виде гистограммы. Такие переменные изучаются, например, в символьном анализе [90, 106, 122, 137] (глава 4). В последнее время наблюдается растущий интерес к моделированию и анализу интервально-значных и гистограммно-значных [63, 105, 106]. Однако анализ публикаций по данной теме исследований показал, что существующие методы и подходы к регрессионному моделированию на гистограммно-значных переменных встречают ряд трудностей [64]. Например, для линейных моделей регрессии для этого типа данных отмечается, что ее параметры не могут быть отрицательными. Для определения параметров этой модели необходимо решить квадратичную задачу оптимизации, при условии неотрицательности ограничений на неизвестных. Определенную проблему составляет задача выбора и вычисления меры погрешности между предсказанными и наблюдаемыми распределения-

ми. Избежать этих трудностей можно, используя численные операции над функциями плотностей, что как раз может быть успешно реализовано в рамках ВВА.

Глава 5 посвящена функциям от случайных аргументов, где определяется новое и одно из основных понятий ВВА — вероятностное расширение, здесь также рассматриваются вопросы построения надежных оценок для функций распределения.

В главе 6 обсуждаются вопросы использования численных вероятностных арифметик, реализующих численные операции над кусочно-полиномиальными представлениями функций плотности вероятности и вероятностных расширения для решения систем линейных и нелинейных алгебраических уравнений со случайными параметрами.

В главе 8 рассматривается применение ВВА к решению задач оптимизации со случайными входными параметрами (случайная оптимизация). В результате решения подобных задач методами математического программирования получаются оптимальные решения, зависящие от этих параметров. В тех случаях, когда известны плотности вероятности входных параметров, на основе вычислительного вероятностного анализа возможно построение совместной функции плотности вероятности оптимального решения. В отличие от стохастического программирования [129], где оптимальное решение представляет собой некоторое фиксированное решение, данный подход позволяет построить все множество решений оптимизационной задачи, определяемое построенной совместной функцией плотности вероятности. Методы, позволяющие строить множество решений оптимизационной задачи со случайными входными параметрами на основе применения численного вероятностного анализа, назовем случайным программированием.

В главе 9 исследуются вычислительные проблемы построения регрессионных моделей над эмпирическими распределениями. Исследуются вопросы агрегированного представления данных и методы построения регрессионных моделей с агрегированными входными параметрами и функциональными временными рядами. Изучаются различные метрики в пространстве распределений.

Глава 10 посвящена приложениям ВВА для практических задач. Например, рассматриваются основные вычислительные аспекты, характерные для задач цифровой экономики. Первый аспект связан с необходимостью обработки данных больших объемов. Для его реализации предлагается использовать процедуры агрегирования данных, основанные на

применении математических моделей представления данных. Переход к более обобщенному представлению с помощью агрегирования необходим по нескольким причинам. Во-первых, агрегация существенным образом может снизить объем данных. Во-вторых, детализированные данные часто оказываются очень изменчивыми из-за воздействия различных случайных факторов, разброса значений и поэтому слабо отражают общие тенденции и свойства исследуемого множества. Агрегация в этом случае позволяет увидеть имеющиеся тенденции и закономерности. Второй аспект связан с организацией вычислительного процесса, обеспечивающей необходимую для решения соответствующей практической задачи оперативность получения необходимой информации. Для преодоления этой проблемы предлагается использовать рекурсивную схему организации вычислительного процесса. Третий аспект отражает требование к достоверности полученных результатов моделирования, обеспеченных надежными вычислительными процедурами, адекватными тем типам неопределенности, которые содержатся в сырых данных.

Рассматривается задача построения достоверных оценок показателей надежности оборудования в условиях малых выборок статистических данных об отказах. Применение вычисленного вероятностного анализа (ВВА) позволяет получить гарантированные оценки показателей надежности функционирования технических объектов в условиях неопределенности и ограниченного объема информации.

Глава 1

Краткий обзор теории вероятностей

Основные понятия и определения теории вероятностей, необходимые для дальнейшего изложения, рассматриваются в этом разделе. Опираясь на работы Rudin [123], Loéve [96] и Rao и Swift [120], представлено краткое введение к основам теории вероятностей, а затем исследовано несколько важных понятий, таких как вещественные случайные величины и векторы, понятие моментных операторов и случайных процессов. Дальнейшие концепции теории вероятностей можно найти, например, в [134], [96] и [92].

1.1. Понятие измеримости

Класс непрерывных функций играет фундаментальную роль в топологической теории. Он имеет несколько элементарных свойств, общих с измеримыми функциями, которые играют важную роль в теории интегрирования. Далее будет представлен материал, подчеркивающий аналогии между понятиями топологического пространства, открытого множества и непрерывных функций с измеримыми пространствами, измеримыми множествами и измеримыми функциями. Здесь Ω определяется как непустое множество с конечным или бесконечным (счетным или несчетным) количеством элементов ω.

Определение 1 (топологическое пространство). *А топологией $\mathcal{F}$ на непустом множестве Ω является коллекция подмножеств Ω такая, что*

1) $\emptyset \in \mathcal{F}$, $\Omega \in \mathcal{F}$;

2) если $A_i \in \mathcal{F}$, $i = 1, 2 \ldots, n$ $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$;

3) если $A_\alpha \in \mathcal{F}$ для $\alpha \in \mathcal{A}$, для произвольного набора индексов $\mathcal{A}$, то $\bigcup_{\alpha\in\mathcal{A}} A_\alpha \in \mathcal{F}$,

где члены $\mathcal{F}$ называются открытыми множествами Ω, а упорядоченная пара $(\Omega, \mathcal{F})$ называется топологическое пространство.

Определение 2 (σ-алгебра и измеримое пространство). *Коллекция $\mathcal{F}$ подмножества непустого множества Ω называется σ-алгеброй Ω, если $\mathcal{F}$ удовлетворяет*

1) $\Omega \in \mathcal{F}$;

2) если $A \in \mathcal{F}$, то $\Omega \setminus A \in \mathcal{F}$;

3) если $\{A_n\}_{i=1}^n \subset \mathcal{F}$, то $\bigcup_{i=1}^n A_n \subset \mathcal{F}$,

в этом случае упорядоченная пара $(\Omega, \mathcal{F})$ называется измеримым пространством, а члены $\mathcal{F}$ называются измеримые множества в Ω.

Определение 3 (измеримая функция). *Пусть $(\Omega, \mathcal{F})$ и (Υ, Σ) — измеримые пространства. Тогда функция $\mu : \Omega \to \Upsilon$ измерима, если для каждого $A \in \Upsilon$, прообраз A под μ находится в $\mathcal{F}$, т. е.*

$$\mu^{-1}(A) \equiv \{\omega \in \Omega | \mu(\omega) \in A\} \subset \mathcal{F}.$$

Определение 4 (положительная мера и пространство меры). *Пусть $(\Omega, \mathcal{F})$ измеримое пространство. Функция $\mu : \mathcal{F} \to [0, \infty]$ называется положительной мерой, если μ удовлетворяет следующему.*

1) Неотрицательность: для всех $A \in \mathcal{F}$, $\mu(A) > 0$.

2) Пустое множество: $\mu(\emptyset) = 0$.

3) Счетная аддитивность: если $A_1, A_2, ... \in \mathcal{F}$ и $A_i \cap A_j = \emptyset$ для $i \neq j$, то

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Тройка $(\Omega, \mathcal{F}, \mu)$ называется измеримым пространством.

Замечание 1. *Измеримое пространство часто называют «упорядоченными тройками» $(\Omega, \mathcal{F}, \mu)$, где Ω — множество, $\mathcal{F}$ — σ-алгебра в Ω, а μ — мера, определенная на $\mathcal{F}$. Аналогично измеримые пространства часто называют «упорядоченными парами» $(\Omega, \mathcal{F})$.*

Эти соглашения имеют здравый смысл и являются логически правильными, несмотря на то, что они несколько избыточны. Например, с учетом вышеупомянутого упорядоченная пара, множество Ω является просто наибольшим членом $\mathcal{F}$; следовательно, учитывая $\mathcal{F}$, мы можем построить Ω. Более того, по определению каждая мера принимает σ-алгебру как его область, так что, учитывая меру μ, мы можем вывести σ-алгебру F, в которой μ определено, и мы также знаем множество Ω, в котором $\mathcal{F}$ является σ-алгеброй, поэтому допустимо использовать выражения «пусть μ будет мерой» или «пусть μ будет мера на Ω, если мы хотим подчеркнуть множество, или даже пусть μ будет мерой на $\mathcal{F}$, если мы хотим подчеркнуть σ-алгебру». Обычный подход, который логически довольно бессмысленен, это сказать «пусть Ω будет пространство меры», даже когда понятно, что есть мера, определенная на $\mathcal{F}$ в Ω, и это мера, которая нас математически интересует.

1.2. Борелевские σ-алгебры

σ-алгебра Бореля является важным примером σ-алгебры, которая используется в теории функций, интеграла Лебега и теории вероятности. Дадим определение и сформулируем классическую теорему о σ-алгебрах.

Теорема 1. *Пусть Ω — множество, а $\mathcal{V}$ — непустая совокупность подмножества Ω. В Ω существует наименьшая σ-алгебра, обозначаемая $\sigma(\mathcal{V})$ такой, что $\mathcal{V} \subset \sigma(\mathcal{V})$, а именно*

$$\sigma(\mathcal{V}) = \bigcap \{\mathcal{F} : \mathcal{F} \text{ является } \sigma\text{-алгеброй } \Omega, \mathcal{V} \subset \mathcal{F}\},$$

которая также называется σ-алгеброй, порожденной $\mathcal{V}$.

Теперь пусть Ω — топологическое пространство. По теореме 1 если $\mathcal{V}$ является набором открытых множеств (или, что то же самое, все замкнутые множества) Ω, то наименьшая σ-алгебра $\mathcal{B} = \sigma(\mathcal{V})$ называется борелевской σ-алгеброй на Ω. Элементы $B \in \mathcal{B}$ называются *борелевскими множествами*, которые могут быть сформированы из открытых множеств (или, что то же самое, из закрытых множеств) через операции счетного пересечения, счетного объединения и относительного дополнений. Поскольку $\mathcal{B}$ является σ-алгеброй, мы можем рассматривать $(\Omega, \mathcal{B})$ как измеримое пространство, где борелевские множества играют роль

измеримых множеств. Если $\mu: \Omega \to \Upsilon$ является непрерывным отображением Ω, где Υ — другое топологическое пространство, тогда из определений получаем, что $\mu^{-1}(A) \in \mathcal{B}$ для каждого открытого множества $A \in \Upsilon$.

В заключение, каждое непрерывное отображение Ω измеримо по Борелю. Измеримые по Борелю отображения часто называют борелевскими отображениями, или борелевскими функциями.

1.3. Вероятностные пространства и случайные величины

Вероятностная мера

По сути, вероятность является числовой мерой неопределенности результатов действия или эксперимента. Фактическое присвоение этих значений должно быть основано на опыте и поддаваться проверке при проведении эксперимента, если это возможно, повторяться при практически одинаковых условиях. Чтобы построить аксиоматическое представление, мы сначала представляем все возможные результаты эксперимента как отдельные точки непустого множества. С момента сбора все такие возможности могут быть бесконечно большими, различные комбинации их, полезные для экспериментов, необходимо учитывать. Затем мы определяем комбинации таких результатов как *события* и рассматриваем алгебру событий как первичные данные, которые включают в себя все мыслимое использование для эксперимента. Наконец, каждому событию присваивается числовая мера, которая соответствует «количеству» неопределенности таким образом, что эта неопределенность обладает аддитивными свойствами. Математически эта аксиоматическая формулировка была создана Колмогоровым.

Определение 5 (вероятностная мера и вероятностное пространство). *Пусть $(\Omega, \mathcal{F})$ — измеримое пространство, представляющее все возможные результаты эксперимента, где члены σ-алгебры $\mathcal{F}$, называемые событиями, являются коллекциями результатов эксперимента.*

$\mathsf{P}: \mathcal{F} \to [0,1]$ *называется вероятностная мера, или просто вероятность, если мера на $(\Omega, \mathcal{F})$ удовлетворяет условиям*

$$\mathsf{P}(A) > 0 \text{ для всех } A \in \mathcal{F}$$

и

$$\mathsf{P}(\Omega) = 1.$$

А пространство вероятностей является тройкой $(\Omega, \mathcal{F}, \mathsf{P})$*.*

Таким образом, вероятностное пространство — это пространство с конечной мерой, у которого функция меры нормируется так, чтобы мера всего пространства была равна единице. Пространство $(\Omega, \mathcal{F}, \mathsf{P})$ называется полное пространство вероятностей, если $\mathcal{F}$ содержит все подмножества A в Ω с внутренней мерой P:

$$\mathsf{P}^*(A) = \inf\{\mathsf{P}(F) : F \in \mathcal{F}, A \subset F\} = 0.$$

Подмножества A в Ω, которые принадлежат $\mathcal{F}$, называются $\mathcal{F}$-измеримыми. Однако в контексте вероятности интерпретации эти событий разные. Например, когда мы пишем $\mathsf{P}(A)$, что означает «вероятность того, что событие A произойдет». В частности, если $\mathsf{P}(A) = 1$, мы говорим, что «А происходит с вероятностью 1», или «почти наверняка».

Условная вероятность

Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ обозначает вероятностное пространство, и пусть A_1, $A_1 \in \mathcal{F}$ события с $\mathsf{P}(A_1) > 0$ и $\mathsf{P}(A_2) > 0$. Обозначим пересечение $A_1 \cap A_2$ A_1A_2. Тогда отношение $\mathsf{P}(A_1A_2)/\,\mathsf{P}(A_1)$ называется *условная вероятность* A_2 при заданном A_1, или просто вероятность A_2, заданная A_1, и обозначается через $\mathsf{P}(A_2|A_1)$, так что

$$\mathsf{P}(A_1A_2) = \mathsf{P}(A_1)\,\mathsf{P}(A_2|A_1). \tag{1.1}$$

Тогда по индукции для $A_1, A_2, ..., A_N \in \mathcal{F}$ получаем правило цепочки

$$P(\bigcap_{i=1}^{N} A_i) = P(A_1)P(A_2|A_1)\dots P(A_1A_2\dots A_{N-1}|A_N). \tag{1.2}$$

Более того, если $\cup_{i=1}^{N} A_i = \Omega$ с $A_i \cap A_j = \emptyset$ и A_i и $B \in \mathcal{F}$,

$$P(B) = P(\Omega B) = \sum P(A_iB). \tag{1.3}$$

Тогда правило полной вероятности следует из (A.1), а именно

$$P(B) = \sum P(A_i)P(B|A_i). \tag{1.4}$$

Наконец, используя (1.1)–(1.4), мы приходим к *теореме Байеса*:

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum P(A_i)P(B|A_i)}.$$

Случайные величины

Определение 6 (случайные величины). *Обозначим через* $(\Omega, \mathcal{F}, \mathsf{P})$ *пространство вероятностей. Функция* $X : \Omega \to R$ *является случайной величиной, если* X *удовлетворяет*

$$X^{-1}(\mathcal{B}) \subset \mathcal{F},$$

или

$$X^{-1}(A) = \{\omega \in \Omega | X(\omega) \in A\} \in \mathcal{F},$$

где $\mathcal{B}$ *— борелевская* σ*-алгебра* R*, и* $A = (-\infty, x)$*,* $x \in R$*.*

Таким образом, случайная величина — это функция из абстрактного множества Ω в вещественное пространство, где каждому результату $\omega \in \Omega$ назначается действительное число $X(\omega) \in R$.

Случайная величина представляет реальный интерес, когда она связана с ее мерой образа или с функцией распределения в контексте теории вероятностей.

Определение 7 (мера образа). *Пусть* $(\Omega, \mathcal{F}, \mathsf{P})$ *обозначает пространство вероятностей и* $X : \Omega \to R$ *обозначают случайную величину. Мера образа* X*, обозначаемая* P_X*, является мерой в борелевском пространстве* $(R, \mathcal{B})$*, определяется как*

$$\mathsf{P}_X(A) = \mathsf{P}(X^{-1}(A)) = \mathsf{P}(\{\omega \in \Omega | X(\omega) \in A\}),\ \forall A \in \mathcal{B},$$

где P_X *также является вероятностной мерой.*

Обратите внимание, что σ-алгебра $\{X^{-1} | A \in \mathcal{B}\}$ является подмножеством $\mathcal{F}$ и только характеризует вероятностные события, связанные со случайным вектором. Таким образом, σ-алгебру обычно называют σ-алгеброй, порожденной X, и обозначают $\sigma(X)$. Ограничение P на $\sigma(X)$, т. е. P_X, описывает только закон вероятности, связанный с X. Кроме того, P_X определяет уникальную функцию распределения в R.

Определение 8 (функции распределения). *Если* $X : \Omega \to R$ *— случайная величина на* $(\Omega, \mathcal{F}, \mathsf{P})$*, тогда ее функция распределения является отображением* $F_X : R \to R_+$*, определяется как*

$$F_X(x) = \mathsf{P}_X(X < x) = \mathsf{P}(\{\omega \in \Omega | X(\omega) \leq x\}),\ \forall x \in R,$$

непрерывно справа и монотонно возрастает и удовлетворяет

$$\lim_{x \to \infty} F_X(x) = 1,\ \lim_{x \to -\infty} F_X(x) = 0.$$

Из определений 7 и 8 мы видим, что случайная величина однозначно определяет меру образа P_X и функцию распределения F_X. И наоборот, F_X однозначно определяет меру P_X, но существует несколько случайных величин, имеющих одинаковую меру образа P_X. Эти случайные величины называются *одинаково распределенными* случайными величинами.

Определение 9 (функции плотности вероятности). *Пусть X обозначает случайная переменная в $(\Omega, \mathcal{F}, \mathsf{P})$, и пусть F_X будет ее функцией распределения вероятностей, которая абсолютно непрерывна в R. Тогда существует интегрируемая функция $f_X(x)$, называемая функцией плотности вероятности X, такая, что*

$$F_X(b) - F_X(a) = \int_a^b f_X(x)dx.$$

$f_X(x)$ фактически является производной Радона–Никодима от F_X.

1.4. Лемма Doob–Dynkin

Следующий результат измеримости чрезвычайно полезен; это особый случай результата обычно называют леммой Doob–Dynkin.

Лемма 1 (Doob–Dynkin). *Пусть $(\Omega, \mathcal{F})$ и $(\Theta, \mathcal{A})$ обозначают меру пространства и пусть $X : \Omega \to R$ — измерим. Тогда функция $Y : \Omega \to R$ является $\sigma(X)$-измеримым тогда и только тогда, когда существует функция $g : \Theta \to R$ такая, что $Y = g(X)$.*

Возможен ряд специализаций этого результата. Если мера пространства $\Theta = R$, и тогда мы будем использовать борелевскую σ-алгебру $\mathcal{A} = \mathcal{B}$ из R, существует измеримый по Борелю $g : R \to R$, который удовлетворяет требования. Это дает следующий результат.

Следствие 1. *Пусть Ω — измеримое пространство. Если $X, Y : \Omega \to R$ — две заданные измеримые функции, то Y является $\sigma(X)$-измеримым тогда и только тогда, когда существует измеримая по Борелю функция $g : R \to R$ такая, что $Y = g(X)$.*

Если $\mathcal{A}$ заменяется большей σ-алгеброй всех (пополнение $\mathcal{A}$) измеримых по Лебегу подмножеств R, то g будет измеримой по Лебегу функцией.

1.5. Интегрируемость и моменты случайных величин

Для случайной величины X: $(\Omega, \mathcal{F}) \to (R, \mathcal{B})$ интеграл от X по отношению к P по подобласти $D \subset \Omega$ определяется выражением

$$\int_D X(\omega)\,\mathsf{P}(d\omega) = \int_\omega \mathcal{I}_D(\omega) X(\omega)\,\mathsf{P}(d\omega),$$

где $\mathcal{I}_D(\omega)$ — характеристическая функция D. Если такой интеграл существует и конечный, то X является P-интегрируемой по D.

Моменты случайной величины

Определение 10 (математическое ожидание). *Пусть X обозначает случайную величину на вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Тогда*

$$\mathsf{E}(X) = \int_\omega X(\omega)\,\mathsf{P}(d\omega)$$

называется математическим ожиданием X. Если X является P-интегрируемым на Ω, то математическое ожидание X конечно.

Исходя из этого определения мы видим, что для любой измеримой по Борелю функции $Y : R \to R$, которая P_X-интегрируемая, мы имеем

$$\mathsf{E}(Y \circ X) = \int_R Y\, d\,\mathsf{P}_X\,.$$

В частности, когда $Y = X$, ожидание X может быть представлено интегралом

$$\mathsf{E}(X) = \int_R X\,\mathsf{P}_X(dX).$$

Определение 11 (пространство $L^q(\Omega)$). *В вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$ для $q > 1$ обозначим через $L^q(\Omega)$ набор случайных величин X, определенных на $(\Omega, \mathcal{F}, \mathsf{P})$, такой, что*

$$\mathsf{E}[|X|^q] \leq \infty.$$

Определение 12 (моменты порядка q). *Пусть X обозначает случайную переменную в вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$ и $Y = X^q$*

для $q > 1$. Математическое ожидание Y называется моментом порядка q X и определяется как

$$\mathsf{E}(X^q) = \int_\Omega X^q(\omega)\,\mathsf{P}(d\omega) = \int_R x^q dF_X(x),$$

где $x \in R$. Если $X \in L^q(\Omega)$, то его момент порядка q конечен.

1.6. Случайные векторы и их вероятностные распределения

Подобно определению случайной величины, случайный вектор, определенный на вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$, является вектором N случайных переменных, обозначаемым как $X(\omega) = (X_1(\omega), \ldots, X_N(\omega))$, чьи компоненты определяются в том же вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Мера образа X, обозначаемая P_X, является мерой в борелевском пространстве $(R^N, \mathcal{B}^N)$, определенной как

$$\mathsf{P}_X(A) = \mathsf{P}(X^{-1}(A)) = \mathsf{P}(\{\omega \in \Omega | X(\omega) \in A\}),\ \forall x \in R^N.$$

Определение 13 (функции совместного и предельного распределения). *Функция совместного распределения X является прямым расширением функции распределения скалярной случайной величины, т. е.*

$$F_X(x) = \mathsf{P}(\{\omega \in \Omega | X_1(\omega) \le x_1, \ldots, X_N(\omega) \le x_N\}) \forall x \in R^N.$$

Функция предельного распределения в X_n, обозначаемая как $F_{Xn}(x_n)$, определяется как

$$Fx_n(x_n) = F_X(\infty, \ldots, \infty, x_n, \infty, \ldots, \infty).$$

Определение 14 (функции совместной и предельной плотности). *Функция совместной плотности в X имеет вид прямого расширения определения функции плотности скалярной случайной величины, т. е. производной Радона–Никодима от Fx, представленная*

$$f_X(x) = \frac{\partial F_X(x)}{\partial x_1 \ldots \partial x_N}.$$

Функция предельной плотности для X_n, обозначаемая как $f_{X_n}(x_n)$, определяется как

$$f_{X_n}(x_n) = \int_{R^{N-1}} F_x(x_1, ..., x_N) dx_1 \ldots dx_{n-1} dx_{n+1} \ldots dx_N.$$

1.7. Независимость и корреляция случайных величин

Определение 15 (независимые события). *Семейство событий* $\{A_i\}_{i\in I}$ *в* $\mathcal{F}$ *называется независимыми относительно меры* P*, если для каждого непустого конечного набора индексов* $\{i_1, ...i_n\} \subset I$ *имеем*

$$\mathsf{P}(A_{i_1} \cap \ldots \cap A_{i_n}) = \mathsf{P}(A_{i_1}) \times \ldots \times \mathsf{P}(A_{i_n}).$$

Определение 16 (независимые случайные величины). *Пусть* $\{X_n\}_{n=1}^N$ *— N случайных величин в* $(\Omega, \mathcal{F}, \mathsf{P})$*. Если для любого* $x = (x_1, \ldots, x_n) \in R^N$*,*

$$\mathsf{P}\left(\bigcap_{n=1}^{N}\{X_n \le x_n\}\right) = \prod_{n=1}^{N}\mathsf{P}(X_n \le x_n),$$

тогда $\{X_n\}_{n=1}^N$ *называются независимыми случайными величинами.*

Эквивалентное определение независимых случайных величин состоит в том, что X_n, $n = 1, ..., N$ независимы тогда и только тогда, когда их совместное распределение — произведение их предельных распределений, т. е.

$$F_X(x) = \prod_{n=1}^{N} F_{X_n}(x_n).$$

С другой стороны, если существует совместная функция плотности f_X, то она также удовлетворяет правилу

$$f_X(x) = \prod_{n=1}^{N} f_{X_n}(x_n).$$

Определение 17 (ковариация). *Пусть* $X = (X_1, ..., X_n)$ *— N-мерный случайный вектор на вероятностном пространстве* $(\Omega, \mathcal{F}, \mathsf{P})$*. Тогда матрица*

$$cov(X) = \mathsf{E}[(X - \mathsf{E}(X))(X - \mathsf{E}(X))] \in R^{N\times N}$$

называется матрицей ковариаций случайного вектора X*. Ковариация* $cov(X) < \infty$*, если и только если* X *является квадратично интегрируемым. Каждая из диагональных записей* $cov(X)(X)$ *называется дисперсией* X_n*,* $n = 1, ..., N$ *и обозначается*

$$\mathsf{D}(X_n) = \mathsf{E}[(X_n - \mathsf{E}(X_n))^2] = \mathsf{E}(X_n^2) - \mathsf{E}(X_n)^2.$$

Каждый из недиагональных записей

$$c_{ij} = \mathsf{E}[(X_i - \mathsf{E}(X_i))(X_j - \mathsf{E}(X_j))],$$

ковариация $\mathrm{cov}(X_i, X_j)$. Если $c_{ij} = 0$ для $i \neq j$, то говорят, что X_i и X_j *некоррелированные.*

1.8. Произведение вероятностных пространств

Теперь рассмотрим семейство вероятностных пространств $(\Omega_k, \mathcal{F}_k, \mathsf{P}_k)$, $k = 1, ..., K_d$, из которого мы хотели бы построить произведение вероятностных пространств $(\Omega, \mathcal{F}, \mathsf{P})$, представленное

$$(\Omega, \mathcal{F}, \mathsf{P}) = \prod_{k=1}^{d} (\Omega_k, \mathcal{F}_k, \mathsf{P}_k),$$

где Ω, $\mathcal{F}$ и P — произведение пробных пространств, σ-алгебр, вероятностных мер соответственно. Их определения приведены ниже.

Определение 18 (произведение пробных пространств). *Произведение пробных пространств Ω определяется*

$$\Omega = \Omega_1 \times \ldots \times \Omega_K = \{(\omega_1, ...\omega_K) | \omega_k \in \Omega_k, k = 1, ..., K\}.$$

Определение 19 (произведение σ-алгебр). *Пусть $(\Omega_k, \mathcal{F}_k)$, $k = 1, ..., K$, обозначим K пространства мер и определим набор подмножеств $\mathcal{C}$ в $\prod_{k=1}^{K} \Omega_k$:*

$$\mathcal{C} = \{\prod_{i=1}^{d} A_k | A_k \in \mathcal{F}_k, k = 1, \ldots, K\}.$$

Тогда произведение σ-алгебр является σ-алгеброй, порожденной $\mathcal{C}$, т. е.

$$\mathcal{F} = \mathcal{F}_1 \times ... \times \mathcal{F}_d = \sigma(\mathcal{C}).$$

Теорема 2. *Пусть $(\Omega_k, \mathcal{F}_k)$, $k = 1, ..., K$ обозначают K вероятностных пространств. Тогда существует уникальная мера вероятности P, определенная на произведении σ-алгебр $\otimes_{k=1}^{K} = \mathcal{F}_k$, удовлетворяющая*

$$\mathsf{P}\left(\prod_{k=1}^{K} A_k\right) = \mathsf{P}_1(A_1) \ldots \mathsf{P}_K(A_k), A_k \in \mathcal{F}_k, \ k = 1, \ldots, K.$$

Для общего события $A \in \mathcal{F} = \otimes_{k=1}^{K} = \mathcal{F}_k$, $\mathsf{P}(A)$ *определяется*

$$\mathsf{P}(A) = \int_{\Omega_{i_K}} \left(... \left(\int_{\Omega_{i_1}} \mathcal{I}_A(\omega_1, ..., \omega_K)\, \mathsf{P}_{i_1}(d\omega_{i_1}) \right) ... \right) \mathsf{P}_{i_d}(d\omega_{i_d}).$$

Теорема 3. *Пусть* $(\Omega_k, \mathcal{F}_k, \mathsf{P}_k)$, $k = 1, ..., K$ *обозначим* K *вероятностных пространств и пусть* $(\Omega, \mathcal{F}, \mathsf{P})$ *будет произведением вероятностных пространств. Если* f *— измеримая функция на* $\mathcal{F} = \mathcal{F}_1 \times ... \times \mathcal{F}_K$ *и является интегрируемой по отношению к мере произведений* $\mathsf{P} = \mathsf{P}_1 \times ... \times \mathsf{P}_K$, *тогда*

$$\int_{\Omega_1 \times ... \times \Omega_K} f(\omega_1, \ldots, \omega_K) d(\mathsf{P}_1 \times ... \times \mathsf{P}_K) =$$

$$= \int_{\Omega_{i_K}} \left(... \left(\int_{\Omega_{i_1}} f(\omega_1, \ldots, \omega_K)\, \mathsf{P}_{i_1}(d\omega_{i_1}) \right) ... \right) \mathsf{P}_{i_d}(d\omega_{i_d}),$$

где $(i_1, ...i_K)$ *— произвольное изменение порядка* $(1, ..., K)$.

1.9. Случайные поля

В этом разделе рассмотрим расширение понятия случайных величин путем включения пространственной зависимости. Для удобства используем обозначения D для представления пространственной области и $x = (x_1, \ldots, x_d)$ для представления пространственных координат. Тогда в вероятностном пространстве $(\Omega, \mathcal{F}, P)$ случайный процесс представляет собой набор случайных величин

$$\{a(x, \omega), x \in D, \omega \in \Omega\}. \tag{1.5}$$

Термин «случайное поле» обычно относится к случайному процессу, принимающему значения в евклидовом пространстве R^d $d = 1, 2, 3$. Случайное поле можно рассмотреть двумя способами:

– для фиксированного $x \in D$, $a(x, \cdot)$ является случайной величиной в Ω;

– для фиксированного $\omega \in \Omega$, $a(\cdot, \omega)$ является реализацией случайного поля в D.

Естественно и полезно изучать статистику случайного поля. Например, *математическое ожидание* случайного поля $a(x, \omega)$ определяется как

$$\bar{a}(x) = \mathsf{M}[a(x, \cdot)]$$

для каждого $x \in D$, и *ковариационная функция* определяется как

$$\operatorname{cov}(x, x') = \mathsf{M}[(a(x, \cdot) - \bar{a}(x, \cdot))(a(x', \cdot) - \bar{a}(x', \cdot))]$$

для каждой пары $x, x' \in D$.

Разложение Karhunen–Loéve

Дан набор вещественных функций $\{b_n(x)\}_{n=1}^{\infty}$, определенный для $x \in D$, и набор некоррелированных случайных величин

$$\{\hat{\xi}_n(\omega)\}_{n=1}^{\infty}$$

для удобства с нулевым матожиданием и дисперсией $\{\sigma^2\}_{n=1}^{\infty}$, линейная комбинация

$$a(x, \omega) = \sum_{n=1}^{\infty} b_n(x)\hat{\xi}_n(\omega) \tag{1.6}$$

— случайное поле, которое можно использовать как простой способ представления данного коррелированное случайного поля, представленного в виде бесконечной суммы с независимыми случайными переменными. Ковариационная функция случайного поля (1.6) определяется как

$$\operatorname{cov}_a(x, x') = \sum_{n=1}^{\infty} \sigma_n^2 b_n(x) b_n(x').$$

Вышеуказанная структура является привлекательной, потому что случайные величины ξ_n не коррелированы или независимы, поэтому с ними легко работать на практике.

Положим $\xi_n(\omega) = \xi_n(\omega)/\sigma_n$ — набор независимых случайных величин со средним нулем и дисперсией 1. Теперь (1.6)

$$a(x, \omega) = \sum_{n=1}^{\infty} b_n(x)\xi_n(\omega).$$

Если функции $\{b_n\}_{n=1}^{\infty}$ являются ортонормированными

$$\int_D b_n(x) b_{n'}(x) dx = \delta_{nn'},$$

тогда

$$\int_D \operatorname{cov}_a(x, x') dx = \int_D \left(\sum_{i=1}^{\infty} \sigma_i^2 b_i(x) b_i(x') \right) b_n(x') dx =$$

$$= \sum_{i=1}^{\infty} \sigma_i^2 b_i(x) b_i(x') \delta_{ni} = \sigma_n^2 b_n(x) b_n(x').$$

Таким образом, мы видим, что σ_n^2 и $b_n(x)$, $n = 1, 2, ...$, являются собственными парами корреляционной функция $\mathrm{cov}_a(x, x')$.

Мы показали, что, учитывая ковариационную функцию $\mathrm{cov}_a(x, x')$, случайное поле $a(x, \omega)$ можно выразить как бесконечную сумму

$$a(x, \omega) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} b_n(x) \xi_n(\omega), \tag{1.7}$$

где $\{\lambda_n, b_n(x)\}_{n=1}^{\infty}$ обозначают пары собственных значений и функций данной ковариационной функции и $\xi_n(\omega)_{n=1}^{\infty}$ — независимые случайные величины с нулевым средним и единичной дисперсией. Разложение (1.7) хорошо известно как Karhunen–Loéve (KL) разложение случайного поля [96]. Разложение KL также известно как правильные ортогональные разложения (POD) и анализ главных компонентов (PCA).

Усеченные разложения KL дают возможность аппроксимировать случайные поля, т. е. у нас есть

$$a(x, \omega) \approx a(x, \omega)_N = \sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x) \xi_n(\omega).$$

Сходимость $a(x, \omega) \to a(x, \omega)_N$ гарантируется следующей теоремой Mercer.

Теорема 4. *Пусть область $D \subset R^d$ замкнута, пусть μ — строго положительная борелевская мера на D, пусть $\mathrm{cov}_a(x, x')$ — непрерывная функция на $D \times D$, которая является симметричной:*

$$\mathrm{cov}_a(x, x') = \mathrm{Cov}_a(x', x), \forall x, x' \in D,$$

неотрицательно определенной:

$$\int_D \int_D \mathrm{cov}_a(x, x') v(x) v(x') dx dx' \geq 0, \forall v(x),$$

и интегрируемой с квадратом.

Тогда

$$\lim_{N \to \infty} \max_{(x,x') \in D \times D} |\mathrm{cov}_a(x, x') - \sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x) b_n(x')| = 0.$$

Более того, ошибка монотонно уменьшается с ростом числа слагаемых в разложении.

1.10. Параметризация случайных коэффициентов

В этом разделе рассмотрено представление входных случайных параметров с использованием конечного числа некоррелированных или даже независимых случайных величин. Сделаем следующие предположения относительно стохастических входных данных, например случайного коэффициента $a(x, \omega_a)$.

Предположение 1. *Случайные входные данные удовлетворяют следующим свойствам.*

1) Функции $a(x, \omega_a)$ ограничены сверху и снизу с вероятностью 1, т. е. для $a(x, \omega_a)$ существует $a_{\min} > -\infty$ и $a_{\max} < \infty$ так что

$$P(\omega_a \in \Omega_a : a_{\min} \leq a(x, \omega_a) \leq f_{\max} \forall x \in \bar{D}) = 1.$$

2) Входные данные $a(x, \omega_a)$ имеют вид

$$a(x, \omega_a) = a(x, y_a(\omega_a) \text{ в } \bar{D} \times \Omega_a,$$

где N_a — целое положительное, $y_a(\omega_a) = y_{a,1}(\omega_a), \ldots, y_{a,N_a}(\omega_a)$ — вектор действительных некоррелированных случайных величин.

3) Случайные функции $a(x, \omega_a)$ σ-измеримы по отношению к y_a.

Далее мы приведем два примера случайных входных данных, которые удовлетворяют предположению.

Пример 1. Кусочно-постоянные случайные поля. Предположим, что пространственная область D представляет собой объединение непересекающихся подобластей $D_n, n = 1, ..., N_a$. Затем рассмотрим коэффициент $a(x, \omega)$, который является случайной постоянной в каждой подобласти D_n, т. е. $a(x, \omega)$ является кусочно-постоянной функцией

$$a(x, \omega) = a_0 + \sum_{n=1}^{N} a_n y_n(\omega) 1_{D_n}(x),$$

где $a_n, n = 0, ..., N$ обозначают константы, 1_{D_n} обозначает индикаторную функцию множества D_n, и случайные величины $y_n(\omega)$ ограничены и независимы. Обратите внимание, что допущение 1 требует ограничений на константы a_n и оценки случайных величин $y_n(\omega)$.

Пример 2. Karhunen–Loéve разложение. Согласно теореме 4 Mercer, любое коррелированное случайное поле второго порядка $a(x, \omega)$ с непрерывной ковариационной функцией $\mathrm{cov}(x, x')$ можно представить

в виде бесконечной суммы случайных величин. Одним из наиболее часто используемых примеров является разложение Karhunen–Loéve. Тогда случайное поле $a(x,\omega)$ может быть аппроксимировано усеченным разложением Karhunen–Loéve, имеющим форму

$$a(x,\omega) \approx a(x,\omega)_N = \mathsf{M}[a(x,\cdot)] + \sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x)\xi_n(\omega),$$

где λ_n и $b_n(x)$ для $n = 1, ..., N$ — собственные значения и соответствующие собственные функции для ковариационной функции, и $y_n(\omega)$ для $n = 1, ..., N_a$ обозначают некоррелированные вещественные случайные величины. Обратите внимание, что если процесс гауссовский, тогда случайные величины y_n равны стандартным независимым одинаково распределенным случайным величинам.

Предположение 1 и лемма Дуба–Дынкина 1 гарантируют, что $a(x,\omega)$ является измеримой по Борелю функцией случайного вектора y.

По определению случайные величины $\{y_n\}_{n=1}^N$ являются отображениями из пространства Ω в пространство R^N, поэтому мы обозначаем $\Gamma_n = y_n(\omega) \subset R$ образ случайной величины y_n и множество $\Gamma = \Pi_{n=1}^N \Gamma_n$, где $N \in N_+$.

Если мера распределения $y(\omega)$ абсолютно непрерывна, тогда существует совместная функция плотности вероятности для y_n, обозначаемая

$$\rho(y) : \Gamma \to R_+,\ \rho(y) \in L_\infty(\Gamma).$$

Таким образом, на основании предположения 1 пространство вероятностей $(\Omega, \mathcal{F}, P)$ отображается на $(\Gamma, \mathcal{B}(\Gamma), \rho(y)dy)$, где $\mathcal{B}(\Gamma)$ — борелевская σ-алгебра на Γ, $\rho(y)dy$ — конечная мера.

Глава 2

Непараметрические оценки функций плотности вероятности

2.1. Гистограммы

В тех случаях, когда желательно по данным эксперимента построить оценку плотности вероятностей, экспериментаторы чаще всего прибегают к построению гистограммы. Процедура ее построения проста и состоит из следующих шагов [40].

В области возможных значений измеряемой величины X строится сетка $\omega = \{x_i | i = 1, 2, \ldots, n\}$.

Определяется, сколько выборочных значений m_i от общего числа N оказалось в каждом интервале $(x_{i-1}, x_i]$.

Над каждым из интервалов строится вертикальный прямоугольник с площадью m_i/N. Высота прямоугольника $P_i = m_i/(N(x_i - x_{i-1}))$. Полученная совокупность прямоугольников и называется *гистограммой*. Другими словами гистограмма — кусочно-постоянная функция, определяемая своей сеткой ω, значениями $\{P_i\}$, принимающая на каждом интервале $(x_{i-1}, x_i]$ постоянное значение P_i.

Основанием для использования гистограммы $p_h(x)$ в качестве оценки неизвестной плотности вероятностей $p(x)$ является кусочно-интегральная сходимость $p_h(x)$ к $p(x)$, которая следует из того, что относительная частота m_i/N события $X \in (x_{i-1}, x_i]$ сходится к его вероятности p_i:

$$\frac{m_i}{N} \to \int_{x_{i-1}}^{x_i} p(x)dx.$$

Такой сходимости в ряде практических случаев оказывается достаточно; однако всегда желательно по возможности улучшить оценку при заданных ограничениях. Несколько факторов влияют на качество ги-

стограммы: объем выборки N, величина интервалов группировки. Все осложняется еще и тем, что степень влияния этих факторов зависит от неизвестного экспериментатору до опыта истинного распределения вероятностей $p(x)$. Поэтому на практике гистограммы строят с некоторым учетом свойств полученной выборки; например, величина интервала группировки выбирается так, чтобы не сгладить существенные особенности распределения; объем выборки связывают с тем, чтобы в ячейке с наименьшим числом измерений их насчитывалось не менее пяти; размещение интервалов связывают с положением наименьшего и наибольшего выборочных значений и т. п. Отметим также, что на качество гистограммы влияет и точность измерений.

Теоретическая задача оптимизации гистограммы может быть сформулирована в нескольких вариантах, однако ее решение связано с трудностями, так что конкретных результатов можно добиться лишь при некоторых частных предположениях.

Основной трудностью при построении гистограммы является выбор разбиения ω. Рекомендуется интервалы $(x_i, x_{i+1}]$ выбирать так, чтобы в каждый из них попадало одинаковое количество членов выборки (5–10, если N порядка 100). При увеличении N количество членов выборки в каждом интервале необходимо увеличивать. Если обозначить $h_{\max}$ длину наибольшего интервала, то $h_{\max}$ играет роль параметра регуляризации при оценивании плотности по гистограмме. При $N \to \infty$ $h_{\max}$ должен уменьшаться согласованно с ростом N.

Правило Стёрджеса — эмпирическое правило определения оптимального количества интервалов, на которые разбивается наблюдаемый диапазон изменения случайной величины при построении гистограммы плотности её распределения. Названо по имени американского статистика Герберта Стёрджеса (Herbert Arthur Sturges, 1882–1958). Количество интервалов n определяется как:

$$n = 1 + \lfloor \log_2 N \rfloor.$$

Обоснование правила следует из примера. Построим гистограмму со столбцами, каждый шириной 1, и по центру точки $i = 0, 1, ..., k-1$. Предположим, что i-й столбец имеет высоту, равную биномиальному коэффициенту $C_i^{k-1} = \binom{k-1}{i}$. С ростом k гистограмма принимает форму нормальной плотности с средним $(k-1)/2$ и дисперсией $(k-1)/4$. Об-

щий размер выборки вытекает из биномиального разложения

$$n = \sum_{i=0}^{k-1} C_i^{k-1} = (1+1)^{k-1} = 2^{k-1},$$

из которого следует правило Стёрджеса.

Среднеквадратичные оценки погрешности в L_2

В этом разделе представлены среднеквадратичные оценки погрешности гистограмм. Далее будем рассматривать гистограммы с постоянной шириной столбцов h [127], n — размерность выборки и

$$\hat{f}(x) = \frac{v_k}{nh},\ x \in [x_{k-1}, x_k).$$

Анализ гистограмм основывается на том факте, что случайная переменная v_k распределена по биномиальному закону:

$$v_k \sim B(n, p_k),\ \text{где } p_k = \int_{x_{k-1}}^{x_k} f(x)dx.$$

Рассмотрим среднеквадратичные оценки погрешности $\hat{f}(x)$, $x \in B_k = [x_{k-1}, x_k]$. Математическое ожидание $\mathrm{E}[v_k] = np_k$ и дисперсия $\mathrm{Var}[v_k] = np_k(1-p_k)$. Следовательно,

$$\mathsf{D}\hat{f}(x) = \frac{\mathsf{D}v_k}{(nh)^2} = \frac{p_k(1-p_k)}{nh^2}, x \in [x_{k-1}, x_k]$$

и

$$\mathrm{Bias}\hat{f}(x) = \mathsf{M}\hat{f}(x) - f(x) = \frac{1}{nh}\mathsf{M}v_k - f(x) = \frac{p_k}{h} - f(x).$$

Далее будем предполагать, что f — Липшиц-непрерывна на $[x_{k-1}, x_k]$ с константой γ_k.

Тогда по теореме о среднем

$$p_k = \int_{x_{k-1}}^{x_k} f(x)dx = hf(\xi_k), \xi_k \in [x_{k-1}, x_k];$$

$$\mathsf{D}\hat{f}(x) = \frac{p_k(1-p_k)}{nh^2} \le \frac{p_k}{nh^2} = \frac{f(\xi_k)}{nh}.$$

В силу Липшиц-непрерывности

$$|f(\xi_k) - f(x)| \le \gamma_k h.$$

Среднеквадратичная оценка погрешности

$$\mathsf{M}[\hat{f}(x) - f(x)]^2 = \mathsf{D}\hat{f}(x) + (\mathsf{M}[\hat{f}(x)] - f(x))^2 \leq \frac{f(\xi_k)}{nh} + (\gamma_k h)^2. \quad (2.1)$$

Определение 20. *Оценка плотности $\hat{f}$ называется сходящейся среднеквадратично, если*

$$\mathrm{MSE}[\hat{f}] \to 0, n \to \infty.$$

Теорема 5. *Предположим, что x — фиксированная точка из отрезка $[x_{k-1}, x_k]$, f — Липшиц-непрерывная функция на $[x_{k-1}, x_k]$ с константой γ_k. Тогда гистограмма сходится к f среднеквадратично, если при $n \to \infty$ выполнено $h \to 0$, $nh \to \infty$.*

Замечание 2. *MSE(x) ограничено (2.1) и минимально, когда*

$$h^*(x) = \left[\frac{f(\xi_k)}{2\gamma_k^2 n}\right],$$

MSE$(x) = O(n^{-2/3})$.*

Эти результаты заслуживают тщательного изучения. Оптимальная ширина столбца гистограммы уменьшается пропорционально $n^{-1/3}$. Эта скорость намного быстрее, чем в правиле Стёрджеса.

Интегральные оценки погрешности в L_2

Рассмотрим интегральную оценку дисперсии [Integrated Variance (IV)]

$$\mathrm{IV} = \int_{-\infty}^{\infty} \mathsf{D}\hat{f}(x)dx = \sum_{k=-\infty}^{k=\infty} \int_{[x_{k-1},x_k]} \mathsf{D}\hat{f}(x)dx. \quad (2.2)$$

Заметим

$$\int_{[x_{k-1},x_k]} \mathsf{D}\hat{f}(x)dx = \frac{p_k(1-p_k)}{nh},$$

и

$$\sum p_k = \int_{-\infty}^{\infty} f(x)dx = 1.$$

Кроме того,

$$\sum p_k^2 = \sum f^2(\xi_k)h^2 = h\int f^2(x)dx + O(h).$$

Комбинируя, получаем

$$\mathrm{IV} = \frac{1}{nh} - \frac{||f||_2^2}{n} + o(1/n). \tag{2.3}$$

Для вычисления $\mathrm{Bias}(\hat{f}(x)) = \mathsf{E}[\hat{f}(x)] - f(x)$ рассмотрим для определенности гистограмму на отрезке $x \in [0, h)$. Тогда

$$p_0 = \int_0^h f(\xi)d\xi = \int_0^h (f(x) + (\xi - x)f'(x) + (t - x)^2 f''(x) + \ldots)d\xi.$$

Таким образом,

$$\mathrm{Bias}\hat{f}(x) = p_0/h - f(x) = (h/2 - x)f'(x) + O(h^2). \tag{2.4}$$

$$\int_0^h (h/2 - x)^2 f'(x)^2 dx = \frac{h^2}{12} f'(\eta)^2,\ \eta \in [0, h).$$

$$\mathrm{ISB} = \frac{h^2}{12} h \sum f'(\eta_k)^2 = \frac{h^2}{12} \int f'(x)^2 dx + o(h^2). \tag{2.5}$$

Из соотношений (2.3) и (2.5) вытекает следующая теорема.

Теорема 6 ([127]). *Предположим, что $f \in C^1$ $||f'|| < \infty$. Тогда*

$$\mathrm{AMISE}(h) = \frac{1}{nh} + \frac{h^2}{12}||f'||_2^2, \tag{2.6}$$

и асимптотически оптимальное значение параметра

$$h^* = \left(\frac{6}{n||f'||_2^2}\right)^{1/3}.$$

Соответствующая оптимальная ошибка,

$$\mathrm{AMISE}^* = \mathrm{AMISE}(h^*) = \left(\frac{3||f'||}{4n}\right)^{2/3},$$

уменьшается с той же скоростью, что и оценка в следствии 2.

На практике, учитывая реальные данные от неизвестной плотности, сглаживающий параметр h выбирается не равным h^*, но в виде $h = ch^*$. В среднем, если $c \ll 1$, то гистограмма будет иметь высокую дисперсию и будет «слишком грубой», если $c \gg 1$, то оценка будет иметь большие уклонения или систематические ошибки и будет «слишком гладкой». Насколько чувствительна MISE к локальным отклонениям c от 1?

Асимптотическое MISE гистограммы в теореме 6 имеет вид

$$\mathrm{AMISE}(h) = \frac{a}{nh} + \frac{b}{2}h^2, \tag{2.7}$$

где a, b — положительные постоянные. Вместо того, чтобы рассматривать только этот частный случай, будет рассмотрен более общий вид AMISE [127]:

$$\mathrm{AMISE}(h) = \frac{a}{(d+r)nh^{d+r}} + \frac{b}{2p}h^{2p}, \tag{2.8}$$

где (d, p, r) — положительные целые, a, b — положительные постоянные, которые зависят от оценивателя плотности и неизвестной функции плотности.

Апостериорная оценка погрешности

Пусть известны $\Xi^l = \{\xi_1^l, \xi_2^l, \ldots, \xi_N^l\} \quad l = 1, 2 \ldots$ — повторные выборки случайной величины x с функцией плотности вероятности $p(x)$, $[a, b]$ — носитель. Построим на $[a, b]$ сетку $\omega = \{a = x_0 < x_1 < \ldots < x_n = b\}$ и определим число элементов m_k^l выборки Ξ^l, попавших в интервал $(x_{k-1}, x_k], k = 1, 2, \ldots, n$.

Заметим, что

$$z_k^l = \frac{m_k^l}{N} \to p_k = \int_{x_{k-1}}^{x_k} p(x)dx. \tag{2.9}$$

Таким образом, получим значения

$$\frac{z_k^l}{x_k - x_{k-1}}, k = 1, 2, \ldots, n,$$

которые определяют на сетке ω гистограммы H^l, аппроксимирующие плотность вероятности $p(x)$. Тогда $z_k^l, l = 1, 2$ можно интерпретировать как повторную выборку случайной величины z_k с математическим ожиданием, равным p_k и дисперсией σ_k.

Сравним случайные величины $z_k^1 - z_k^2$ и $p_k - z_k^1$. Случайная величина $z_k^1 - z_k^2$ имеет математическое ожидание равное нулю и дисперсию $2\sigma_k$, случайная величина $p_k - z_k^1$ имеет математическое ожидание равное нулю и дисперсию σ_k. Вычислим величины

$$s_1 = \sqrt{\sum_{k=1}^{n}(z_k^1 - z_k^2)^2},$$

$$s_2 = \sqrt{\sum_{k=1}^{n}(p_k - z_k^1)^2},$$

$$s = \sqrt{\sum_{k=1}^{n}(p_k - (z_k^1 + z_k^2)^2}.$$

Таким образом, математическое ожидание $\mathsf{M}[s] = \mathsf{M}[s_1]/2$.

При этом $s_1/2$ будем интерпретировать как апостериорную оценку точности гистограммы H, построенной по объединенной выборке $\Xi = \{\xi_1, \xi_2, \ldots, \xi_{2N}\}$.

Правило *апостериорной оценки погрешности* для гистограмм можно сформулировать следующим образом.

Пусть известна повторная выборка $\Xi = \{\xi_1, \xi_2, \ldots, \xi_{2N}\}$. На сетке $\omega = \{a = x_0 < x_1 < \ldots < x_n = b\}$ построим гистограмму H. Оценкой погрешности построенной гистограммы можно считать норму l_2

$$s = ||p - z||_2 = \sqrt{\sum_{k=1}^{n}(p_k - z_k)^2}.$$

Разобьем выборку Ξ на две равные по мощности выборки Ξ^1, Ξ^2 и построим гистограммы H^1, H^2. Вычислим оценку

$$s_1 = \sqrt{\sum_{k=1}^{n}(z_k^1 - z_k^2)^2},$$

при этом $s \approx s_1/2$.

Численные эксперименты

Для простоты эксперимента строились случайные величины с плотностью вероятности, соответствующей сумме четырех равномерно распределенных случайных величин на [0,1].

Например, при размерности выборки $2N = 10^4$ и размерности гистограммы $n = 50$, апостериорная оценка точности гистограммы $s_1/2 = 0.144563/2 = 0.07228$, при этом математическое ожидание ошибки $s = 0.06952$.

Для повышения надежности апостериорных оценок в работе использовался бутстрэп-подход (bootstrap, bootsrapping) [77]. Для этих целей

исходная выборка разбивалась на две равномощные выборки порядка 100–1000 раз, результаты усреднялись.

Результаты бутстрэп $n = 50$, $2N = 10^4$ $s = 0.06952$, $s_1/2 = 0.06973$. Следует заметить, что средняя апостериорная оценка при этом практически не зависит от исходной выборки и весьма точно приближает математическое ожидание ошибки. Апостериорная оценка погрешности $s_1/2 = 0.1394/2 = 0.0697$, математическое ожидание оценки погрешности $s = 0.06952$.

Представленный подход в сочетании с бутстрэп-методом можно использовать при оценке ошибок гистограмм дистанционного зондирования Земли, анализе изображений и снимков, в задачах стохастической гидрологии.

2.2. Частотные полигоны

Гистограмма — кусочно-постоянная функция, не является непрерывной и это ограничивает ее полезность в качестве инструмента для отображения многомерных данных. Частотный полигон (ЧП) является непрерывной оценкой плотности на основе гистограммы, в форме линейной интерполяции. В работе [127] D. Scott исследовал теоретические свойства одномерных и двумерных частотных полигонов и обнаружил, что у них есть существенные улучшения по сравнению с гистограммами.

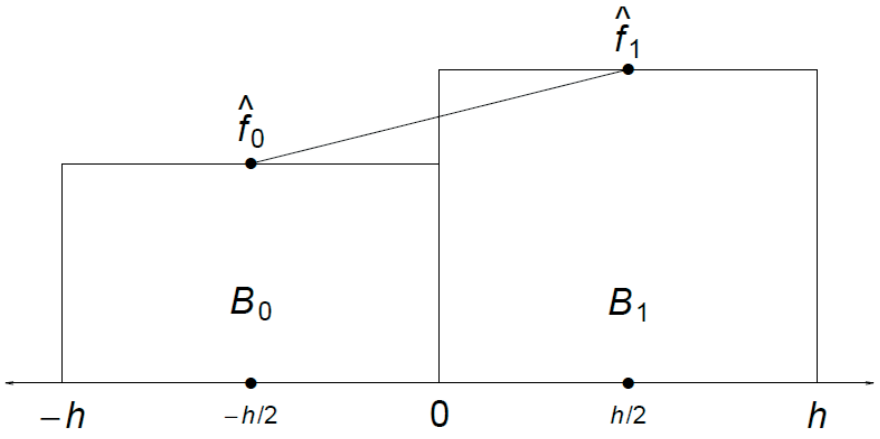


Рис. 2.1. Частотный полигон в области $(-h/2, h/2)$

Частотный полигон является кусочно-линейным интерполянтом середин столбцов гистограммы. Как таковой частотный полигон может выходить за пределы гистограммы в пустую область с каждого края.

Далее мы будем рассматривать гистограммы с одинаковыми по ширине столбцами.

Асимптотическая оценка MISE легко вычисляется на основе подхода «область к области», рассматривая типичную пару столбцов гистограммы, изображенных на рис. 2.1. Частотный полигон соединяет два смежных значения гистограммы $\hat{f}_0$ и $\hat{f}_1$, между центрами областей, как показано на рисунке. Он описывается уравнением

$$\hat{f}(x) = (1/2 - x/h)\hat{f}_0 + (1/2 + x/h)\hat{f}_1, x \in (-h/2, h/2).$$

Случайность в частотном полигоне полностью зависит от свойств $\hat{f}_i$ гистограмм. Заметим, что математическое ожидание $\mathsf{M}\hat{f}_i = p_i/h$. Следовательно,

$$\mathsf{M}\hat{f}(x) = (1/2 - x/h)p_0/h + (1/2 + x/h)p_1/h \approx f(0) + xf'(0) + h^2 f''(0)/6.$$

$$\text{Bias}\{\hat{f}(x)\} \approx (h^2 - 3x^2)f''(0)/6.$$

$$\int_{-h/2}^{h/2} (\text{Bias}\{\hat{f}(x)\})^2 dx = [49h^4 f''(0)^2/2880]h.$$

$$\text{ISB} \approx [\sum_i \frac{49}{2880} h^4 f''(ih)]h = \frac{49}{2880} h^4 ||f''||_2 + O(h^6).$$

Вычисление дисперсии аналогично. Из определения ЧП дисперсия $\hat{f}(x)$ равна

$$(1/2 - x/h)^2 \mathsf{D}\hat{f}_0 + (1/2 + x/h)^2 \mathsf{D}\hat{f}_1 + 2\text{cov}(\hat{f}_0, \hat{f}_1).$$

$$\mathsf{D}(\hat{f}_i) = \frac{nh_i(1 - p_i)}{(nh)^2} \approx \frac{f(0)(1 - hf(0))}{nh}.$$

$$\text{cov}(\hat{f}_0, \hat{f}_1) = \frac{-np_0p_1}{(nh)^2} \approx -\frac{f(0)^2}{n}.$$

$$\mathsf{D}\hat{f}(x) = \left(\frac{2x^2}{nh^3} - \frac{1}{2nh}\right) f(0) - \frac{f(0)^2}{n} + o(1/n).$$

$$\text{IV} \approx \sum_i \left[\frac{2f(ih)}{3nh} - \frac{f(ih)^2}{n}\right] h = \frac{2}{3nh} - \frac{1}{n}||f|| + o(1/n).$$

Теорема 7 ([127]). *Пусть f'' абсолютно непрерывна и $||f^{(3)}||_2 < \infty$.*

$$\mathrm{AMISE}(h) = \frac{2}{3nh} + \frac{49}{2880}h^4||f''||_2.$$

$$h^* = 2\left[\frac{15}{49||f''||}\right]^{1/5} n^{-1/5}.$$

2.3. Ядерные оценки функции плотности вероятности

Для оценки функции плотности вероятности часто используются непараметрические методы. Отметим, что вплоть до середины 50-х годов в качестве единственного подхода для построения непараметрической оценки функции плотности вероятности использовалась гистограмма. Первые важные результаты в области применения ядерных оценок для функции плотности вероятности были получены в работах М. Розенблатта, Э. Парзена и Н. Ченцова.

В общем виде ядерная оценка может быть записана в виде

$$\hat{f}^h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x-\xi_i}{h}) = \frac{1}{Nh}\sum_{i=1}^{N} K_h(x-\xi_i),$$

где $K_h(t) = K(t/h)/h$.

Обозначим

$$K_h(x,\xi_i) = K(\frac{x-\xi_i}{h}),$$

где ξ — случайная величина с функцией плотности вероятности $f(x)$.

Тогда математическое ожидание

$$\mathsf{M}[\hat{f}^h(x)] = \mathsf{M}[K_h(x,\xi)]$$

и

$$\sigma_N = \mathsf{D}[\hat{f}^h(x)] = \frac{1}{N}\mathsf{D}[K_h(x,\xi)].$$

Значение математического ожидания можно записать как

$$\mathsf{M}[K_h(x,\xi)] = \frac{1}{h}\int_{-\infty}^{\infty} K(\frac{x-t}{h})f(t)dt = \int_{-\infty}^{\infty} K(\eta)f(x-h\eta)d\eta.$$

Заметим, что

$$f(x-h\eta) = f(x) - hf'(x)\eta + \frac{h^2}{2}f''(x)\eta^2 + \frac{h^3}{6}f^{(3)}(x)\eta^3 + O(h^4).$$

$$\mathsf{M}[K_h(x,\xi)] = f(x)\int_{-\infty}^{\infty} K(\eta)d\eta - hf'(x)\int_{-\infty}^{\infty}\eta K(\eta)d\eta +$$
$$+\frac{h^2}{2}f''(x)\int_{-\infty}^{\infty}K(\eta)\eta^2 d\eta + \frac{h^3}{6}f^{(3)}(x)\int_{-\infty}^{\infty}K(\eta)\eta^3 d\eta + O(h^4).$$

Пусть ядро K удовлетворяет требованиям

$$\int_{-\infty}^{\infty}K(\eta)d\eta = 1, \int_{-\infty}^{\infty}\eta K(\eta)d\eta = 0$$

и

$$\int_{-\infty}^{\infty}\eta^3 K(\eta)d\eta = 0.$$

Обозначим

$$\int_{-\infty}^{\infty}\eta^2 K(\eta)d\eta = \sigma_K^2.$$

Тогда

$$\mathsf{M}[\hat{f}^h(x)] = \mathsf{E}[K_h(x,\xi)] = f(x) + \sigma^2 h^2 f''(x)/2 + O(h^4)$$

и

$$\mathsf{M}[\hat{f}^h(x) - f(x)] = \sigma^2 h^2 f''(x)/2 + O(h^4).$$

Определим

$$f^h(x) = \mathsf{M}[\hat{f}^h(x)] = f(x) + \sigma^2 h^2 f''(x)/2 + O(h^4). \tag{2.10}$$

и

$$f^{2h}(x) = \mathsf{M}[\hat{f}^{2h}(x)] = f(x) + 4\sigma^2 h^2 f''(x)/2 + O(h^4). \tag{2.11}$$

Далее оценим

$$\mathsf{D}[K_h(x,\xi)] = \mathsf{M}[(\frac{1}{h}K(\frac{x-\xi}{h}))^2] - (\mathsf{M}[\frac{1}{h}K(\frac{x-\xi}{h})])^2.$$

$$\mathsf{M}[(\frac{1}{h}K(\frac{x-\xi}{h}))^2] = \frac{1}{h}\int_{-\infty}^{\infty}K^2(\eta)f(x-h\eta)d\eta =$$
$$= \frac{f(x)}{h}\int_{-\infty}^{\infty}K^2(\eta)d\eta + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty}K^2(\eta)\eta^2 d\eta + O(h^4).$$

В итоге получим

$$\sigma^2(x) = \mathsf{D}[\hat{f}(x)] = \frac{f(x)}{Nh}||K||_2^2 + \frac{f(x)^2}{N} + O(h/N).$$

Без ограничения общности, можно записать, что [127]

$$\mathsf{M}[(\hat{f}^h(x) - f(x))^2] = \sigma^2(x) + f''(x)^2 h^4/4 + O(h^6)$$

и

$$\sigma_I^2 = \int_{-\infty}^{\infty} \sigma^2(x)dx = \frac{||K||_2^2}{Nh} - \frac{||f||_2^2}{N} + O(h/N).$$

$$\mathsf{M}||\hat{f}^h - f||_2^2 = \sigma_I^2 + \frac{||f''||^2 h^4}{4} + O(h^6). \tag{2.12}$$

2.4. Экстраполяция Ричардсона и правило Рунге

В разделе рассматривается новый подход к повышению точности в задачах построения аппроксимации функции плотности вероятности и оценке ее погрешности. Подход основан на комбинации ядерных оценок с различными параметрами сглаживания h. Для этих целей используются экстраполяция Ричардсона и правило Рунге [71, 114].

Оценка производных плотности имеет важное значение для приложений. Это было отмечено еще в первых работах по оценке плотности вероятности. В данной статье рассматриваются применение правила Рунге для вычисления второй производной оценки функции плотности вероятности. В отличие от известных методов, этот подход не требует дифференцирования ядерных оценок или вычисления конечных разностей от эмпирической функции плотности вероятности. Подробный обзор существующих методов оценки производных и библиография представлены в [127].

Использование оценок вторых производных позволяет получить реалистичные оценки математических ожиданий в l_2-норме погрешности аппроксимации функции плотности вероятности. Знание этих оценок позволяет рассчитать оптимальный параметр сглаживания h [127].

Экстраполяция Ричардсона — очень мощный подход, который может эффективно использоваться для повышения производительности численных расчетов. Он был представлен Ричардсоном в начале XX века, и после этого многие ученые и инженеры многократно использовали его для повышения точности численных расчетов математических моделей. В большинстве приложений этот метод до сих пор использовался главным образом в усилиях либо для повышения точности результатов модели. Однако этот подход в виде правила Рунге можно использовать для

оценки и проверки величины вычислительных ошибок, попытаться добиться большей точности и повысить эффективность вычислительного процесса.

Экстраполяция Ричардсона является общим подходом для получения результатов высокой точности по формулам низкого порядка [98, 121]. Она получила широкое распространение для повышения точности разностных методов решения задач Коши для систем обыкновенных дифференциальных уравнений и краевых задач для дифференциальных уравнений в частных производных.

В книге [98] рассмотрены аспекты применения экстраполяции Ричардсона к разностным методам решения задач математической физики. В монографии [121] представлены главным образом задачи Коши для систем ОДУ.

Экстраполяция Ричардсона основана на разложении по степеням h приближенного решения u^h как суммы

$$u^h = u + h^k v + O(h^{k+m}), \tag{2.13}$$

где u есть искомое точное решение, v — неизвестная функция и h — малый параметр дискретизации, который чаще всего рассматривается как шаг разностной сетки. Целое значение k характеризует порядок точности приближенного решения, и $m > 0$ характеризует порядок точности приближенного решения членом погрешности $h^k v$. Поскольку u и v не зависят от h, для параметра $h/2$ справедливо разложение:

$$u^{h/2} = u + \left(\frac{h}{2}\right)^k v + O(h^{k+m}). \tag{2.14}$$

Объединим два разложения таким образом, чтобы исключить ошибку порядка h^k. Умножим (2.14) на 2^k и вычтем из (2.13); получаем

$$u = \frac{2^k}{2^k - 1} u^{h/2} - \frac{1}{2^k - 1} u^h + O(h^{k+m}).$$

Таким образом получено приближение точного решения с более высоким порядком точности.

Одно из первых правил для практической оценки погрешности было предложено К. Рунге в начале XX века. Это правило широко использовалось сначала в области квадратурных вычислений, затем в разностных методах и методе конечных элементов.

Вычтем (2.14) из (2.13), избавляясь от u:

$$u^h - u^{h/2} = v\left(\frac{h}{2}\right)^k (2^k - 1) + O(h^{k+m}).$$

Отсюда можно определить главный член погрешности:

$$u^{h/2} - u \approx \frac{u^h - u^{h/2}}{2^k - 1}. \tag{2.15}$$

Поскольку в формуле (2.15) отброшен остаточный член порядка $O(h^{k+m})$, то полученное выражение не приводит к гарантированной оценке, но при достаточно малых h дает представление о величине погрешности численного решения [21].

Традиционно экстраполяция Ричардсона используется для повышения точности численных решений дифференциальных уравнений. В этих работах, поскольку h можно выбрать достаточно малым, как правило использовались два разложения h, $2h$. Это позволяет пренебречь величиной отбрасываемых слагаемых. Особенности ядерных оценок функций плотности вероятности заключаются в том, что параметр h не может быть выбран произвольно малым [127]. Следовательно, в этих случаях надо увеличивать число слагаемых разложения по параметру h.

Теоретические аспекты экстраполяции Ричардсона. Рассмотрим некоторый процесс вычисления $y(h)$, который зависит от параметра $h > 0$ и для которого справедливо разложение в ряд по параметру $h > 0$:

$$y(h) = y(0) + c_1 h^{p_1} + c_2 h^{p_2} + c_3 h^{p_3} + \ldots, \tag{2.16}$$

где $y(0)$ есть точное решение, константы c_i не зависят от h, p_i — целые числа. Решим задачу (2.16) при h_i $i = 1, 2, \ldots l$. Получаем

$$\begin{array}{l} y(h_1) = y(0) + c_1 h_1^{p_1} + c_2 h_1^{p_2} + c_3 h_1^{p_3} + \ldots + c_l h_1^{p_l} + \ldots, \\ y(h_2) = y(0) + c_1 h_2^{p_1} + c_2 h_2^{p_2} + c_3 h_2^{p_3} + \ldots + c_l h_2^{p_l} + \ldots, \\ \vdots \qquad \ldots \qquad \vdots \\ y(h_l) = y(0) + c_1 h_l^{p_1} + c_2 h_l^{p_2} + c_3 h_l^{p_3} + \ldots + c_l h_l^{p_l} + \ldots \end{array} \tag{2.17}$$

Отбрасывая слагаемые порядка выше p_l, получаем систему линейных алгебраических уравнений, которую можно разрешить относительно $y(0)$ и c_i, $i = 1, \ldots, l-1$ через известные значения $y(h_i), i = 1, \ldots, l$.

Рассмотрим частный случай для $l = 3$, $h_i = ht^{i-1}$, $i = 1, 2, 3$:

$$\begin{array}{l} s(0) + c_1 h^2 + c_2 h^4 = s(h), \\ s(0) + c_1 h^2 t^2 + c_2 h^4 t^4 = s(ht), \\ s(0) + c_1 h^2 t^4 + c_2 h^4 t^8 = s(ht^2). \end{array}$$

Решив систему, получаем

$$s(0) = \frac{s(h)t^6 - s(ht)t^4 - s(ht)t^2 + s(ht^2)}{t^6 - t^4 - t^2 + 1},$$

$$c_1 = ((s(ht) - s(h))t^4 - s(ht^2) + s(ht))/(t^6 - 2t^4 + t^2),$$
$$c_2 = -((s(ht) - s(h))t^2 - s(ht^2) + s(ht))/(t^8 - t^6 - t^4 + t^2).$$

В частности, полагая $t = \sqrt{2}$, получаем

$$s(0) = \frac{8s(h) - 6s(\sqrt{2}h) + s(2h)}{3};$$

используя только два разложения, получаем хорошо известное решение для h и $2h$:

$$s(0) = \frac{4s(h) - s(2h)}{3}.$$

Таким образом, получено решение системы линейных алгебраических уравнений с h, $\sqrt{2}h$ и $2h$. Такой подход уменьшает погрешность решения системы линейных алгебраических уравнений и позволяет использовать экстраполяцию Ричардсона для повышения точности ядерных оценок.

Разложение ядерных оценок для экстраполяции Ричардсона. Для использования экстраполяции Ричардсона в ядерных оценках функции плотности вероятности необходимо получить разложения общего вида (2.17). Для этого рассмотрим свойства ядерных оценок.

Ядерная оценка может быть компактно записана, как [127]:

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x - \xi_i}{h}) = \frac{1}{Nh}\sum_{i=1}^{N} K_h(x - \xi_i),$$

где $K_h(t) = K(t/h)/h$.

Заметим, что

$$K_h(x, \xi_i) = K(\frac{x - \xi_i}{h}),$$

где ξ есть случайная величина с плотностью вероятности $f(x)$.

Тогда

$$\mathsf{M}[\hat{f}(x)] = \mathsf{M}[K_h(x, \xi)]$$

и

$$\sigma_N = \mathsf{D}[\hat{f}(x)] = \frac{1}{N}\mathsf{D}[K_h(x, \xi)].$$

Математическое ожидание может быть записано в виде

$$\mathsf{M}[K_h(x,\xi)] = \frac{1}{h}\int_{-\infty}^{\infty} K(\frac{x-t}{h})f(t)dt =$$

$$= \int_{-\infty}^{\infty} K(\eta)f(x-h\eta)d\eta.$$

Отметим, что

$$f(x-h\eta) = f(x) - hf'(x)\eta + \frac{h^2}{2}f''(x)\eta^2 +$$

$$+\frac{h^3}{6}f^{(3)}(x)\eta^3 + \frac{h^4}{24}f^{(3)}(x)\eta^4 + O(h^6).$$

$$\mathsf{M}[K_h(x,\xi)] = f(x)\int_{-\infty}^{\infty} K(\eta)d\eta - hf'(x)\int_{-\infty}^{\infty}\eta K(\eta)d\eta +$$

$$+\frac{h^2}{2}f''(x)\int_{-\infty}^{\infty} K(\eta)\eta^2 d\eta +$$

$$+\frac{h^3}{6}f^{(3)}(x)\int_{-\infty}^{\infty} K(\eta)\eta^3 d\eta + \frac{h^4}{6}f^{(4)}(x)\int_{-\infty}^{\infty} K(\eta)\eta^4 d\eta + O(h^5).$$

Предположим, что $f \in C^{2n+2}$ и ядро K удовлетворяет условиям

$$\int_{-\infty}^{\infty} K(\eta)d\eta = 1,$$

ядра симметричные $K(\eta) = K(-\eta)$, тогда для всех нечетных α выполнено соотношение

$$\int_{-\infty}^{\infty} \eta^\alpha K(\eta)d\eta = 0.$$

Обозначим

$$\int_{-\infty}^{\infty} \eta^i K(\eta)d\eta = \sigma_i^2.$$

Тогда

$$\bar{f} = \mathsf{M}[K_h(x,\xi)] = f(x) + \sum_{i=1}^{n}\sigma_{2i}^2 h^{2i} f^{(2i)}(x)/(2i!) + O(h^{2n+2})$$

и

$$\mathsf{M}[\hat{f}(x) - f(x)] = \sum_{i=1}^{n}\sigma_{2i}^2 h^{2i} f^{(2i)}(x)/2 + O(h^{2n+2}).$$

Определим

$$f^h(x) = \mathsf{M}[\hat{f}(x)] = f(x) + \sum_{i=1}^{n} \sigma_{2i}^2 h^{2i} f^{(2i)}(x)/2 + O(h^{2n+2}) \tag{2.18}$$

и

$$f^{th}(x) = \mathsf{M}[\hat{f}(x)] = f(x) + \sum_{i=1}^{n} \sigma_{2i}^2 t^{2i} h^{2i} f^{(2i)}(x)/2 + O(h^{2n+2}). \tag{2.19}$$

Далее применим экстраполяцию Ричардсона для $f^h(x)$ [71]. Решив полученную систему линейных алгебраических уравнений, получаем для $t = 2$ и двух значений h, $2h$

$$f(x) = \frac{4}{3} f^h(x) - \frac{1}{3} f^{2h}(x) + O(h^4).$$

Заметим, что мы получили аппроксимацию $f(x)$

$$f^h_{\text{cor}}(x) = \frac{4}{3} \hat{f}^h(x) - \frac{1}{3} \hat{f}^{2h}(x) \tag{2.20}$$

с точностью $O(h^4)$.

Используя три значения $h, \sqrt{2}h, 2h$, получаем аппроксимацию $f(x)$

$$f^h_{\text{cor}}(x) = \frac{8\hat{f}^h(x) - 6\hat{f}^{\sqrt{2}h}(x) + \hat{f}^{2h}(x)}{3}$$

с точностью $O(h^6)$.

Оценим дисперсию $f^h_{\text{cor}}(x)$. Известна формула для суммы случайных величин

$$\mathsf{D}[\sum_{i=1}^{n} a_i f_i] = \sum_{i=1}^{n} a_i^2 \mathsf{D}[f_i] + 2 \sum_{1 \le i < j \le n} \text{cov}(f_i, f_j),$$

ковариацию можно оценить

$$|\text{cov}(f_i, f_j)| \le \sqrt{\mathsf{D}[f_i]\mathsf{D}[f_j]}.$$

Известна дисперсия ядерных оценок $\hat{f}^h(x)$ [127]:

$$\mathsf{D}[\hat{f}^h(x)] = \frac{f(x)}{Nh} ||K||_2^2 + \frac{f(x)^2}{N} + O(h/N).$$

Полагая $\mathsf{D}[\hat{f}^h(x)] \ge \mathsf{D}[\hat{f}^{\sqrt{2}h}(x)]$, $\mathsf{D}[\hat{f}^h(x)] \ge \mathsf{D}[\hat{f}^{2h}(x)]$, окончательно получаем

$$\mathsf{D}[f^h_{\text{cor}}(x)] \le C\mathsf{D}[\hat{f}^h(x)],$$

где $C \leq (101/9 + 6)$. Последнее обстоятельство следует учитывать при выборе h, N.

Правило Рунге и оценки погрешности. Как уже было показано, используя правило Рунге, можно получить реалистичные оценки погрешности. При использовании двух ядерных оценок с параметрами h, $2h$ погрешность будет иметь точность порядка $O(h^4)$, т. е. асимптотически точные оценки при малых h. В случае достаточно гладких функций f для повышения точности оценок следует учитывать большее число членов в разложении и соответственно большее число ядерных оценок с параметрами h_i.

Таким образом, учитывая ядерные оценки при трех различных параметрах $h, \sqrt{2}h, 2h$, получаем оценку погрешности в виде $c_1 + c_2$, точность порядка $O(h^6)$, где

$$c_1 = (4(\hat{f}^{\sqrt{2}h}(x) - \hat{f}^{h}(x)) - \hat{f}^{2h}(x) + \hat{f}^{\sqrt{2}h}(x))/2,$$

$$c_2 = -(2(\hat{f}^{\sqrt{2}h}(x) - \hat{f}^{h}(x)) - \hat{f}^{2h}(x) + \hat{f}^{\sqrt{2}h}(x))/6.$$

Оценка производных. Заметим, что

$$c_i = \sigma_{2i}^2 h^{2i} f^{(2i)}(x)/(2i!), //i = 1, 2, ...$$

Таким образом, зная, например, c_1, σ_2, можно определить значение второй производной $f''(x)$:

$$f''(x) = \mathsf{E}\left[\frac{2c_1(x)}{\sigma_2 h^2}\right].$$

Численные примеры. В качестве примера рассмотрим функцию плотности вероятности для случайной величины, определяемой как сумма числа n независимых случайных величин, каждая из которых имеет равномерное распределение. Такое распределение известно в теории вероятностей и статистике как распределение Ирвина–Холла, названное в честь Джозефа Оскара Ирвина и Филиппа Холла.

При этом заметим, что плотность вероятности суммы n равномерно распределенных случайных величин может быть представлена как

$$p_n(x) = \frac{1}{(n-1)!}(x^{n-1} - C_n^1(x-1)^{n-1} + C_n^2(x-2)^{n-1} - ...), \qquad (2.21)$$

где C_n^k есть биномиальные коэффициенты, и для каждого фиксированного аргумента x сумма в скобках относится только к тем переменным, для которых значение $(x-k), k = 1, 2, ...$ неотрицательно.

Так, когда $n = 6$, мы имеем:

$$p(x) = \begin{cases} x^5/120, & x \in [0,1]; \\ (-5x^5 + 30x^4 - 60x^3 + 60x^2 - 30x + 6)/120, & x \in [1,2]; \\ (10x^5 - 120x^4 + 540x^3 - 1140x^2 + 1170x - 474)/120, & x \in [2,3]; \\ (-10x^5 + 180x^4 - 1260x^3 + 4260x^2 - 6930x + 4386)/120, & x \in [3,4]; \\ (5x^5 - 120x^4 + 1140x^3 - 5340x^2 + 12\ 270x - 10\ 974)/120, & x \in [4,5]; \\ (-x^5 + 30x^4 - 360x^3 + 2160x^2 - 6480x + 7776)/120, & x \in [5,6]. \end{cases}$$

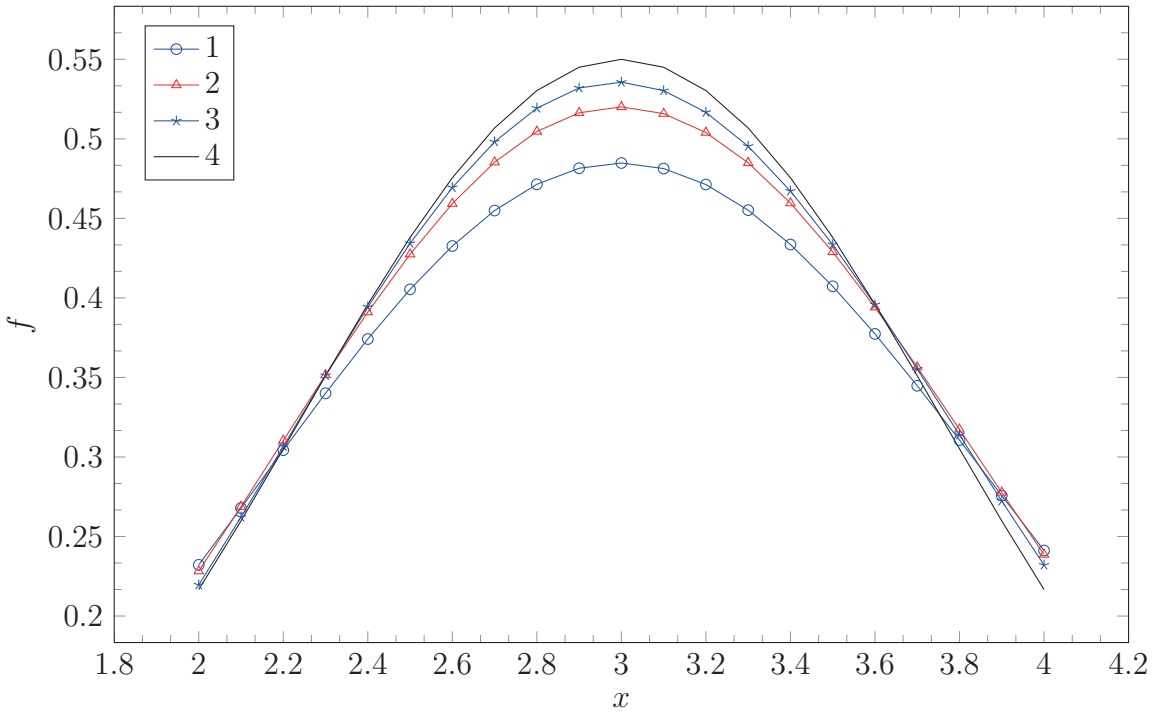


Рис. 2.2. Уточнение ядерной оценки

На рис. 2.2 показано уточнение ядерной оценки распределения Ирвина–Холла, на интервале $(2, 4)$. Линия 1 — ядерная оценка кубическими ядрами с $h = 1$, линия 2 — уточнение, использующее две ядерные оценки кубическими ядрами с $h = 1, 2$, линия 3 — уточнение, использующее три ядерные оценки кубическими ядрами с $h = 1, \sqrt{2}, 2,\ 4$ — распределения Ирвина–Холла при $n = 6$.

Для этого примера приведем погрешность ядерной оценки кубическими ядрами с $h = 1$

$$\mathrm{Err}^2 = \int_2^4 (p(x) - \hat{f}^h(x))^2 dx$$

и ее аппроксимацию

$$\mathrm{Err}^2_{\mathrm{app}} = \int_2^4 (c_1(x) + c_2(x))^2 dx.$$

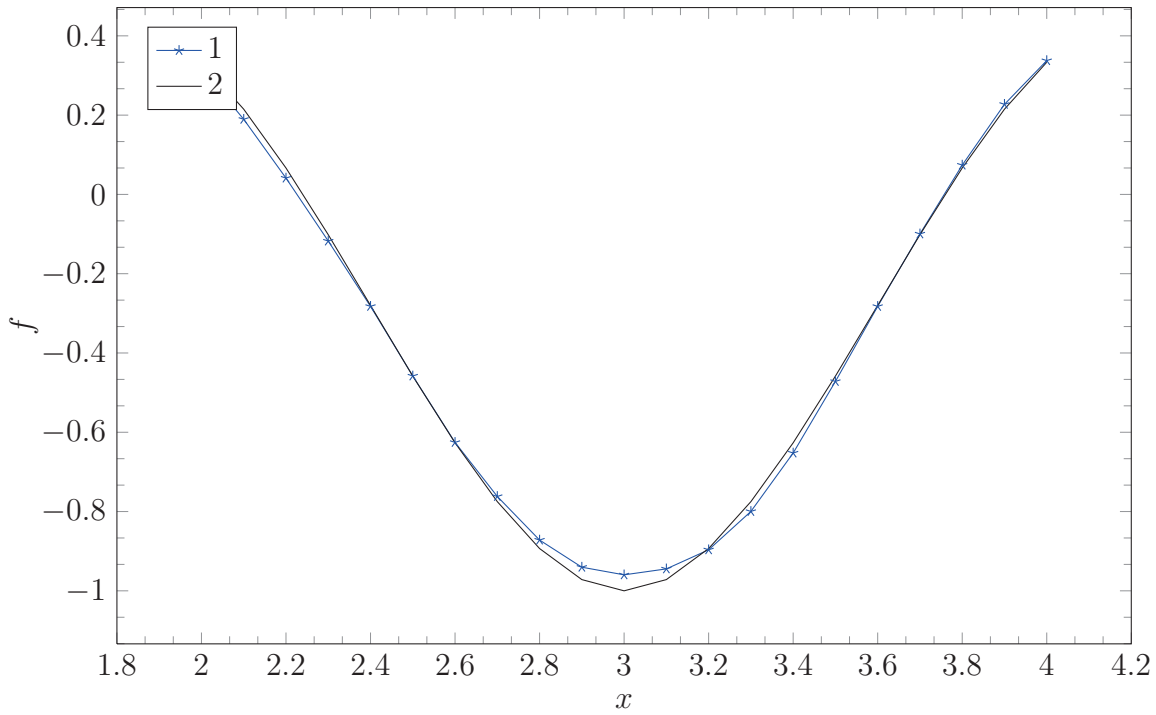


Рис. 2.3. Оценка второй производной

Были получены следующие оценки математических ожиданий $\mathsf{E}[\mathrm{Err}] = 0.0369$, $\mathsf{E}[\mathrm{Err}_{\mathrm{app}}] = 0.0358$. Таким образом получены реалистичные оценки погрешности, позволяющие судить об истинной погрешности ядерных оценок.

На рис. 2.3 показана оценка второй производной (линия 1) распределения Ирвина–Холла (линия 2), на интервале $(2, 4)$, размерность выборки $N = 10^6$.

Представленные результаты демонстрируют новый подход для задач, связанных с проблемой повышения точности оценки функции плотности вероятности по эмпирическим данным в условиях случайной неопределенности. Это достигается за счет численных процедур, основанных на идее экстраполяции Ричардсона. Для оценивания погрешностей ядерных оценок и значений вторых производных функции плотности вероятности используется правило Рунге. Поскольку эти подходы не требуют значительного увеличения вычислительных затрат, то они будут весьма полезны для практических задач обработки и анализа эмпирических данных.

Глава 3

Функциональный анализ данных

3.1. Введение

С развитием современных технологий все больше и больше данных записывается непрерывно. Во многих случаях их удобно представлять в виде функций. Данные, представленные с использованием функций, становятся широко распространенным типом функциональных данных. Для их изучения предназначен функциональный анализ данных (ФАД), который охватывает как теорию, так и методы их исследования.

Эта глава представляет обзор понятий, моделей, методов, излагающих суть функционального анализа данных, начиная с простых статистических понятий, таких как среднее значение и ковариационные функции, и далее модели функциональной линейной регрессии, методы кластеризации и классификации функциональных данных.

Функциональный анализ данных (ФАД) имеет дело с анализом и теорией данных, которые представлены в виде функций, изображений или более общих объектов. Элементы функциональных данных — функции, где для каждого субъекта в случайной выборке записывается одна или несколько функций. Хотя термин “functional data analysis” был придуман Ramsay, Dalzell [116, 117], история этой области намного старше. Функциональные данные по своей природе бесконечномерные. Высокая внутренняя размерность этих данных создает проблемы как для теории, так и для вычислений. Эти проблемы зависят от того, как были собраны функциональные данные. Структура данных с большой или бесконечной размерностью является богатым источником информации и приносит много возможностей для исследований и анализа данных.

Функциональные данные первого поколения обычно состоят из случайной выборки независимых вещественных функций, $X_1(t), \ldots, X_n(t)$

на компактном интервале $I = [0, T]$ на вещественной прямой. Такие данные были названы данными кривой (curve data) (Gasser et al. 1984, Rice & Silverman 1991, Gasser & Kneip 1995) [88, 118, 119]. Эти вещественные функции можно рассматривать как реализации одномерного стохастического процесса, который часто предполагается в гильбертовом пространстве, например в пространстве $L_2(I)$. Здесь говорят, что случайный процесс $X(t)$ является процессом в пространстве L_2 тогда и только тогда, когда он удовлетворяет условию

$$\mathsf{M}[\int_I X^2(t)dt] < \infty.$$

Для исследования функциональных данных применяются параметрические подходы, в то время как необходимость гибкости анализа таких данных в сочетании с естественным порядком (во времени) способствует применению непараметрических подходов. Например, гладкость отдельных случайных функций (как реализаций случайного процесса), существование непрерывных вторых производных часто используется при регуляризации и является особенно полезной при применении непараметрических методов сглаживания.

В этой главе рассматриваются функциональные данные первого поколения. Функциональные данные следующего поколения — это функциональные данные, которые являются частью сложных объектов данных, и они могут быть многомерными, коррелированными или включать изображения или формы. Примеры функциональных данных следующего поколения включают данные о функционировании мозга и нейровизуализации. Краткое обсуждение функциональных данных следующего поколения представлено в отчете London workshop on the Future of Statistical Sciences (Int. Year Stat. 2013, p. 23).

Основная и общепринятая основа в ФАД — это рассматривать функциональные данные как реализации основного стохастического процесса. В реальных приложениях анализа таких данных основной процесс часто не может наблюдаться напрямую, так как данные могут быть собраны дискретно с течением времени, либо на фиксированной или случайной сетке времени. В таких ситуациях основной процесс считается скрытым. Временная сетка, где проводятся наблюдения, может быть плотной, разреженной или пустой, и она может отличаться от предмета к предмету. Первоначально функциональные данные рассматривались как образцы полностью наблюдаемых траекторий. Немного более общее предположение состоит в том, что функциональные данные записываются на той

же плотной временной сетке, упорядоченной по времени $t_1, \ldots, t_p$ для всех n предметов. Если запись выполняется инструментом, таким как электроэнцефалограмма или функциональное устройство регистрации магнитно-резонансной томографии, временная сетка обычно равномерная, т. е. $t_j - t_{j-1} = \text{const}$ для всех j. В асимптотическом анализе предполагается, что $t_{j+1} - t_j$ стремится к нулю при n, стремящемся к бесконечности, поэтому $p = p_n$ — это последовательность, которая стремится к бесконечности. Хотя большие значения p приводят к большим сложностям, здесь это скорее благо. Это благо реализуется путем принятия предположения о гладкости процессов в пространстве L_2, так что информация во времени может быть объединена, чтобы преодолеть проклятие размерности. Таким образом, сглаживание служит как инструмент для регуляризации.

Хотя формального определения плотных функциональных данных не существует, соглашение заключалось в том, чтобы объявить функциональные данные как плотные (в отличие от разреженных) выборки, когда p_n сходится к бесконечности достаточно быстро, чтобы позволить соответствующую оценку для средней функции $\mu(t) = \mathsf{E}X(t)$, где X — базовый процесс, чтобы достичь параметрического $\sqrt{n}$ коэффициента сходимости для стандартных метрик, таких как норма L_2.

Разреженные функциональные данные возникают в исследованиях, для которых измеряются объекты в разные моменты времени, и количество измерений n_i для объекта i может быть ограничено некоторой константой C, т. е. $\sup_{1 \le i \le n} n_i < C < \infty$. Строгое определение типов функциональных данных на основе метрик и количества наблюдений по-прежнему отсутствует. Zhang & Wang описывают возможный подход [146, 147].

В действительности многие наблюдаемые данные загрязнены случайным шумом, обусловленным ошибками измерения, которые часто считаются независимым между субъектами. Ошибки измерения могут рассматриваться как случайные колебания вокруг гладкой траектории или как фактические ошибки в измерении. Сильной стороной ФАД является то, что он легко приспосабливается к ошибкам измерения, потому что наблюдает повторные измерения для каждого субъекта. Разреженные и нерегулярно выбранные функциональные данные, которые соответствуют общему типу «продольных данных», обычно требуют больше усилий в теории и методологии, чем плотно выбранные функциональные данные, которые записываются непрерывно.

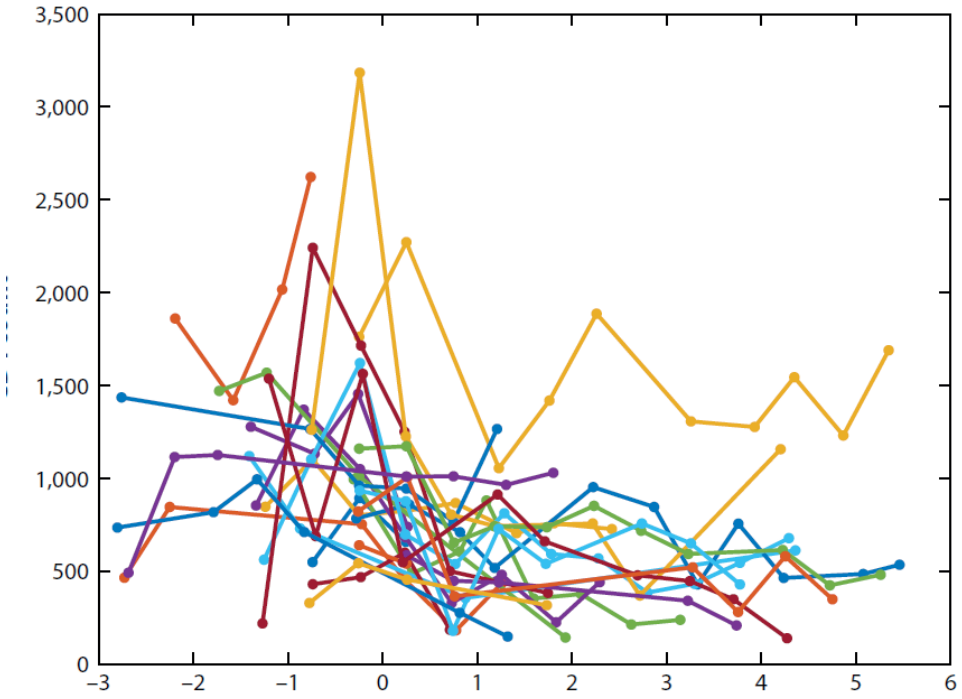


Рис. 3.1. Число клеток Т-лимфоцитов для 25 субъектов показаны разными цветами

Пример таких данных представлен на рис. 3.1 (число клеток Т-лимфоцитов).

Функциональные данные, которые наблюдаются непрерывно без ошибки, являются самым простым типом для обработки в качестве теории для случайных процессов, к ним легко применимы функциональные законы больших чисел и функциональные центральные предельные теоремы.

Одна из проблем в ФАД заключается в том, что многие основные задачи приводят к обратным задачам, особенно для функциональной линейной регрессии и многих функциональных мер корреляции. Поскольку функциональные данные по своей сути бесконечномерны, сокращение размерности является ключевым для моделирования и анализа данных.

Кластеризация и классификация функциональных данных являются полезными и важными инструментами с широким приложением в ФАД. Методы включают в себя расширения классических k-средних и иерархической кластеризации, байесовский и модельный подходы к кластеризации и классификации с помощью функциональной регрессии.

Инструменты исследования, которые полезны для ФАД, включают различные методы сглаживания, особенно ядерные, метод наименьших квадратов и сплайновое сглаживание, для которых существуют различные превосходные справочники (Wand & Jones 1995, Fan & Gijbels 1996,

Eubank 1999, de Boor 2001) [140, 83, 80]; функциональный анализ (Конвей 1994, Hsing & Eubank 2015); и случайные процессы (Ash & Gardner 1975). Несколько пакетов программ доступны для анализа функциональных данных, в том числе программного обеспечения в функциональных данных. (Аналитический сайт Джеймса Рамсея, пакет ФАД из проекта CRAN (комплексная сеть архивов R) с использованием программного языка R и среды (R Core Team 2013). Пакет Matlab PACE на сайте статистического факультета Калифорнийского университета, Дэвис, а некоторые функции также доступны в версии R, FDApace.)

3.2. Примеры функциональных данных

На рис. 3.2 представлен пример типа данных, которые мы рассмотрим. Он показывает рост 10 девушек, измеренных 31 раз [118]. Хотя каждая запись включает в себя только дискретные значения, эти значения отражают плавное изменение высоты, которое можно оценить, в принципе, так часто, как хотелось бы, и, следовательно, является функцией высоты. Таким образом, данные состоят из выборки из 10 функциональных наблюдений роста. В этих данных есть свойства, слишком тонкие, чтобы их можно было увидеть на графике такого типа.

На рис. 3.3 показаны кривые ускорения D^2H_i, рассчитанные по этим данным [119]. Мы используем обозначение дифференциального оператора D

$$D^2H = \frac{d^2H}{dt^2}.$$

На рис. 3.3 скачок роста проявляется в виде сильного положительного ускорения, которое сопровождается резким отрицательным замедлением. В большинстве записей также наблюдается скачок около шести лет, который называется серединой рывка. Поэтому мы заключаем, что некоторые вариации от кривой к кривой могут объясняться на уровне определенных производных. Тот факт, что производные представляют интерес — это еще одна причина думать о данных как о функциях, а не векторах наблюдений в дискретном времени.

Сами возрасты также должны играть явную роль в нашем анализе, потому что они не одинаково разнесены. Хотя соотнести рост в возрасте 9, 10 и 10,5 лет может быть тоже интересно. Действительно, хотя в этом конкретном примере возраст, в котором проводятся наблюдения, номинально одинаков для каждой девушки, в этом нет особой необходимости.

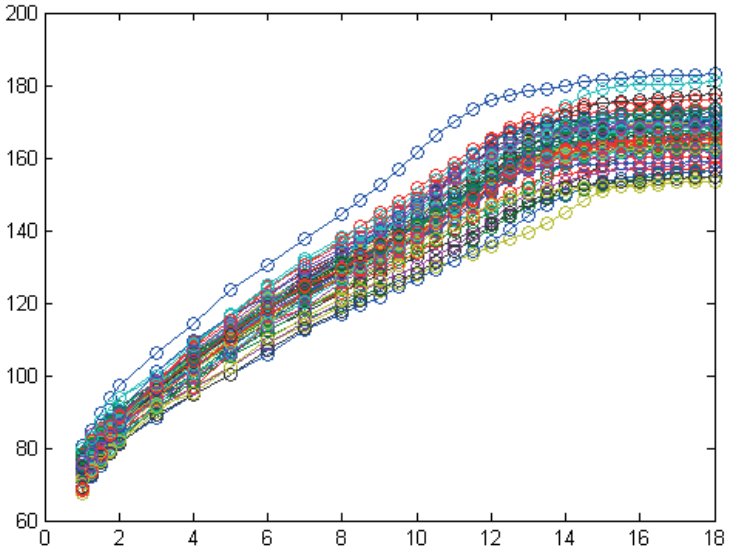


Рис. 3.2. Рост 10 девушек

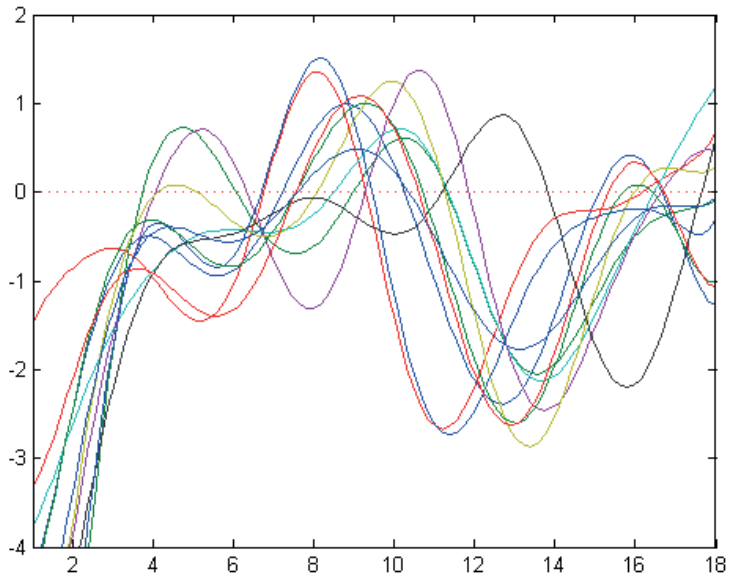


Рис. 3.3. Кривые ускорения роста девушек

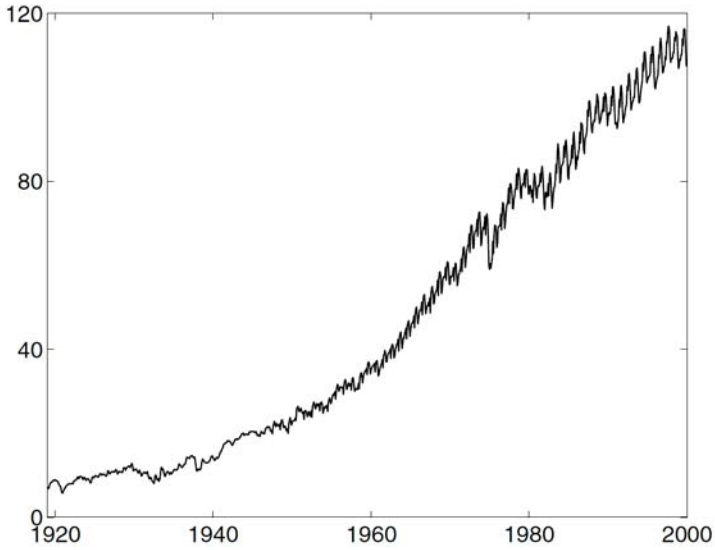


Рис. 3.4. Индекс производства товаров длительного пользования

Точки, в которых наблюдаются функции, могут сильно отличаться.

Репликация этих кривых высоты предлагает исследование путей, в которых кривые меняются. Это потенциально сложно. Например, быстрый рост в период полового созревания виден на всех кривых, но время и интенсивность роста отличается от девочки к девочке. Какой-то тип анализа основных компонентов, несомненно, будет полезным, но необходимо адаптировать процедуру с учетом неравного возрастного интервала и учитывать гладкость функций высоты.

Не все функциональные данные включают независимые записи; часто приходится работать с одной длинной записью. Рис. 3.4 показывает важный экономический индикатор, такой как индекс производства товаров длительного пользования для США. Данные, подобные этим, часто отражают вариации в виде нескольких уровней. Есть тенденция для индекса, чтобы выявить геометрическое или экспоненциальное увеличение по всему периоду. Но в более мелком масштабе мы видим отклонения от этой тенденции вследствие Великой депрессии, Второй мировой войны, конца войны во Вьетнаме и других локальных событий. Более того, в еще более тонком масштабе вариации мы можем задаться вопросом, показывает ли сама эта сезонная тенденция долгосрочные изменения. Хотя здесь нет независимых копий, есть еще много повторений информации, которую мы можем использовать, чтобы получить стабильные оценки интересных характеристик кривой.

Функциональные данные также возникают в виде пар «ввод / вывод», таких как данные в рис. 3.5, собранные на нефтеперерабатывающем заво-

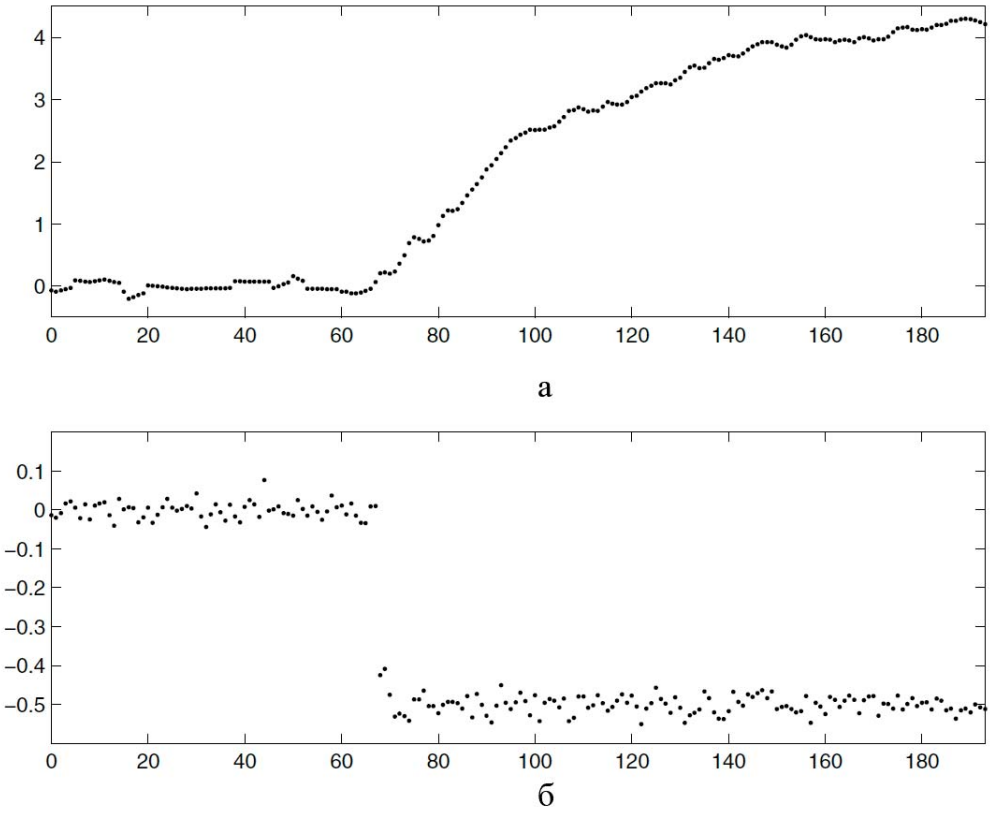


Рис. 3.5. Количество нефти в ректификационной колонне (а), изменение потока пара в поддоне (б)

де. Количество нефти в ректификационной колонне или крекинг-колонне приводится на рис. 3.5,а, изменение потока пара в поддоне показано на нижней части рис. 3.5,б. Очевидно, что уровень нефти реагирует на изменение потока пара. Заметим, что линейные модели не описывают эти отношения, но эти зависимости можно явно описать довольно простыми соотношениями.

3.3. Функциональные модели данных

Представленные выше примеры данных, по-видимому, заслуживают ярлыка «функциональный», поскольку они настолько четко отражают гладкие кривые, что мы предполагаем, что они так и были «созданы». Но не все данные, подлежащие функциональному анализу, сами по себе являются функциональными.

Рассмотрим задачу оценки функции плотности вероятности p для описания распределения выборки наблюдений $x_1, \ldots, x_n$. Классический подход к этой проблеме заключается в том, что после рассмотрения

основных принципов и внимательного изучения данных, параметрическая модель со значениями $p(x|\theta)$ определена фиксированным и обычно с небольшим количеством параметров вектора θ. Например, мы могли бы рассмотреть нормальное распределение как подходящее для данных, так что $\theta = (\mu, \sigma^2)$. Сами параметры обычно выбираются для описания формы функции плотности вероятности и, следовательно, они в центре внимания анализа.

Но предположим, что мы не хотим или не удается заранее выбрать одно из многих стандартных видов функций плотности, потому что, возможно, ни одно из них не показывает особенности поведения данных, которые мы можем видеть на гистограммах или других графических представлениях функций плотности вероятности. Непараметрические методы оценки плотности предполагают только гладкость и допускают большую гибкость в оценке $p(x)$, как того требуют данные. Безусловно, параметры часто используются в методах оценки плотности вероятности, но количество параметров анализа данных не фиксируется заранее, и наше внимание сосредоточено на самой функции p, а не на значениях параметров. Большая часть технологии для оценки гладких функциональных параметров была изначально разработана и отточена в контексте оценки функции плотности.

В качестве примера отметим, что результаты анализа психического теста, как правило, сильно зависят от выбора функциональных моделей, на первый взгляд, нефункциональных данных. Данные обычно являются дискретными и указывают на неудачные и правильные ответы на тестовые задания, но модели состоят из набора функций ответа на вопросы, по одной на элемент теста, отображая гладкую связь между вероятностью успеха на предмет и предполагаемой скрытой способностью.

Рис. 3.6 показывает три таких функциональных параметра для теста по математике, оцениваемых по функциональному методу анализа данных [119].

Пример функционального анализа данных

Данные во многих практических областях приходят к нам через процесс, который естественно описывается как функциональный. Чтобы перейти к совершенно другому контексту, рассмотрим рис. 3.7 [119], где построены среднемесячные температуры для четырех метеостанций. Он показывает оценки соответствующих гладких функций температуры. Ме-

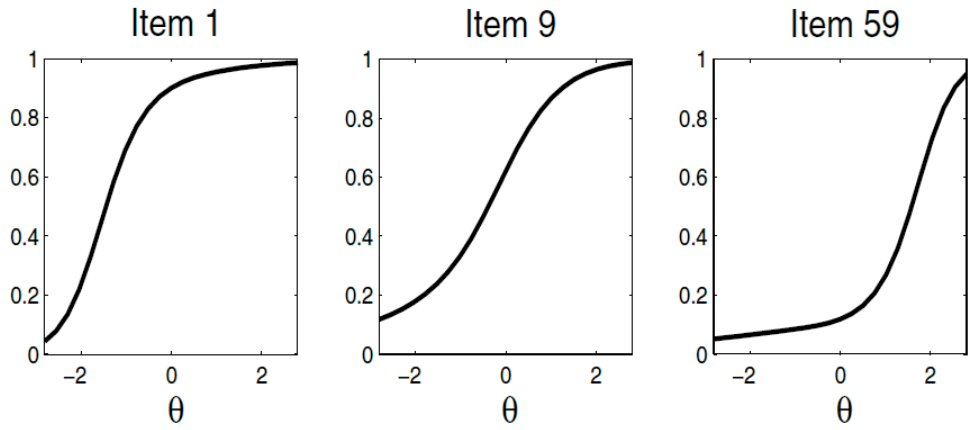


Рис. 3.6. Функции ответа на вопрос

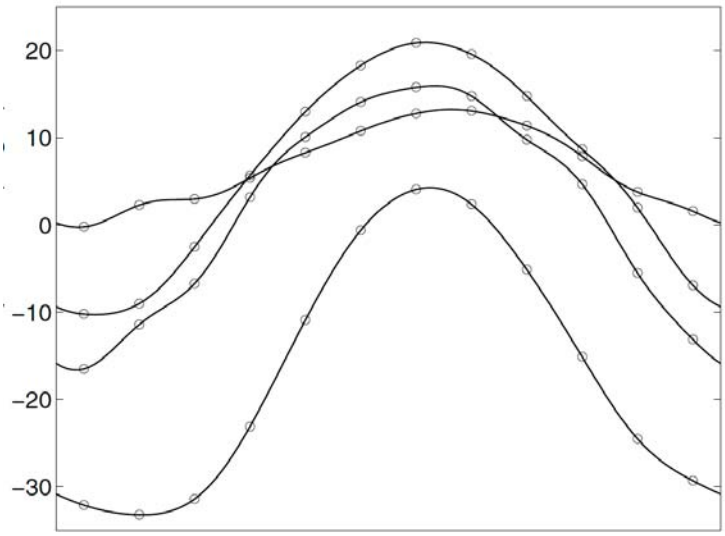


Рис. 3.7. Среднемесячные температуры для метеостанций

теостанция в г. Монреале с самой теплой летней температурой имеет красивый синусоидальный вид функции температур. В г. Эдмонтоне со следующей самой теплой летней температурой наблюдаются некоторые отличительные отклонения от синусоидального вида, обусловленные морским климатом.

Можно ожидать что температура будет в основном синусоидальной по своему характеру и, безусловно, периодической в течение годового цикла. Есть некоторые изменения в фазе, потому что самый холодный день в году окажется позже в Монреале, чем в Эдмонтоне. Следовательно, модель для этих данных имеет вид

$$T_i(t) \approx c_{i1} + c_{i2}\sin(\pi t/6) + c_{i3}\cos(\pi t/6), \tag{3.1}$$

где T_i — функция температуры для i-й метеостанции, и (c_{i1}, c_{i2}, c_{i3}) пред-

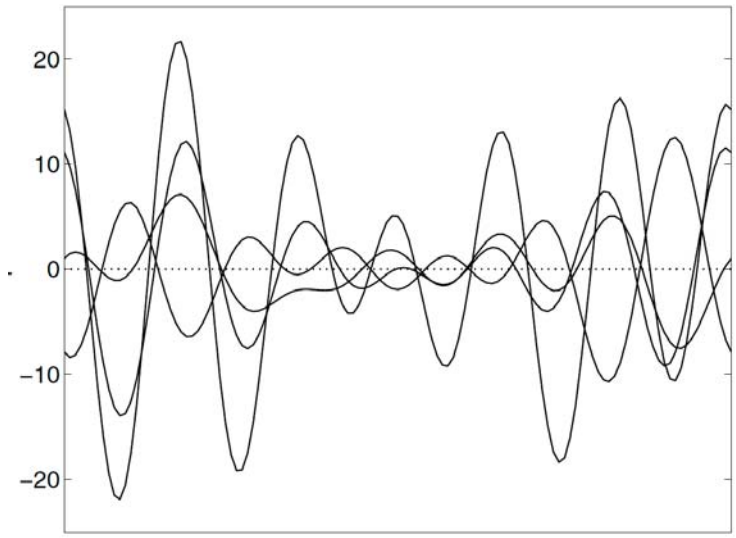


Рис. 3.8. Результат применения дифференциального оператора L

ставляет собой вектор трех параметров, связанных с этой станцией.

Обозначение LT — это функция, получающаяся в результате применения линейного дифференциального оператора $L = (\pi/6)^2 D + D^3$ к функции T. Теперь, если температурная функция действительно синусоидальная, тогда LT должен быть точно нулем, как это было бы для любой функции вида (3.1). То есть это будет соответствовать дифференциальному уравнению:

$$D^3 T = -(\pi/6)^2 DT.$$

Но рис. 3.8 указывает на то, что функции LT_i отображают систематические особенности, которые особенно сильны в весенние и осенние месяцы. Заметим, что температура на конкретной метеостанции может быть описана как решение неоднородного дифференциального уравнения, соответствующего в $LT = u$, где форсирующую функцию u можно рассматривать вне системы, или экзогенное влияние. Метеорологи предполагают, например, что эти весенние и осенние эффекты частично связаны с изменением отражательной способности земли, когда снег или лед тает, и это будет соответствовать тому факту, что наименее синусоидальные записи связаны с континентальными станциями, далеко отделенными от больших водоемов.

Здесь дело в том, что нам часто бывает интересно, используя внутреннюю гладкость в процессе и применяя дифференциальный оператор, удалить эффекты простой природы. Многолетний опыт в области естественных и технических наук позволяет предположить, что это может

стать ближе к основным факторам в работе, чем просто добавление и вычитание эффектов, как это обычно делается в многомерном анализе данных.

3.4. Цели функционального анализа данных

Цели функционального анализа данных в основном такие же, как и у любой другой ветви статистики. Они включают в себя:

- представление данных способами, которые помогают дальнейшему анализу;
- отображение данных, чтобы выделить различные характеристики;
- изучение важных источников закономерностей и вариаций среди данных;
- объяснение изменения в результатах с помощью входной информации;
- сравнение двух или более наборов данных в отношении определенных типов вариации, где два набора данных могут содержать разные наборы дубликатов одних и тех же функций.

3.5. Функциональная регрессия

Функциональная регрессия — это версия регрессионного анализа, когда предикторы (независимые переменные) и переменные отклика могут быть функциями [102]. Модели функциональной регрессии могут быть классифицированы по типу зависимостей, являются ли выходные переменные функциональными или скалярными. Кроме того, модели функциональной регрессии могут быть линейными, частично линейными или нелинейными. В частности, функциональные полиномиальные модели, функциональные модели с одним и несколькими индексами и функциональные аддитивные модели являются тремя частными случаями функциональных нелинейных моделей.

Функциональные линейные модели

Функциональные линейные модели (ФЛМ) являются продолжением линейных моделей. Линейную модель со скалярным откликом $Y \in \mathbb{R}$ и скалярными предикторами $X \in \mathbb{R}^p$ можно записать как

$$Y = \beta_0 + \langle X, \beta \rangle + \varepsilon, \tag{3.2}$$

где $\langle \cdot, \cdot \rangle$ обозначает скалярное произведение в евклидовом пространстве, $\beta_0 \in \mathbb{R}$ и $\beta \in \mathbb{R}^p$ обозначают коэффициенты регрессии, а ε — случайная ошибка со средним нулем и конечной дисперсией. FLM могут быть разделены на два типа на основе ответов.

Функциональные линейные модели со скалярными откликами

Функциональные линейные модели со скалярными откликами можно получить, заменив скалярные предикторы X и вектор коэффициентов β в модели (3.2) на центрированный функциональный предиктор

$$X^c(\cdot) = X(\cdot) - \mathsf{M}(X(\cdot)),$$

и функция коэффициентов $\beta = \beta(\cdot)$ с областью $\mathcal{T}$ соответственно и замена скалярного произведения в евклидовом пространстве на произведение в гильбертовом пространстве L_2,

$$Y = \beta_0 + \langle X^c, \beta \rangle + \varepsilon = \beta_0 + \int_{\mathcal{T}} X^c(t)\beta(t)\, dt + \varepsilon, \tag{3.3}$$

где $\langle \cdot, \cdot \rangle$ здесь обозначает скалярное произведение в L^2. Один из подходов к оценке β_0 и $\beta(\cdot)$ заключается в расширении центрированного предиктора $X^c(\cdot)$ и функции коэффициентов $\beta(\cdot)$ на той же функциональной основе, например, опираясь на базис в пространстве сплайнов или базис, используемый в разложении Karhunen–Loéve. Предположим, $\{\phi_k\}_{k=1}^{\infty}$ является ортонормированным базисом в L^2.

Расширяя X^c и β на этой основе,

$$X^c(\cdot) = \sum_{k=1}^{\infty} x_k \phi_k(\cdot), \beta(\cdot) = \sum_{k=1}^{\infty} \beta_k \phi_k(\cdot),$$

модель (3.3) становится

$$Y = \beta_0 + \sum_{k=1}^{\infty} \beta_k x_k + \varepsilon.$$

Для реализации необходима регуляризация, которая может быть выполнена путем усечения L^2 или L^1 [102]. Кроме того, подход воспроизведения ядра гильбертова пространства (Reproducing kernel Hilbert space RKHS) также можно использовать для оценки β_0 и $\beta(\cdot)$ в модели (3.3).

Добавляя несколько функциональных и скалярных предикторов, модель (3.3) может быть расширена до

$$Y = \sum_{k=1}^{q} Z_k \alpha_k + \sum_{j=1}^{p} \int_{\mathcal{T}_j} X_j^c(t)\beta_j(t)\, dt + \varepsilon, \tag{3.4}$$

где $Z_1, \ldots, Z_q$ являются скалярными предикторами с $Z_1 = 1$, $\alpha_1, \ldots, \alpha_q$ являются коэффициентами регрессии для $Z_1, \ldots, Z_q$ соответственно, X_j^c является центрированным функциональным предиктором, заданным $X_j^c(\cdot) = X_j(\cdot) - \mathbb{E}(X_j(\cdot))$, β_j является функцией коэффициента регрессии для $X_j^c(\cdot)$ и $\mathcal{T}_j$ является доменом X_j и β_j для $j = 1, \ldots, p$. Однако из-за параметрического компонента α методы оценки для модели (3.3) не могут быть использованы в этом случае, и доступны альтернативные методы оценки для модели (3.4) [141].

Функциональные линейные модели с функциональными реакциями

Для функционального ответа $Y(\cdot)$ с областью $\mathcal{T}$ и функциональным предиктором $X(\cdot)$ с областью $\mathcal{S}$ две функциональные линейные модели с $Y(\cdot)$ $X(\cdot)$ были рассмотрены в [119, 141]. Одна из этих двух моделей имеет вид

$$Y(t) = \beta_0(t) + \int_{\mathcal{S}} \beta(s,t) X^c(s)\, ds + \varepsilon(t), \text{ for } t \in \mathcal{T}, \tag{3.5}$$

где $X^c(\cdot) = X(\cdot) - \mathbb{E}(X(\cdot))$ по-прежнему является центрированным функциональным предиктором, $\beta_0(\cdot)$, $\beta(\cdot,\cdot)$ являются функциями коэффициентов, а $\varepsilon(\cdot)$ обычно считается случайным процессом со средним нулем и конечной дисперсией. В этом случае в любой момент времени $t \in \mathcal{T}$ значение Y, т. е. $Y(t)$, зависит от всей траектории X. Модель (3.5) для любого заданного времени t является расширением многомерной линейной регрессии с скалярным произведением в евклидовом пространстве и заменяется на соответствующее в L^2. Оценочное уравнение, мотивированное многомерной линейной регрессией,

$$r_{XY} = R_{XX}\beta, \text{ for } \beta \in L^2(\mathcal{S} \times \mathcal{S}),$$

где $r_{XY}(s,t) = \mathrm{cov}(X(s), Y(t))$,

$R_{XX} : L^2(\mathcal{S} \times \mathcal{S}) \to L^2(\mathcal{S} \times \mathcal{T})$ определяется как

$$(R_{XX}\beta)(s,t) = \int_{\mathcal{S}} r_{XX}(s,w)\beta(w,t)dw,$$
$$r_{XX}(s,w) = \mathrm{cov}(X(s), X(w))s, w \in \mathcal{S}.$$

Когда X и Y наблюдаются одновременно, т. е. $\mathcal{S} = \mathcal{T}$, [91], целесообразно рассмотреть историческую функциональную линейную модель, где текущее значение Y зависит только от истории X, т. е. $\beta(s,t) = 0$ для $s > t$ в модели (3.5) [119, 141]. Более простой версией исторической функциональной линейной модели является функциональная параллельная модель.

После добавления нескольких функциональных предикторов модель (3.5) может быть расширена до

$$Y(t) = \beta_0(t) + \sum_{j=1}^{p} \int_{\mathcal{S}_j} \beta_j(s,t) X_j^c(s)\, ds + \varepsilon(t),\ t \in \mathcal{T},$$

где для $j = 1, \ldots, p$, $X_j^c(\cdot) = X_j(\cdot) - \mathsf{M}(X_j(\cdot))$ является центрированным функциональным предиктором с областью $\mathcal{S}_j$ $\beta_j(\cdot,\cdot)$.

Функциональные совместные модели

Предполагая, что $\mathcal{S} = \mathcal{T}$, другая модель, известная как функциональная совместная модель, иногда называемая моделью с переменным коэффициентом, имеет вид

$$Y(t) = \alpha_0(t) + \alpha(t)X(t) + \varepsilon(t),\ \text{for } t \in \mathcal{T}, \tag{3.6}$$

где α_0 и α являются функциями коэффициентов [145].

3.6. Прогноз плотности

В последние годы наблюдается значительный интерес к прогнозам плотности. Это было вызвано быстро расширяющейся областью управления финансовыми рисками и прогнозированием инфляции. Например, если целью является достижение уровня инфляции в определенном диапазоне или целевой полосе, точечный прогноз инфляции имеет ограниченную ценность. Кусочно-полиномиальная функция прогноза плотности плотности будет более информативной оценкой вероятности достижения цели. Тем не менее кусочно-полиномиальный прогноз плотности

будет иметь значение только в той степени, в которой вероятности прогноза точно отражают истинные вероятности. Внимание будет уделено оценке, где прогнозы являются кусочно-полиномиальный аппроксимацией плотности вероятностей.

Точечный прогноз против прогноза плотности

Прогноз плотности дает полное описание неопределенности и отличается от точечного прогноза, который его не содержит.

Согласно Elliott и Timmermann [79], точечные прогнозы применяются в качестве решения такой проблемы, как указано ниже:

$$\min_{f(z)} \int L(f(z), y) p_Y(y|z) dy.$$

Это зависит как от функции потерь $L(f(z), y)$, так и от плотности результата, зависящей от имеющихся данных, $p_Y(y|z)$. Точечный прогноз является особенностью прогнозируемой плотности, $p_Y(y|z)$, и в большинстве случаев этот вид сводной статистики достаточен для большинства пользователей.

Почему в какой-то ситуации мы должны выбирать прогноз плотности, а не точечный прогноз?

Во-первых, прогнозы будут использоваться различными пользователями с различными функциями потерь (точечный прогноз — специфическая функция потерь). Таким образом, прогнозируемая плотность может быть объединена с функцией потерь.

Во-вторых, точечный прогноз дает мало информации о точности прогноза (пример инфляции — прогноз сообщается политическому деятелю).

В-третьих, ситуации, когда люди интересуются многоэтапным прогнозированием по нелинейным моделям. Полная плотность имеет значение всякий раз, когда мы выполняем итерацию на модели нелинейного прогнозирования, поскольку нелинейные эффекты обычно не зависят только от условного среднего, но также и от того, где (в наборе возможных результатов) встречаются будущие значения.

Основная проблема прогнозирования плотности

Elliott и Timmermann [79] указали, что основная проблема прогнозирования плотности связана с одним результатом переменной y_{t+1}; и условные переменные, z_t, и прогноз плотности очерчивают условное распределение y_{t+1}, учитывая Z_t:

$$p_Y(y_{t+1}|Z_t) = \mathsf{P}(y_{t+1}|Z_t).$$

Как видно из приведенного выше уравнения, условное распределение (или прогнозная плотность) использует все соответствующие переменные в наборе информации.

На практике обычно используется подмножество переменных кондиционирования. В простейшем случае, когда рассматривается только авторегрессия динамики, это будет включать в себя прошлую историю переменной как таковой.

Построение метода

Очень популярный двухэтапный подход к построению прогноза плотности сделан Elliott и Timmermann [79]:

во-первых, смоделируйте любую условную динамику в средней и условной волатильности ряда;

во-вторых, примените гибкие параметрические или непараметрические методы для моделирования «нормализованного» остатка, полученного после вычитания условного среднего и деления на условную волатильность.

Оценка плотности может быть либо параметрической, если данные взяты из известного набора данных, либо непараметрической, когда строится оценка неизвестного распределения.

Одним из ключевых инструментов по оценке прогноза плотности является интегральное преобразование вероятности M. Clements [61]. Предположим, у нас есть серия из 1-го шага прогноза плотности для значения случайной величины $\{Y_t\}$, обозначенной $p_{Y,t-1}(y)$, где $t = 1, \ldots, n$. Вероятностное интегральное преобразование переменной относительно плотностей прогноза:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(y_t) \tag{3.7}$$

для $t = 1, \ldots, n$, где $P_{Y,t-1}(y_t)$ — вероятность прогноза Y_t. В терминах случайных величин $\{Y_t\}$, вместо их реализованных значений $\{y_t\}$, мы получаем случайные величины $\{Z_t\}$:

$$Z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(Y_t).$$

Когда прогнозируемая плотность равна истинной плотности, $f_{Y,t-1}(y)$, следует, что $Z_t \sim U(0,1)$, где $U(0,1)$ — равномерное распределение на (0, 1). Хотя фактические условные плотности могут меняться со временем, при условии, что прогнозируемые плотности соответствуют фактическим значениям плотности при каждом t, тогда $Z_t \sim U(0,1)$ для каждого t и Z_t независимо распределены, а реализованный временной ряд $\{z_t\}_{t=1}^n$ распределен на $U(0,1)$.

Если мы обозначим $q_{Z,t}(z)$ плотность Z_t, то получим результат, $q_{Z,t}(\cdot)$ — равномерное распределение. Для $Z_t = P_{Y,t-1)}(Y_t)$ получаем $q_{Z,t}(\cdot)$:

$$q_{Z,t}(z) = \frac{f_{Y,t-1}(P_{Y,t-1}^{-1}(z))}{p_{Y,t-1}(P_{Y,t-1}^{-1}(z))}.$$

Когда $f_{Y,t-1}(\cdot)$ и $p_{Y,t-1}(\cdot)$ одинаковы, $q_{Z,t}(z) = 1$ для $z \in [0,1]$, т. е. $Z_t \sin U(0,1)$. Это верно для каждого $t, t = 1, \ldots, n$, так что время индекса на $q_{Z,t}$ избыточно.

Это говорит о том, что мы можем оценить, является ли условным прогноз плотности, и сопоставить истинные условные плотности, проверяя, являются ли $\{z_t\}_{t=1}^n$ независимыми и одинаково распределенными на $U(0,1)$. Независимость можно оценить, изучив $\{z_t - \bar{z}\}$. Могут быть выполнены формальные испытания автокорреляции.

Применение прогноза плотности в макроэкономике

Tay and Wallis [133] рассмотрели обзор применения прогноза плотности в экономике. Ранее известный прогноз плотности проводился Федеральным резервным банком Филадельфии с 1968 года под названием «Survey of Professional Forecasters». Помимо вопросов, задаваемых для точечных прогнозов, также запрашиваются прогнозы плотности для инфляции и роста производства. В каждом случае у каждого прогнозиста просят сообщить свой соответствующий прогноз вероятности, т. е. количество интервалов, в которые будущие значения переменных могут попасть.

Таким образом, респонденты предоставляют прогнозы плотности двух переменных в виде гистограмм вероятности, затем усредняются по респондентам для получения прогнозов средней плотности, снова в виде гистограмм, которые публикуются. Другой более свежий пример прогноза плотности инфляции показан в табл. 1 ниже.

Прогнозы плотности инфляции в США. Средняя вероятность (из 28 прогнозистов) прилагается к возможному процентному изменению индекса цен ВВП, 1998–1999 гг.

Таблица 1

Уровень инфляции (%)	Вероятность (%)
8,0 или более	0,07
7,0–7,9	0,11
6,0–6,9	0,14
5,0–5,9	0,25
4,0–4,9	1,21
3,0–3,9	8,96
2,0–2,9	20
1,0–1,9	49,54
0,0–0,9	17,89

Источник: Philadelphia Fed., Survey of Professional Forecasters, 20 November 1998.

Представление прогнозов плотности

Часто возникающая проблема в прогнозировании плотности заключается в том, как сообщить прогнозную плотность. Балльная оценка обычно требует сообщать только одно число, например, прогнозируемое количество рабочих мест, созданных в следующем году, может быть 1 миллион. Это число, легко сообщаемое пользователю прогноза.

С другой стороны, Elliott и Timmermann [79] подчеркнули, что для прогнозирования плотности требуется знание функции распределения. Если прогноз плотности параметрический, функциональная форма плотности вместе с расчетными параметрами может быть представлена, но для непараметрических подходов это становится намного сложнее. Часто только частичная информация передается в конце процесса прогнозирования, например, график распределения прогноза.

Иногда возможно представить аналитическую форму прогнозов плотности, но они доступны только в том случае, если используются стандартные распределения, и в любом случае особенности прогноза могут быть не сразу очевидны из алгебраических выражений. Более распространенный способ представления прогнозов плотности — построение графика плотности. Обычно это относится к прогнозам, полученным из полупараметрических подходов к плотности. Часто полезно дискретизировать плотность, представив ее в виде гистограммы, графически или в виде таблицы.

Глава 4

Символьный анализ данных

Символьный анализ данных (Symbolic Data Analysis) обеспечивает основу для представления и анализа данных, в которых присутствует внутренняя изменчивость. В то время как в Data Mining и в классической статистике анализируемые данные обычно представляют одно значение для каждой переменной, это больше тот случай, когда анализируемые объекты собраны в группы на основе некоторых заданных критериев. Тогда каждая переменная содержит изменчивость, присущую каждой группе. При исследовании таких разделов реального мира, как ботанические виды, описания болезней, модели автомобилей и т. д., данные содержат в себе внутреннюю изменчивость, которая должна быть явно учтена. Для анализа таких данных были введены новые типы переменных, реализация которых — не отдельные реальные значения или категории, но наборы, интервалы или, в более общем смысле, распределения в данной области. Символьный анализ данных (САД) предоставляет методы для (многомерного) анализа данных, где принимается во внимание изменчивость, выраженная в представлении данных.

В интеллектуальном анализе данных, многомерном анализе данных и классической статистике данные для анализа обычно представлены в массиве данных $n \times p$, где каждая строка представляет собой субъект («дело» или «физическое лицо»), каждый столбец относится к переменной (также называемой «атрибут»), которая может быть числовой или категориальной, и одно значение записывается для каждой переменной и для каждого из сущностей. Эта модель представления имеет ограничения, когда анализируемые данные имеют изменчивость. Это тот случай, когда субъекты анализа не отдельные элементы, а группы, которые формируются на основе некоторых заданных общих свойств. Затем для каждой переменной наблюдаемая изменчивость, присущая каждой группе,

должна приниматься во внимание, чтобы избежать потери важной информации. Такие случаи имеют место при анализе концепций как таковых — ботанический вид, а не данный образец; модель автомобиля, а не конкретное транспортное средство, и т. д., опять же, изменчивость присуща данным, которые должны быть явно учтены. Например, предположим, что при исследовании деятельности аэропорта необходимы данные, которые собираются для каждого рейса, прибывающего в разные аэропорты, (количество пассажиров, задержка прибытия, авиастроительная компания и т. д.). В качестве статистических единиц интерес представляют аэропорты и не каждый отдельный рейс: данные, касающиеся рейсов, прибывающие в один и тот же аэропорт, должны как-то агрегироваться. Похожая ситуация возникает, когда проводится исследование для сравнения разных школ, и данные собираются для отдельных студентов — возраст, пол, оценки на разных экзаменах и т. д. Далее данные отдельных студентов должны быть агрегированы для соответствующей школы, так что описание каждой школы может быть получено и впоследствии проанализировано и сравнено. В ситуациях такого рода, где статистическая информация находится на более высоком уровне, чем тот, на котором были собраны данные, агрегирование наблюдаемых значений должно быть выполнено до анализа данных. Стандартный подход заключается в вычислении суммы показателей, таких как средние значения, медианы или моды — так, чтобы данные помещались в обычный массив данных $n \times p$. Эта модель и классические методы могут быть применены, однако это часто влечет за собой потерю важной информации.

Анализ символьных данных [58] обеспечивает основу для представления и систематизации данных с присущей им изменчивостью. Для этого были введены новые типы переменных, реализация которых есть не единичные реальные значения или категории, но наборы, интервалы или, в более общем смысле, распределения по данным областям. Естественно, анализ таких данных ставит новые вопросы, потому что большинство концепций и методов в первую очередь предназначены для однозначных наблюдений. До сих пор, тем не менее, много методов для (многомерного) анализа символьных данных были разработаны с учетом различных подходов и используя четкие критерии, которые позволяют принимать во внимание изменчивость в представлении данных.

Другой подход, при котором данные представляются в агрегированном виде — Granular Computing (см., например, [112]). Информационные гранулы определяются как группы отдельных наблюдений, которые

отражают семантику абстрактных объектов, представляющих интерес. Как правило, с учетом набора данных D, в результате грануляции получается набор гранул, образованных на основе сходства или близости, которая может быть достигнута, например, с помощью алгоритмов кластеризации. Когда данные — числовые, гранулы часто принимают форму гиперкубов. Информационные гранулы, представленные в теории нечетких множеств, отображены с помощью функции принадлежности.

Конечно, есть общие точки зрения между этими двумя подходами в проведении агрегации исходных данных. Тем не менее, могут быть указаны четкие различия. В САД исходное лицо и агрегированные данные представлены в одной и той же структуре данных — массиве $n \times p$, где новые типы переменных описывают сформированные группы, выражая их явно в пределах изменчивости. Как прямое следствие, методы анализа, предназначенные для символьных данных, хотя, возможно, стремятся к той же цели (например, кластеризация, классификация и т. д.), полагаются на разные свойства.

В следующем разделе мы представим и мотивируем появление ФАД более подробно и обсудим разные источники символьных данных.

4.1. Символьные данные

Символьные данные (Symbolic data), т. е. данные, которые содержат внутреннюю изменчивость, возникают в разных контекстах. В большинстве случаев из совокупности отдельных наблюдений — называемых микроданные.

Мы можем выделить два разных типа агрегации:

• временная агрегация: когда данные записываются в разные моменты времени для одного и того же физического лица — например, покупки, сделанные в данном магазине, но при этом не учитывая время (т. е. мы не заинтересованы в хронологическом порядке наблюдений). Результаты затем должны быть агрегированы так, чтобы использовался весь набор значений (или их распределений), и при этом сохраняется не только среднее значение, медиана или значение моды. В этой ситуации наблюдения проводятся по времени, и анализируемые статистические единицы («случаи», т. е. клиенты в данном примере) остаются неизменными до и после агрегации;

• одновременная агрегация: это касается ситуации, когда данные записаны в одной точке времени, но мы заинтересованы в анализе сущно-

стей на более высоком уровне, чем тот, на котором они были изначально собраны, например когда собираются данные для студентов, и мы заинтересованы в сравнении класса или школы в целом. В этом случае статистические единицы, которые будут проанализированы, составляют конкретные группы. Обратите внимание, что этот случай также включает в себя ситуацию пространственной агрегации, когда данные собираются в одной точке для статистических единиц — например, отдельных граждан, как в официальных статистических обзорах — по разным регионам, а затем данные агрегируются на уровне региона.

В качестве иллюстрации временной агрегации рассмотрим три человека, Альберта, Барбару и Кэролайн, которые характеризуются количеством времени (в мин), которое они тратят на дорогу на работу. Это меняется изо дня в день, и наблюдаемое изменение может быть выражено интервалами как в табл. 2.

Таблица 2

Персона	Время (min)
Альберт	[15,20]
Барбара	[25,30]
Кэролайн	[10,20]

Другими такими ситуациями являются, например:

• артериальное давление неоднократно измеряется для разных пациентов;

• некоторые технические измерения, которые взяты в разных точках данных объектов.

Таблица 3

Школы	Возраст	Пол	Оценки
A	[12,14]	{F,M}	{<10, (0.2); [10–15], (0.6);>15, (0.2)}
B	[11,13]	{F}	{<10, (0.3); [10–15], (0.3);>15, (0.4)}

Во всех этих случаях статистические единицы одинаковы до и после агрегирования данных.

Теперь предположим, как упоминалось выше, что исследование проводится в разных школах, и эти данные собираются для учащихся, посещающих эти школы, например возраст, пол, оценки на разных экзаменах и т. д. Для статистики представляют интерес школы, а не отдельные студенты, и, следовательно, данные, касающихся учащихся, посещающих

одну и ту же школу, должны быть объединены. В табл. 3 представлены данные, агрегированные по школам.

Другие примеры подобных ситуаций могут быть упомянуты: когда данные собираются для отдельных игроков, но изучены будут команды в целом; когда данные собираются для отдельных граждан (как в официальных статистических опросах), но исследования должны проводиться на уровне городов, специальных групп по интересам и пр.

Это особенно интересно в приложениях Data Mining, в которых собираются огромные наборы данных, и они должны быть проанализированы на более высоком уровне (супермаркеты или большой отдел магазина где записывают данные о каждой совершенной покупке, например, потраченной сумме, приобретенных товарах, количестве каждого предмета и т. д.). Как правило, администрация и исследователи маркетинга не особенно заинтересованы в отдельных покупках, а скорее в потребительском поведении. То есть они хотят иметь информацию об общих покупках каждого клиента или определенных групп клиентов. Чтобы получить такую информацию, данные, собранные об отдельных покупках, должны быть агрегированы. Другие примеры, рассматриваемые в исследованиях Data Mining, касаются:

• данных об отдельных телефонных звонках, совокупность которых оператор связи хочет проанализировать на уровне клиента;

• данных о веб-журналах, агрегированных пользователем или вебсайтом;

• данных о медицинских рецептах, представляющих интерес для врачей.

Структура САД предоставляет возможность агрегировать отдельные «микроданные», которые сохраняют изменчивость записей.

В последние годы появился термин «большие данные», наборы данных, настолько большие и сложные, что они становятся трудными для обработки традиционными методами анализ заявок за разумное время. САД предлагет возможность агрегирования данных на выбранной пользователем степени детализации, в то время как хранение информации о внутренней изменчивости, а затем анализ полученных (символьных) массивов данных может сыграть важную роль в этом контексте.

Giordano and Brito [90] используют САД для исследования и сравнения социальных сетей. В этой работе символьное описание сети определяется согласно статистической характеристике сети и ее топологическим

свойствам. Анализ многомерных данных позволяет использовать представление сети как точки в метрическом пространстве и последующий анализ (например, кластеризация).

САД также могут представлять интерес для (большого) анализа опросов, когда наблюдаемый образец делится на конкретные группы, которые находятся в центре внимания. Это может иметь место, например, в социологических или маркетинговых обследованиях, когда группы, определенные, например, по возрасту, полу, уровню образования и/или профессиональному статусу должны сравниваться и анализироваться вместе. Кроме того, САД позволяет объединять независимые опросы, сделанные на том же населении, на макроуровне. В этом случае «микроданные» нельзя анализировать вместе, так как наблюдаемые лица в различных опросах не являются идентичными. Объединяя соответствующие опросы, используя одни и те же критерии (т. е. формируя те же «группы»), мы получаем данные, которые могут быть собраны вместе.

4.2. Типы переменных

Для представления изменчивости данных в САД указаны новые типы переменных, их реализации в настоящее время не ограничиваются реальными значениями (в числовом случае) или отдельными категориями (в качественном случае). Рассматриваются разные переменные типы, в том числе классические, которые могут изучаться как особые случаи, символьные типы определены ниже.

Как и в классической статистике, мы различаем численные и категориальные переменные. Однозначная переменная (действительная или целая) является числом (или количественной), как и в классических рамках, если она занимает одно значение из базовой области для каждого объекта. Она является многозначной, если ее значения — конечные подмножества области, и является интервальной переменной, если ее значения — интервалы. Когда распределение происходит по заданному набору подинтервалов, переменная называется кусочно-полиномиальной (гистограммной, сплайновой). Категорическая (или качественная) переменная является однозначной (порядковой или номинальной), когда она берет одну категорию из данной конечной категории $O = \{m_1, ..., m_k\}$ для каждой сущности; многозначной, если ее значения являются конечными подмножествами из O. Категориальная модальная переменная Y с конечной областью $O = \{m_1, ..., m_k\}$ равна многозначной переменной, где

для каждого элемента дан набор категорий, и для каждой категории m_l указываем частоту или вероятность.

Пусть $Y_1, ..., Y_p$ — множество переменных, O_j — базовая область Y_j, а B_j — множество, где Y_j принимает его значение для каждой сущности, для $j = 1, ..., p$. Описание d определяется как кортеж $d = (d_1, ..., d_p)$ с $d_j \in B_j$, $j = 1, ..., p$. Пусть $S = \{s_1, ..., s_n\}$ будет множеством сущности (статистические единицы), которое анализируется, тогда $Y_j(s_i) \in Bj$ для $j = 1, ..., p$, $i = 1, ..., n$. Анализируемый массив данных состоит из n описаний, по одному для каждого $s_i \in S : d_i = (Y_1(s_i), ..., Y_p(s_i))$, $i = 1, ..., n$.

4.3. Классические переменные

Количественные однозначные переменные

Учитывая набор из n объектов $S = \{s_1, \ldots, s_n\}$, количественная однозначная переменная Y определяется отображением $Y : S \to O$ таким, что $s_i \mapsto Y(s_i) = c \in O \subset R$. Это классический числовой случай, и B идентичен базовому набору O, $B \equiv O$.

Категориальные однозначные переменные

Категориальная однозначная переменная (Categorical Single-Valued Variables) является стандартной категориальной переменной. Учитывая $S = \{s_1, \ldots, s_n\}$ и конечный набор категорий, $O = \{m_1, \ldots, m_k\}$, категориальная однозначная переменная определяется отображением $Y : S \to O$ таким, что $s_i \mapsto Y(si) = m_l$ (т. е. в этом случае опять $B \equiv O$). Если категории естественно упорядоченные, переменная называется порядковой, в противном случае она является номинальной. Такая категориальная переменная может быть использована для создания новых концепций или объектов путем агрегирования случаев, относящихся к одной категории.

4.4. Новые типы переменных

Количественные многозначные переменные

Дано множество S, количественная многозначная переменная Y определяется отношением $Y : S \to B$ таким, что $s_i \mapsto Y(s_i) = \{c_{i1}, \ldots, c_{in_i}\}$.

Здесь B — множество всех подмножеств базового множества O (исключая пустой набор $\emptyset$). $Y(s_i)$ теперь конечный непустой набор действительных чисел.

Интервальные переменные

Учитывая $S = \{s_1, \ldots, s_n\}$, интервальная переменная определяется отношением $Y : S \to B$ таким, что $s_i \mapsto Y(s_i) = [l_i, u_i]$, B — множество интервалов; базовым набором $O \subseteq R$. Пусть I будет матрица $n \times p$, представляющая значения интервальных переменных на S. Каждый $s_i \in S$ представлен p-кортежем интервалов, $I_i = (I_{i1}, \ldots, I_{ip})$, $i = 1, \ldots, n$, с $I_{ij} = [l_{ij}, u_{ij}]$, $j = 1, \ldots, p$.

Таблица 4

Аэропорты	Число пассажиров	Номера компаний
A	[150,200]	{1, 2}
B	[180,300]	{1, 2, 3}
C	[200,400]	{1, 3}

Пример: рассмотрим набор данных, содержащий информацию о прибывающих рейсах в некоторых аэропортах; в табл. 4 представлены данные трех аэропортов. В аэропорту A, где количество пассажиров на прибывающих рейсах колеблется от 150 до 200, а количество компаний-участников — 1 или 2, количество пассажиров — интервальная переменная, тогда как номера компаний — многозначная количественная переменная. Аналогичное описание может быть получено для остальных аэропортов. Следует подчеркнуть, что в этом примере анализируемые объекты являются аэропортами, для каждого из которых мы собрали информацию, а не индивидуальные рейсы.

Кусочно-полиномиальные переменные

Когда реальные данные агрегируются с помощью интервалов, информация о распределении внутри интервалов не учитывается. Один из способов сохранить более подробную информацию, чтобы определить подинтервалы между глобальными нижними (LB) и верхними (UB) границами и аппроксимировать полиномом плотности для этих интервалов, получаем в случае констант — гистограмму с k-классов (и k-частот),

где k является количеством рассмотренных подинтервалов. Естественно, агрегировать числовые микроданные с помощью гистограммы подразумевает, что достаточно большое количество наблюдений доступно на микроуровне. Дано $S = \{s_1, \ldots, s_n\}$, гистограммная переменная определяется с помощью отображения $Y : S \to B$, такого, что $s_i \mapsto Y(s_i) = \{[l_{i1}, U_{i1}], p_{i1}; \ldots [l_{in}, U_{in}], p_{in}\}$, B — сейчас множество частотных распределений.

Пример: рассмотрим еще раз пример аэропортов, с новой переменной, которая записывает задержку (в мин) каждого прибывающего рейса. В этом случае информация, записанная в течение трех отрезков времени (от 0 до 10 мин, от 10 до 30 мин, от 30 мин до 1 ч), соответствующая переменной, равна гистограммной переменной.

Значение кусочно-полиномиальной переменной может быть эквивалентно представлено эмпирической функцией распределения F или ее обратной, квантильной функцией $\Psi = F^{-1}$. Этот последний вариант часто используется, учитывая, что все квантильные функции определены на [0, 1], что удобно для их сравнения.

4.5. Категориальные многозначные переменные

Категориальная многозначная переменная определяется отношением $Y : S \to B$, где B — множество непустых подмножеств $O = \{m_1, \ldots, m_k\}$. «Значения» $Y(s_i)$ — теперь конечные множества категорий.

Категориальные модальные переменные

Категориальная модальная переменная Y с конечной областью $O = \{m_1, \ldots, m_k\}$ является многозначной переменной, где для каждого элемента нам дают набор категорий, и, для каждой категории m_l, вес, частота или вероятность p_l, который указывает частоту или вероятность категории для этого элемента. Если установлено, что сумма p_l составляет 1, хотя это не обязательно следует из определения, тогда B является набором распределений (вероятности, частот или другое) на O, и его элементы обозначены $\{m_1(p_1), \ldots, m_k(p_k)\}$.

Пример: рассмотрим еще раз пример аэропортов и информацию об основных авиакомпаниях. Затем у нас есть категориальная модальная переменная, как показано в табл. 5.

Таблица 5

Аэропорт	Компании
A	{British (0.25), Lufthansa (0.4), Air France (0.35) }
B	{British (0.10), Lufthansa (0.15), Aeroflot (0.6) }
C	{Lufthansa (0.3), Air France (0.5), Aeroflot (0.2) }

На самом деле, веса могут быть чем-то другим, чем вероятности или частоты, такие как объемы, потребности, возможности или доверие. В этих случаях их сумма не обязательно составляет 1. Для более общего обсуждения этих случаев см. [65], где показано, как теория вероятностей и теория возможностей могут быть распространены на анализ символьных данных.

Обобщение описаний с участием гистограммных или категориальных модальных переменных операторы максимума или минимума приводят к распределению, где сумма значений для каждого класса категорий не может быть равна 1. Эти обобщения были использованы в кластеризации символьных данных, например [60].

Категориальные модальные переменные похожи на гистограммные переменные для количественного случая в том, что их значения характеризуются классами категорий или весами. Далее под «распределенными данными» мы понимаем оба типа как противоположные «множественным» переменным, когда нет распределения. Тем не менее с математической точки зрения зрения они имеют разную природу.

4.6. Квантильное представление

Квантильное представление [93] обеспечивает общие рамки для представления символьных данных, описываемых переменными разных типов.

В качестве примера рассмотрим данные о прибывающих рейсах в трех аэропортах, для каждого рейса представим номера пассажиров, время задержки (в мин) и категории расстояний (от 1 для внутреннего рейса до 5 для очень дальних межконтинентальных полетов). Данные для каждого рейса были агрегированы по аэропортам, в виде символьного представления данных в табл. 6.

Таблица 6

Символьное представление данных

Аэропорт	Число пассажиров	Задержка времени прибытия (мин)	Категория расстояний
A	[150,200]	{[0, 10[, 0.25; [10, 30[, 0.65; [30, 60], 0.10}	{1 (0.40) ; 2 (0.40) ; 3 (0.2)}
B	[180,300]	{[0, 10[, 0.45; [10, 30[, 0.30; [30, 60], 0.25}	{1 (0.10) ; 2 (0.30) ; 3 (0.30); 4 (0.20); 5 (0.10)}
C	[200,400]	{[0, 10[, 0.75; [10, 30[, 0.20; [30, 60], 0.05}	{1 (0.05) ; 2 (0.10) ; 3 (0.15); 4 (0.40); 5 (0.30)}

Квантильное представление данных

Аэропорт	Число пассажиров	Задержка времени прибытия (мин)	Категория расстояний
A	(150, 162.5, 175, 187.5, 200)	(0, 10, 17.7, 25.4, 60)	(1, 1, 2, 2, 3)
B	(180, 210, 240, 270, 300)	(0, 10.6, 13.4, 30, 60)	(1, 2, 3, 4, 5)
C	(200, 250, 300, 350, 400)	(0, 3.3, 6.7, 10, 60)	(1, 3, 4, 5, 5)

Чтобы получить однородный массив данных, квантильное представление может рассматриваться для каждого наблюдения; здесь мы предполагаем равномерное распределение внутри каждого наблюдаемого интервала переменной «Число пассажиров» и в каждом подинтервале переменной «Задержка времени прибытия». Табл. 6 отображает полученное квантильное представление для трех аэропортов.

4.7. Другие типы символьных данных

Таксономические переменные

Переменная $Y \to O$ является таксономической, если O имеет древовидную структуру. Таксономии могут быть приняты во внимание в получении описаний агрегированных данных.

Математически таксономией является древообразная структура классификаций определённого набора объектов. Вверху этой структуры — объединяющая единая классификация — корневой таксон — которая относится ко всем объектам данной таксономии. Таксоны, находящиеся ниже корневого, являются более специфическими классификациями, которые относятся к поднаборам общего набора классифицируемых объектов. Термины «таксономия» и «систематика» нередко используют как синонимы, но в строгом смысле таксономия является лишь частью систематики.

Ограниченные переменные

Переменная Y' иерархически зависит от переменной Y, если ее применение ограничено значениями, принятыми для $Y : Y'$, которые нельзя применить, если Y принимает значения в данном наборе C. Другими словами, переменная Y' иерархически зависит от переменной Y, если Y' не имеет смысла для некоторых значений, которые Y может принимать, и, следовательно, становится «неприменимым». Например, если опрос содержит пункт о времени безработицы человека, переменная не относится к человеку, который никогда не был безработным. Описания, которые допускают несоблюдение правила, называются «некогерентными».

4.8. Методы анализа символьных данных

В последние годы были исследованы разные подходы и предложено много методов для анализа символьных данных. Это, однако, не произошло единообразно по типам данных и методам анализа: интервальные данные на сегодняшний день являются наиболее рассмотренным случаем. Кластерный анализ получил значительно больше внимания, чем другие многомерные методологии.

Далее мы представляем не исчерпывающее исследование различных методов анализа, относящихся к наиболее важным (или впервые предложенным) методам. В новой и динамичной области исследований, как САД, в настоящее время разрабатывается много подходов, есть различные альтернативные методы в любой конкретной области интересов.

Метрики в пространствах символьных данных

Многие многомерные методы основаны на различиях между анализируемыми объектами. Рассмотрены и исследованы несколько метрик для разных типов символьных данных, которые различаются в зависимости от типа переменных.

Когда анализируются сущности по интервальным переменным, для сравнения интервальных наблюдений могут использоваться различные меры расстояния. Наиболее распространенными являются расстояния минковского типа, которые возникают в результате вложения интервалов в R^2, где одно измерение используется для нижней, другое для верх-

ней границы интервалов, а также хаусдорфово расстояние, которое оценивает максимум расстояния от набора до ближайшей точки в другом наборе, т. е. два набора близки по хаусдорфовскому расстоянию, если каждая точка любого набора близка к некоторой точке другого набора. Mahalanobis расстояние описывается векторами интервалов.

В работе [89] представлено исследование различных мер для сравнения вероятностных распределений, например расхождение, расстояние Хеллингера, относительная энтропия (или дивергенция Кульбака–Лейблера), колмогорова метрика, метрика Леви, метрика Прохорова. Расстояние Wasserstein и его L_2, расстояние Mallows являются наиболее широко используемыми для кусочно-полиномиальных данных.

4.9. Символьная регрессия

Линейная регрессия в САД была впервые рассмотрена для случая интервальных переменных. Billard и Diday [59] предложили первую модель линейной регрессии для интервальных переменных, полученную при условии однородности в каждом наблюдаемом интервале; оценочные коэффициенты затем применяются до нижней и верхней границ независимой переменной для оценки нижней и верхней границ зависимой переменной. Neto и De Carvalho [105] предложили другую модель, оценивая среднюю точку и радиус зависимой переменной; потом в работе [106] те же авторы предложили новую модель с неотрицательными ограничениями на средних частотах коэффициентов регрессии.

Что касается гистограммных переменных, то модель [59] использует полученные значения ковариации для оценки коэффициентов регрессии.

Irpino и Verde [139] разработали простую линейную регрессионную модель для гистограммных данных, которая минимизирует расстояние Mallows между наблюдаемыми квантильными функциями зависимой переменной. Предлагаемый метод заключается в использовании свойств разложения расстояния Wasserstein, полученного Irpino и Romano; это применяется для измерения суммы квадратов ошибок.

Dias и Brito [63] предложили модель линейной регрессии для симметричного распределения. Распределения по гистограммным переменным представлены их квантильной функцией: это представление использует расстояние Mallows. Модель включает в себя как квантильные функции, которые представляют распределения, которые независимые гистограммные переменные принимают, и квантильные функции, которые

представляют собой соответствующие симметричные гистограммные переменные. Здесь на параметры накладываются ограничения неотрицательности. Модель требует решения задачи квадратичной оптимизации с учетом ограничений неотрицательности на неизвестные.

4.10. Анализ временных рядов

В первой работе, посвященной интервальным временным рядам, использован подход, основанный на подборе одномерного метода ARIMA [135]. Maia и др. предложили адаптацию одномерных процессов ARIMA для средних точек и радиусов, используя их для прогнозирования границ интервалов, а также подход, основанный на модели искусственной нейронной сети, и их комбинации. В работах [53] авторы определили интервальные случайные процессы, интервальные временные ряды, слабую стационарность для интервальных процессов и, исходя из ранее предложенных примеров, определили эмпирическую автоковариацию и автокорреляцию функции для данных временного ряда интервалов. Прогнозирование в основном основано на векторе авторегрессионной модели сглаживающих фильтров.

Teles и Brito [135] предложили моделирование временного ряд интервалов с пространственно-временной авторегрессией моделей.

Прогнозирование, когда данные описываются гистограммными переменными для финансовых приложений, представлено в [55].

Глава 5

Функции случайных переменных

При численном моделировании часто возникают задачи вычисления функциональных зависимостей. Эти зависимости необходимо исследовать.

Анализ этих зависимостей осложняется в условиях неопределенности. В тех случаях, когда относительно входных переменных известны их функции плотности вероятности, возникает задача оценки законов распределения подобных функций. Поскольку для большинства функций сложно построить аналитические выражения законов распределения, возникает задача численного построения аппроксимаций их функций плотности вероятности. Численные процедуры вычисления аппроксимаций законов распределения функций случайных аргументов мы далее будем называть вероятностными расширениями.

Мы начнем главу с рассмотрения алгебраических свойств случайных переменных, включая арифметические операции, операции сравнения. Во втором параграфе рассмотрены свойства вероятностных расширений. Далее приведены примеры вычисления вероятностных расширений различных типов функций, включая решения краевых задач со случайными коэффициентами, задач интерполяции и построения надежных оценок эмпирических функций распределения.

5.1. Алгебра случайных переменных

Пусть x — случайная величина, тогда ее плотность вероятности будем обозначать жирным шрифтом $\boldsymbol{x}$.

Обозначим через $\boldsymbol{R}$ — множество плотностей вероятности $\{\boldsymbol{x}\}$ случайных величин x, соответственно $\boldsymbol{R}^n$ — пространство плотностей вероятности случайных векторов из R^n.

Носителем функции плотности вероятности $\boldsymbol{f}$ будем называть множество

$$\mathrm{supp}(\boldsymbol{f}) = \{x | \boldsymbol{f}(x) > 0\}.$$

Вероятностная арифметика

Известны аналитические формулы для определения плотности вероятности результатов арифметических действий над случайными величинами. Например, для нахождения плотности вероятности $p_{x_1+x_2}$ суммы двух случайных величин $x_1 + x_2$ используется соотношение [9]

$$p_{x_1+x_2}(x) = \int_{-\infty}^{\infty} p(x-v, v)dv = \int_{-\infty}^{\infty} p(v, x-v)dv. \tag{5.1}$$

Для нахождения плотности вероятности p_{x_1/x_2} частного двух случайных величин x_1/x_2

$$p_{x_1/x_2} = \int_{0}^{\infty} vp(xv, v)dv - \int_{-\infty}^{0} vp(v, xv)dv. \tag{5.2}$$

Плотность вероятности $p_{x_1x_2}$ произведения двух случайных величин x_1x_2 [6] определяется соотношением

$$p_{x_1x_2} = \int_{0}^{\infty} (1/v)p(x/v, v)dv - \int_{-\infty}^{0} (1/v)p(v, x/v)dv. \tag{5.3}$$

Рассмотрим вопрос существования обратных элементов по сложению и умножению. Заметим, что для уравнения $\boldsymbol{a} + \boldsymbol{x} = 0$ существует решение, которое можно представить в виде совместной функции распределения (a, x):

$$p_{-a}(x_1, x_2) = \begin{cases} p_a(x_1) & \text{если } x_1 = -x_2; \\ 0 & \text{если } x_1 \neq -x_2, \end{cases}$$

где $p_a(x)$ — плотность вероятности случайной величины a. Таким образом $p_{-a}(x_1, x_2)$ — определяет обратный элемент по сложению для $\boldsymbol{a}$. Аналогично несложно построить и обратный элемент по умножению.

Рассмотрим операцию $\max(\boldsymbol{x}, \boldsymbol{y})$. Вероятность того, что $\max(x, y) < z$ — определяется через функцию распределения F

$$F(z) = \int_{-\infty}^{z} \boldsymbol{x}(\xi)d\xi \int_{-\infty}^{z} \boldsymbol{y}(\xi)d\xi.$$

Используя функцию распределения F, можно построить функцию плотности вероятности для $\max(\boldsymbol{x}, \boldsymbol{y})$. Тогда

$$\max(\boldsymbol{x}, \boldsymbol{y})(z) = \frac{dF(\xi)}{d\xi} = \boldsymbol{x}(z) \int_{-\infty}^{z} \boldsymbol{y}(\xi) d\xi + \boldsymbol{y}(z) \int_{-\infty}^{z} \boldsymbol{x}(\xi) d\xi.$$

Аналогично можно получить выражение для $\min(\boldsymbol{x}, \boldsymbol{y})$:

$$\min(\boldsymbol{x}, \boldsymbol{y})(z) = \boldsymbol{x}(z) \left(1 - \int_{-\infty}^{z} \boldsymbol{y}(\xi) d\xi\right) + \boldsymbol{y}(z) \left(1 - \int_{-\infty}^{z} \boldsymbol{x}(\xi) d\xi\right).$$

Расширим отношения порядка $* \in \{<, \leq, \geq, >\}$ на случайные переменные:

$$\boldsymbol{x} * \boldsymbol{y} \Leftrightarrow x * y \text{ для всех } x \in supp(\boldsymbol{x}), y \in supp(\boldsymbol{y}).$$

Если носители $\boldsymbol{x}$, $\boldsymbol{y}$ пересекаются, тогда мы можем говорить о вероятности $\boldsymbol{x} * \boldsymbol{y}$

$$P(\boldsymbol{x} * \boldsymbol{y}) = \int_{\Omega} p(x, y) dx dy,$$

где $\Omega = \{(x, y) | x * y\}$, $p(x, y)$ — совместная функция плотности вероятности $\boldsymbol{x}$ и $\boldsymbol{y}$.

5.2. Вероятностные расширения

Рассмотрим задачу определения закона распределения функции нескольких случайных аргументов.

Пусть $(x_1, x_2, \ldots, x_n)$ — система случайных непрерывных переменных с совместной функцией плотности вероятности $\boldsymbol{p}(x_1, x_2, \ldots, x_n)$. Случайная переменная z

$$z = f(x_1, x_2, \ldots, x_n),$$

где функция $f : R^n \to R$.

Определение 21. *Будем говорить, что случайная функция $\boldsymbol{f} : \boldsymbol{R}^n \to \boldsymbol{R}$ является вероятностным продолжением вещественной функции $f : R^n \to R$ на множестве $D \subset R^n$, если $\boldsymbol{f}(x) = f(x)$ для всех точечных аргументов $x \in D$.*

Определение 22. *Случайная функция $\boldsymbol{f} : \boldsymbol{R}^n \to \boldsymbol{R}$ называется вероятностным расширением вещественной функции $f : R^n \to R$ на множестве $D \subset R^n$, если она*

(i) является вероятностным продолжением f на D,

(ii) функция плотности вероятности $\boldsymbol{f}$ совпадает с функцией плотности вероятности $\boldsymbol{z}$ случайной величины z:

$$z = f(x_1, x_2, \ldots, x_n).$$

Таким образом, мы можем записать

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n).$$

В тех случаях, когда надо указать непосредственно значение плотности вероятности $\boldsymbol{f}$ в некоторой точке ξ, будем использовать обозначение

$$\boldsymbol{z}(\xi) = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)(\xi).$$

Пусть имеется система непрерывных случайных величин $(x_1, x_2, \ldots, x_n)$ с совместной плотностью распределения $p(x_1, x_2, \ldots, x_n)$ и

$$y = f(x_1, x_2, \ldots, x_n).$$

Тогда функция распределения F_y для случайной величины y [6, 9]

$$F_y(\xi) = P(y < \xi) = \int_{\Omega_z} p(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n, \qquad (5.4)$$

где $\Omega_\xi = \{(x_1, x_2, \ldots, x_n) | f(x_1, x_2, \ldots, x_n) < \xi\}$. В тех случаях, когда это возможно, дифференцируя F_y, можно получить плотность вероятности для $\boldsymbol{y}$

$$\frac{dF_y(\xi)}{dz} = f_y(\xi).$$

Плотность вероятности случайной величины $y = ax + b$ для вещественных $a \neq 0$, b будет иметь вид [110]

$$\boldsymbol{y}(\xi) = \frac{1}{|a|}\boldsymbol{x}(\frac{\xi - b}{a}). \qquad (5.5)$$

Таким образом для линейной функции $y = ax + b$ построено вероятностное расширение $a\boldsymbol{x} + b$ (5.5). Если необходимо указать конкретное значение плотности вероятности в точке ξ, можно использовать запись $(a\boldsymbol{x} + b)(\xi)$.

Пусть $f(x_1, ..., x_n)$ — рациональная функция, тогда для вычисления $\boldsymbol{f}$ заменим арифметические операции на вероятностные операции, а переменные x_1, x_2, ... , x_n — значениями их функций плотности вероятности. Полученную плотность вероятности $\boldsymbol{f}$ будем называть *естественным вероятностным расширением.*

Теорема 8. *Пусть $f(x_1, \ldots, x_n)$ — рациональная функция, каждая переменная которой встречается только один раз и $x_1, \ldots, x_n$ — независимые случайные величины. Тогда естественное вероятностное расширение совпадает с вероятностным расширением.*

Доказательство проведем по индукции. Для $n = 2$ утверждение справедливо. Пусть справедливо для $n = k$ и $\boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ — вероятностное расширение функции $f(x_1, \ldots, x_k)$. Покажем, что это справедливо и для $n = k + 1$. Действительно, $f(x_1, \ldots, x_k, x_{k+1}) = f(x_1, \ldots, x_k) * x_{k+1}$, где $* \in \{+, -, \cdot, /\}$, но $\boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}) = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) * \boldsymbol{x}_{k+1}$. Теорема доказана.

Для рациональной функции $x(y+z) = xy+xz$ только первое представление попадает под условие теоремы 8 и, следовательно, естественное вероятностное расширение будет совпадать с вероятностном расширением. Таким образом, относительно численных арифметических операций дистрибутивность не выполняется.

Теорема 8 легко обобщается на следующий случай.

Следствие 2 ([16]). *Пусть для функции $f(x_1, ..., x_n)$ возможна замена переменных, такая что $f(z_1, \ldots, z_k)$ — рациональная функция от переменных $z_1, \ldots, z_k$, удовлетворяющая условиям теоремы 8 и z_i — функции от множества переменных $x_i, i \in Ind_i$, причем множества Ind_i попарно не пересекаются. Пусть для каждой z_i существуют вероятностные расширения. Тогда естественное вероятностное расширение $f(z_1, \ldots, z_k)$ будет совпадать с вероятностным расширением.*

Пусть $f(x_1, x_2) = (-x_1^2 + x_1)\sin(x_2)$. Тогда, полагая $z_1 = (-x_1^2 + 1)$ и $z_2 = sin(x_2)$. Заметим, что можно построить вероятностные расширения функций z_1, z_2 и $f = z_1 z_2$ — рациональная функция, попадающая под условия теоремы 8. Следовательно, ее естественное расширение будет совпадать с вероятностным расширением.

Для частного вида случайных функций введем следующие понятия. Пусть случайная функция имеет вид

$$\boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i(x), \qquad g_i \in C^m[a, b].$$

Тогда формальную производную от $\boldsymbol{f}(x)$ определим таким образом:

$$\partial^k \boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i^{(k)}(x), \quad k = 0, ..., m.$$

Соответственно функцию

$$f(x) = \sum_{i=1}^{n} a_i g_i(x), \qquad a_i \in \mathrm{supp}(\boldsymbol{a}_i)$$

будем называть *сужением функции* $\boldsymbol{f}$ по константам $\boldsymbol{a}_i$. Далее значение $a \in \boldsymbol{a}$ будем понимать как некоторую реализацию случайной величины a.

Имеется непрерывная случайная величина x с плотностью распределения $\boldsymbol{x}$. Другая случайная величина y связана с нею функциональной зависимостью:

$$y = f(x).$$

Функция плотности вероятности $\boldsymbol{y}$ в случае монотонной функции f согласно [6] определяется следующим образом:

$$\boldsymbol{y}(\xi) = f(f^{-1}(\xi))|(f^{-1})'(\xi)|,$$

где f^{-1} — функция обратная к f. Однако пользоваться таким представлением в ряде случаев довольно затруднительно.

Одним из возможных способов оценки плотности вероятности $\boldsymbol{z}$ of случайной величины z

$$z = f(x_1, \ldots, x_n) \tag{5.6}$$

есть метод Монте-Карло [27]. Рассмотрим случай независимых случайных величин $(x_1, \ldots, x_n)$ с плотностями вероятности $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Далее генерируем случайные вектора $(x_1^i, \ldots, x_n^i)$ и вычислим $z^i = f(x_1^i, \ldots, x_n^i)$, $i = 1, ..., N$.

Для изучения свойств $\boldsymbol{z}$ построим сетку

$$\{z_0, z_1, ...z_k\} \in \mathrm{supp}(\boldsymbol{z})$$

и вычислим число точек n_l в каждом отрезке z_{l-1}, z_l:

$$p_l = \frac{n_l}{N} \approx \int_{z_{l-1}}^{z_l} \boldsymbol{z}(\xi) d\xi.$$

Построим сетку $\{t_0, t_1, ...t_m\} \subset \mathrm{supp}(\boldsymbol{x}_1)$ и вычислим число попаданий m_j в каждый отрезок t_{j-1}, t_j:

$$p_j^1 = \frac{m_j}{N}.$$

Рассмотрим только те случайные вектора $(x_1^i, \ldots, x_n^i)$, для которых $(x_1^i \in (t_{j-1}, t_j)$. Число таких векторов в точности равно m_j. Если ν_{jl} — число попаданий в каждый отрезок z_{l-1}, z_l, тогда

$$n_l = \sum_j^m \nu_{jl}$$

и

$$p_l = \frac{n_l}{N} = \sum_j^m \frac{m_j}{N} \frac{\nu_{jl}}{m_j}.$$

Переходя к пределу при $N \to \infty$ получаем

$$\boldsymbol{x}_1(t) \approx \frac{m_j}{N}.$$

Далее, пусть для всех $t \in \text{supp}\boldsymbol{x}_1$ мы можем построить вероятностное расширение $\boldsymbol{f}(t, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$

$$\frac{\nu_{jl}}{m_j} \approx \boldsymbol{f}(t, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)(\xi).$$

Увеличивая размерность сеток и переходя к пределу, получаем

$$\boldsymbol{f}(\xi) = \int \boldsymbol{x}_1(t) \boldsymbol{f}(t, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)(\xi) dt. \tag{5.7}$$

Таким образом доказана теорема.

Теорема 9 ([76]). *Пусть $(x_1, \ldots, x_n)$ — независимые случайные величины и $\boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ — вероятностное расширение $f(x_1, x_2, \ldots, x_n)$, и для всех вещественных t функция $\boldsymbol{f}(t, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ — вероятностное расширение $f(t, x_2, \ldots, x_n)$. Тогда*

$$\boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)(\xi) = \int_{\text{supp}(\boldsymbol{x}_1)} \boldsymbol{x}_1(t) \boldsymbol{f}(t, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)(\xi) dt. \tag{5.8}$$

Замечание 3. *Из теоремы 9 вытекает возможность рекурсивного вычисления вероятностных расширений общего вида сведением их к вычислению одномерных вероятностных расширений.*

Рассмотрим вычисление интеграла (5.8). Для простоты представим (5.8) в виде квадратуры

$$\int \boldsymbol{x}_1(t) \boldsymbol{f}(t, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) dt \approx \sum_{l=1}^m \gamma_l \boldsymbol{x}_1(t_l) \boldsymbol{f}(t_l, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n).$$

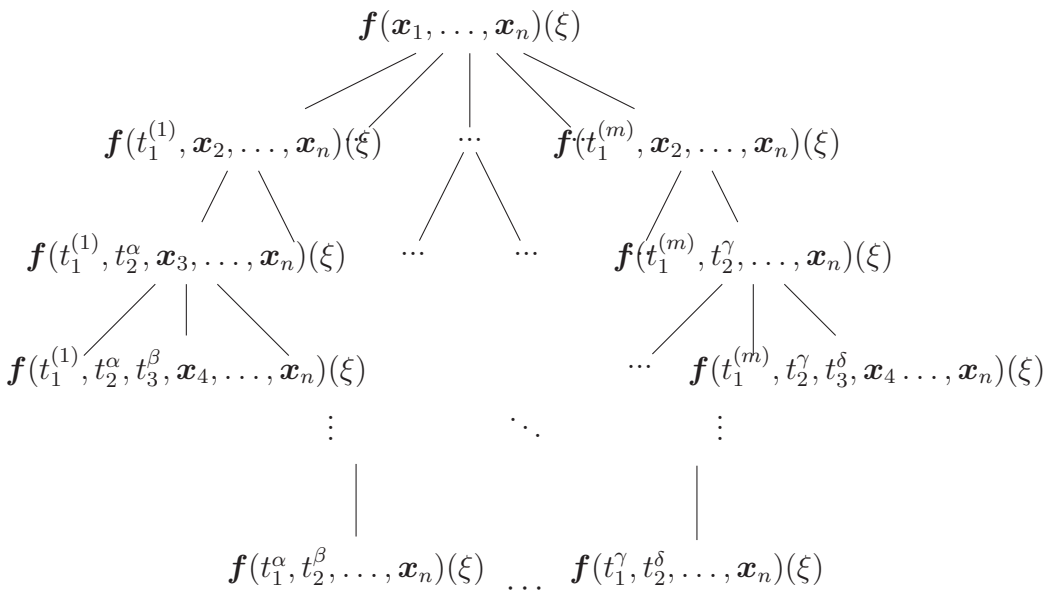


Рис. 5.1. Дерево параллельно-рекурсивной организации вычислительного процесса

Далее, для вычисления $\boldsymbol{f}(z,t_l,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n)$ мы также будем использовать числовые квадратуры и т. д. В общем, это NP-трудная проблема, и необходимо распараллеливание.

На рис. 5.1 показано дерево параллельно-рекурсивной организации вычислительного процесса. Таким образом, на нижнем уровне необходимо вычислять вероятностные расширения для функций только одной переменной. Все расчеты на каждом слое независимы и могут быть рассчитаны одновременно.

5.3. Одномерный случай

Рассмотрим процедуру вычисления вероятностных расширений в одномерном случае, не требуя монотонности функции f.

Пусть дана функциональная зависимость

$$y = f(x),$$

где x — случайная величина; $\boldsymbol{x}$ — функция плотности вероятности случайной величины x с носителем $[\underline{x},\overline{x}]$. Далее $\{x_i(y) \in [\underline{x},\overline{x}] | i = 1,\ldots,n\}$ — корни уравнения $y = f(x)$.

Пусть необходимо найти функцию плотности вероятности $\boldsymbol{y}$ случайной величины y или вероятностное расширение $\boldsymbol{f}(\boldsymbol{x})$ функции $f(x)$. Мы

можем представить решение в виде

$$\boldsymbol{f}(\boldsymbol{x})(\xi) = \sum_{i=1}^{n} \frac{\boldsymbol{x}(x_i(\xi))}{|f'(x_i(\xi))|}.$$

Действительно, вероятность попадания случайной величины в отрезок $[y, y + dy]$:

$$\boldsymbol{f}(\boldsymbol{x})(\xi)dy = \mathsf{P}(x \in [\xi, \xi + dy]) = \sum_{i=1}^{n} \boldsymbol{x}(x_i(\xi))|dx_i|,$$

выразив $|dx_i| = dy/|f'(x_i(\xi))|$ и поделив на dy, получаем доказательство.

Пример 3. Рассмотрим построение вероятностного расширения

$$f = ax^2 + bx, a \geq 0, b \geq 0,$$

где x — случайная переменная распределенная на $[0, 2]$ по треугольному закону

$$\boldsymbol{x}(\xi) = \begin{cases} \xi & \text{if } \xi \in [0, 1), \\ 2 - \xi & \text{if } \xi \in [1, 2]. \end{cases}$$

Далее

$$r_1(z) = -\frac{\sqrt{4az + b^2} + b}{2a}, r_2(z) = \frac{\sqrt{4az + b^2} - b}{2a}.$$

— корни уравнения $z = f(r_i)$. Выберем положительный корень $r(\xi) = r_2(\xi) = \frac{\sqrt{4a\xi+b^2}-b}{2a}$, и

$$[ax^2 + bx]' = 2ax + b = \sqrt{4az + b^2}.$$

Окончательно

$$\boldsymbol{f}(\boldsymbol{x})(\xi) = \boldsymbol{x}(r(\xi))/\sqrt{4a\xi + b^2}.$$

Положим $a = 1$ и $b = 0$, получаем

$$\boldsymbol{f}(\boldsymbol{x})(\xi) = \begin{cases} 1/2 & \text{if } \xi \in [0, 1), \\ 1/\sqrt{\xi} - 1/2 & \text{if } \xi \in [1, 4]. \end{cases}$$

Численный подход

Рассмотрим численный подход для построения вероятностного расширения $\boldsymbol{f}$ функции $f(x)$. Для этих целей построим на носителе $[\underline{x}, \overline{x}]$

функции плотности вероятности $\boldsymbol{x}$ сетку $\{\xi_i | \xi_i \in [\underline{x}, \overline{x}], i = 0, 1, 2...n\}$ и вычислим $\{z_i = f(\xi_i), i = 0, 1, 2...n\}$. Далее положим

$$f_z(z_i) = \frac{\boldsymbol{x}(\xi_i)}{|(f'(\xi_i)|}$$

и, используя $(z_i, f_z(z_i))$, построим кусочно-полиномиальную интерполяцию.

Так, для примера 3 носитель $[0, 2]$, $\xi_i = i/10, i = 0, 1, \ldots, 20$. С $a = 1, b = 0$ получаем

$$(z_i, f_z(z_i)) = \left((i/10)^2, \frac{\boldsymbol{x}(i/10)}{i/5}\right), i = 0, 1, \ldots, 20.$$

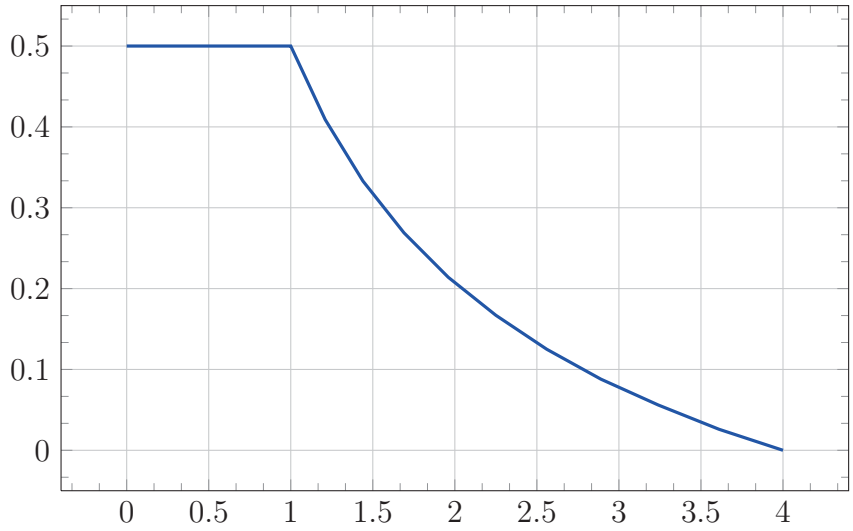


Рис. 5.2. Вероятностное расширение $\boldsymbol{f}(\boldsymbol{x})$

На рис. 5.2 показана функция плотности вероятности $\boldsymbol{f}(\boldsymbol{x})$ из примера 3. В отличие от главного подхода, в данном случае нет необходимости находить корни уравнений.

5.4. Случай двух переменных

Пусть $p(x_1, x_2)$ — совместная плотность распределения системы двух случайных величин (x_1, x_2) и случайная переменная z есть функция $f(x, y)$:

$$z = f(x, y).$$

Необходимо найти функцию плотности вероятности случайной величины z. Определим

$$\Omega_z = \{(x, y) | z > f(x, y)\};$$

и F_z — функция распределения случайной величины z

$$F_z = \int_{\Omega_z} p(x,y)dxdy;$$

функция плотности вероятности z

$$f_z = \frac{d}{dz}F_z.$$

Следовательно,

$$\frac{d}{dz}F_z = \frac{d}{dz}\int_{\Omega_z} p(x,y)dxdy =$$

$$= \lim_{dz\to 0} \frac{\int_{\Omega_{z+dz}} p(x,y)dxdy - \int_{\Omega_z} p(x,y)dxdy}{dz} =$$

$$= \lim_{dz\to 0} \frac{\int_{\Omega_{z+dz}\setminus\Omega_z} p(x,y)dxdy}{dz}.$$

Положим $dx = 0$, тогда

$$dz = f'_y dy$$

и

$$dS = dxdy = dxdz/|f'_y|.$$

Окончательно получаем

$$f_z(z) = \int_{\Gamma_z} \frac{p(x,y)}{|f'_y(x,y)|}dx,$$

где

$$\Gamma_z = \{(x,y)|z = f(x,y)\} = \{(x,y(x))\}.$$

В случае независимых случайных переменных x, y совместную плотность $p(x,y)$ можно представить в виде произведения $p(x,y) = \boldsymbol{xy}$. Тогда

$$f_z(z) = \int_{\Gamma_z} \frac{p(x,y)}{|f'_y(x,y)|}dx = \int_{\mathrm{supp}(\boldsymbol{x})} \frac{\boldsymbol{x}(t)\boldsymbol{y}(\xi)}{|f'_y(t,\xi)|}dt, \tag{5.9}$$

где $\xi(t,z)$ — корень уравнения $z = f(t,\xi(t,z))$. Заметим, что

$$\frac{\boldsymbol{y}(\xi)}{|f'_y(t,\xi)|} = \boldsymbol{f}(t,\boldsymbol{y})(z)$$

— вероятностное расширение функции $f(t,y)$. Таким образом, мы можем переписать

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})(z) = \int_{\mathrm{supp}(\boldsymbol{x})} \boldsymbol{x}(t)\boldsymbol{f}(t,\boldsymbol{y})(z)dt.$$

Рассмотрим вопрос численного вычисления $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})(z)$. Для этих целей в области носителей $\boldsymbol{x}$, $\boldsymbol{y}$ $[\underline{x},\overline{x}]$, $[\underline{y},\overline{y}]$ построим сетки $\omega_x = \{x_0 = \underline{x} < x_1 < \ldots < x_K = \overline{x}\}$, and $\omega_y = \{y_0 = \underline{y} < y_1 < \ldots < y_L = \overline{y}\}$.

Вычислим значения $f_{kl} = f(x_k, y_l)$. Далее построим эрмитовы кубические сплайны $s_l(x), l = 0, 1, \ldots, L$, используя значения f_{kl}. На рис. 5.3 они показаны сплошной линией.

Далее, для некоторого значения ξ мы находим корни ξ_l:

$$s_l(\xi_l) = \xi, l = 0, 1, \ldots, L.$$

Значение $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})(\xi)$ вычисляем, используя численные квадратуры, например Симпсона:

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})(\xi) = h\sum_{l=0}^{L}\gamma_k\boldsymbol{y}(y_l)\boldsymbol{x}(\xi_l)/s'(\xi_l). \tag{5.10}$$

Рассмотрим вычисление $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})(\xi)$ в случае зависимых переменных, когда нам известна совместная функция плотности $p(x, y)$. Поскольку в формуле (5.10) выражение $\boldsymbol{y}(y_l)\boldsymbol{x}(\xi_l)$ определяет значение функции плотности вектора (x, y) в точке (ξ_l, y_l), мы можем заменить на $p(\xi_l, y_l)$ согласно (5.9)

$$\boldsymbol{f}(z) = \int \frac{p(t,\xi(z))}{|s'_y(t,\xi(z))|}dx.$$

Замечание 4. *Основные вычислительные затраты — число вычислений значений функции в узлах сетки $f_{kl} = f(x_k, y_l)$, равное KL. Между тем, построив массив f_{kl}, мы можем относительно быстро вычислить $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})(\xi)$ при различных значениях $\boldsymbol{x},\boldsymbol{y}$.*

В рамках вычислительного вероятностного анализа рассматриваются численные арифметические операции над различными типами плотностей случайных величин. Это позволяет в некоторых случаях вычислять интегралы вида (5.4) с необходимой точностью.

В вычислительном вероятностном анализе используются различные типы представления плотностей случайных величин: дискретные, гистограммы, полиграммы, полигоны, кусочно-полиномиальное и аналитическое представление. Далее рассмотрим изложение на примере кусочно-полиномиальных функций (сплайнов).

В случае, когда случайные величины x, y являются независимыми и имеют плотности вероятности, представленные кубическими сплайнами, s_x s_y, совместную плотность вероятности можно представить в виде

произведения $p(x,y) = s_x s_y$. Поскольку кубический сплайн на каждом отрезке сетки представляет кубический полином, то $p(x,y)$ в случае вычисления интегралов (5.1)–(5.3) будет кусочно-полиномиальной функцией шестой степени. Наиболее удобными в этом случае будут квадратуры Гаусса с четырьмя внутренними узлами, которые точны на полиномах седьмой степени.

В качестве примера рассмотрим построение сплайна, аппроксимирующего $\boldsymbol{x}_1 + \boldsymbol{x}_2$. Для этих целей в области носителя $\boldsymbol{x}_1 + \boldsymbol{x}_2$ построим сетку $\omega = \{x_0, x_1, ..., x_n\}$ и вычислим численно значения $f_i = \boldsymbol{x}_1 + \boldsymbol{x}_2(x_i)$. Используя значения f_i и краевые условия $s'(x_0) = s_0$, $s'(x_n) = s_n$, на сетке ω построим кубический сплайн s. В этом случае справедлива оценка

$$||(\boldsymbol{x}_1 + \boldsymbol{x}_2)^\nu - s^{(\nu)}|| \le K h^{4-\nu} ||(\boldsymbol{x}_1 + \boldsymbol{x}_2)^{(4)}||, \ \nu = 1, 2, 3.$$

Далее вычислим

$$\text{norm} = \int s(x) dx,$$

если norm$\neq 1$, то $s(x) := s(x)/\text{norm}$.

В случае, когда случайные величины x, y являются зависимыми, совместную функцию плотности вероятности необходимо вычислять отдельной процедурой.

Пример 4. Рассмотрим построение вероятностного расширения для функции

$$z = x^2 y + x y^2.$$

Пусть t вещественное, положим $y = t$ и найдем вероятностное расширение для функции

$$z_x(x,t) = x^2 t + x t^2.$$

Заметим, что z_x — функция одной случайной переменной, и вероятностное расширение

$$\boldsymbol{z}_x(\boldsymbol{x}, t)(\xi) = \sum_i \boldsymbol{x}(\eta_i(\xi)) / \sqrt{4t\xi + t^4},$$

где $\eta_1(\xi) = (\sqrt{4t\xi + t^4} - t^2)/(2t), \eta_1(\xi) = -(\sqrt{4t\xi + t^4} + t^2)/(2t)$ — корни квадратного уравнения $x^2 t + x t^2 = \xi$.

Таким образом, вероятностное расширение для $\boldsymbol{z}$ можно представить в виде

$$\boldsymbol{z}(\xi) = \int \boldsymbol{y}(t) \boldsymbol{z}_x(\boldsymbol{x}, t)(\xi) dt =$$

$$= \sum_i \int_{\mathrm{supp}(\boldsymbol{y})} \boldsymbol{y}(t)\boldsymbol{x}(\eta_i(\xi))/\sqrt{4t\xi + t^4}dt. \qquad (5.11)$$

5.5. Многомерный случай

В качестве примера рассматривается задача восстановления функции плотности вероятности суммы n равномерно распределенных на [0,1] случайных величин. Как известно, суммы равномерно распределенных на [0,1] случайных величин имеют функцию плотности вероятности распределения Ирвина–Холла.

Плотность вероятности равномерно распределенной на [0,1] случайной величины можно представить в виде

$$p = \begin{cases} 1 & x \in [0,1] \\ 0 & x \notin [0,1] \end{cases}.$$

Заметим, что операцию сложения случайной величины с некоторой вещественной переменной можно реализовать следующим способом. Пусть $\boldsymbol{x}$ — плотность вероятности случайной величины x, тогда $\boldsymbol{x}_r$ — плотность вероятности случайной величины $x + r$:

$$\boldsymbol{x}_r(x) = \boldsymbol{x}(x - r).$$

Последнее равенство означает, что носитель случайной величины $x + r$ сдвинут на r.

Если x — равномерно распределенная на [0,1] случайная величина, то плотность вероятности случайной величины $x + r$

$$p_r = \begin{cases} 1 & x \in [0 + r, 1 + r] \\ 0 & x \notin [0 + r, 1 + r] \end{cases}.$$

Рассмотрим программную реализацию вычисления плотности вероятности s_m суммы m равномерно распределенных на [0,1] случайных величин. Согласно теореме 9 вычисление плотности вероятности $s_m(\xi)$ можно свести к вычислению повторного интеграла. Для вычисления интегралов будем использовать численные квадратуры. В данном случае — метод прямоугольников.

Реализация численных квадратур выражается в виде вложенных m циклов. На нижнем слое вычисляется значение $f_t(\xi)$, $t = t_1 + t_2 + \ldots + t_m$.

Приведем оценки вычисления $s_4(2.0)$. Точное значение $s_4(2.0) = 2/3$, при $h = 0.1$ значение вероятностного расширения равно 0.6700, при $h =$

0.05 — 0.66750. Таким образом ошибка в первом случае составила $\varepsilon_1 = 0.0034$, во втором — $\varepsilon_2 = 0.00083$. Оценим скорость сходимости α

$$\epsilon(h) \approx Ch^{\alpha}.$$

$$\ln(\epsilon(h_i)) \approx \ln(C) + \alpha \ln(h_i), i = 1, 2,$$

$$\alpha = 2.00434.$$

Оценим число операций, необходимое для достижения точности ε при сложении m равномерно распределенных на [0,1] случайных величин.

Точность метода Монте-Карло можно оценить как $\sim \frac{C_1}{\sqrt{N}}$, где N — число реализаций, C_1 — константа. Таким образом,

$$N \sim \varepsilon^{-2},$$

число обращений к вычислению равномерно распределенных на [0,1] случайных величин

$$K_{MC} \sim m\varepsilon^{-2}.$$

Вероятностные расширения. В этом случае точность $\varepsilon \sim h^{\alpha}$, где h — шаг сетки при использовании квадратур, α — порядок сходимости квадратурной формулы. Таким образом, число $n \sim 1/h$ — число узлов в сетке. Число обращений к вычислению равномерно распределенных на [0,1] случайных величин K_{PE}

$$K_{PE} \sim \varepsilon^{-(m-1)/\alpha}.$$

Сравнивая порядки величин K_{MC} и K_{PE}, можно видеть, что при $\alpha = 2$, при числе слагаемых $m \leq 5$ ВВА эффективней Монте-Карло. При $\alpha = 4$ вероятностные расширения эффективней Монте-Карло при $m \leq 9$. Так, при $m = 4$, $h = 0.1$, $\alpha = 2$ точность $\varepsilon = 0.0034$, число обращений к процедуре вычисления равномерно распределенных на [0,1] случайных величин равно 10^3. Численный эксперимент существенно показал, что для получения подобной точности число реализаций метода Монте-Карло должно быть не меньше 10^6. В этом примере ВВА эффективней метода Монте-Карло примерно в тысячу раз.

Реализация вычисления функции в случае зависимых случайных переменных не отличается от рассмотренного выше случая двух переменных. Как и в случае двух переменных в области носителей случайных величин x_i, строим сетки ω_i. Согласно замечанию 3 вычисление функции плотности вероятности сведется к вычислению интеграла

$$\boldsymbol{f}(z) = \int \frac{p(t_1, t_2, \ldots, t_{n-1}, \xi(z))}{|s'(\xi(z))|} dt_1 dt_2 \ldots dt_{n-1},$$

где s — сплайн, построенный на сетке ω_n, $\xi(z)$ — корень уравнения $z = f(t_1, t_2, \ldots, t_{n-1}, \xi(z))$.

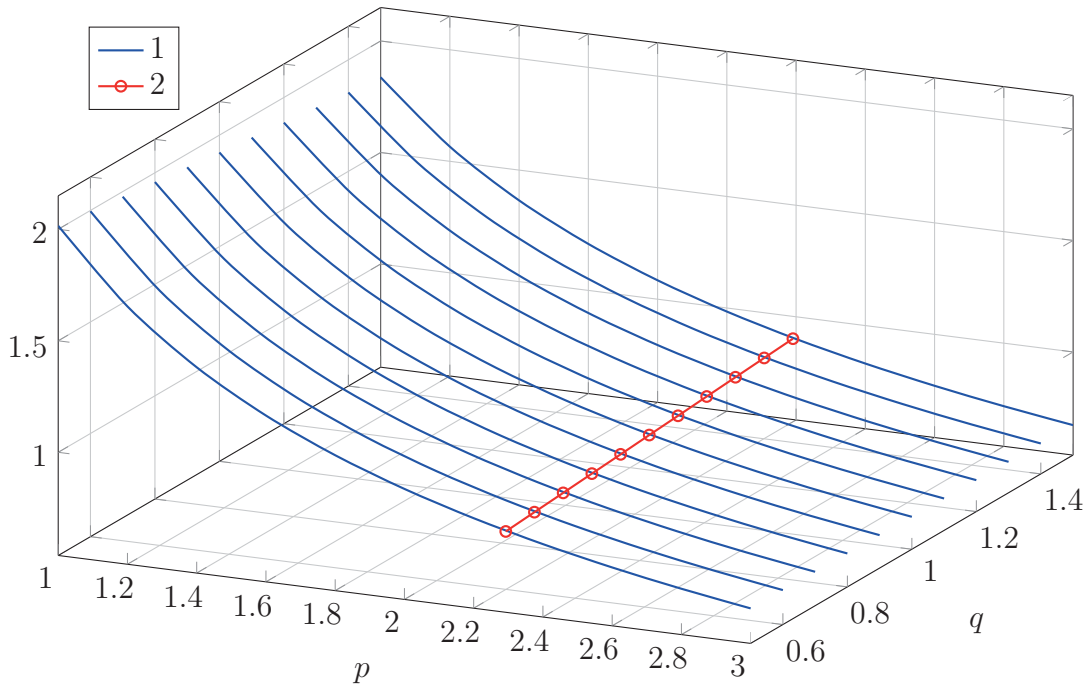


Рис. 5.3. Построение функции плотности вероятности $\boldsymbol{u}_i$, линии 1 — сплайны, 2 — линия интегрирования

5.6. Краевые задачи со случайными коэффициентами

В параграфе рассматривается использование вычислительного вероятностного анализа решения краевых задач для обыкновенных дифференциальных уравнений со случайными коэффициентами [130].

Рассмотрим краевую задачу

$$Lu \equiv -\boldsymbol{p}u'' + \boldsymbol{q}u = f(x), x \in (0,1), \tag{5.12}$$

с граничными условиями

$$u(0) = 0,\ u(1) = 0,$$

где $\boldsymbol{p} > 0$, $\boldsymbol{q} \geq 0$, $\boldsymbol{p}, \boldsymbol{q}$ — независимые случайные константы.

Пусть $\omega_h = \{x_i = ih, i = 1, 2, ..., N-1, h = 1/N\}$ — сетка и

$$L^h u_i^h = -p\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + qu_i = f(x_i), i = 1, 2, ..., N-1$$

— разностная схема. Далее $[\underline{p}, \overline{p}]$, $[\underline{q}, \overline{q}]$ — носители $\boldsymbol{p}$, $\boldsymbol{q}$. Построим сетки $\omega_p = \{p_0 = \underline{p} < p_1 < \ldots < p_K = \overline{p}\}$, и $\omega_q = \{q_0 = \underline{q} < q_1 < \ldots < q_L = \overline{q}\}$.

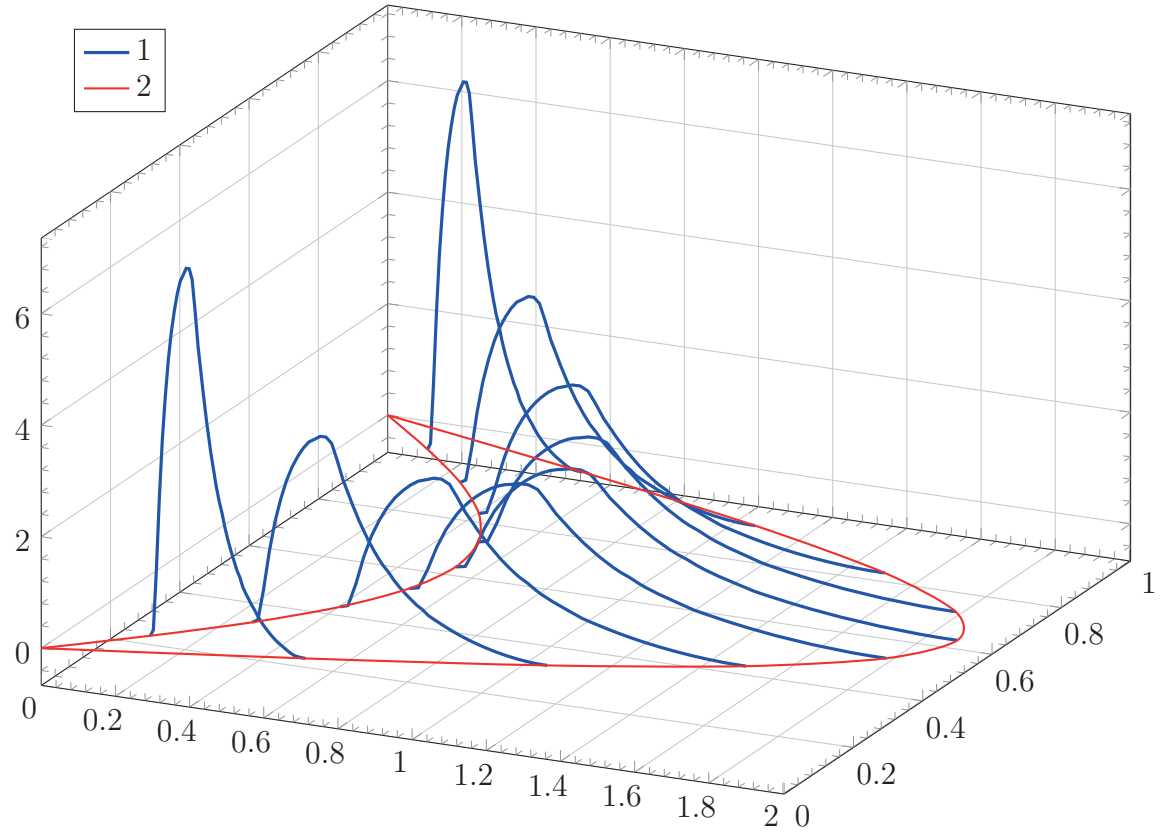


Рис. 5.4. Функции плотности вероятности решений краевой задачи

Решим численно KL задач

$$-p_k \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + q_l u_i = f(x_i), i = 1, 2, ..., N-1.$$

Таким образом, мы получаем массив решений $u_{ikl} = u_i(p_k, q_l)$. Рассмотрим построение функции плотности вероятности решения $\boldsymbol{u}_i$. Для этих целей, используя значения $u_i(p_k, q_l)$, мы построим эрмитовые кубические сплайны $s_l(p), l = 0, 1, \ldots, 10$. На рис. 5.3 линии 1 — эрмитовые кубические сплайны s_l. Для того, чтобы вычислить значение функции плотности вероятности решения $\boldsymbol{u}_i$ в некоторой точке ξ, найдем корни p_l

$$s_l(p_l) = \xi, l = 0, 1, \ldots, 10.$$

Значения $\boldsymbol{u}_i(\xi)$ вычислим, используя численные квадратуры, например квадратуры Симпсона

$$\boldsymbol{u}_i(\xi) = h \sum_{k=0}^{10} \gamma_k \boldsymbol{q}(q_k) \boldsymbol{p}(p_k) / s'(p_k).$$

На рис. 5.3, 2 — линия интегрирования (помечена «о»). Заметим, квадратура Симпсона и кубические сплайны имеют точность $O(h^4)$.

Таким образом, вычисляя значения функции плотности вероятности $\boldsymbol{u}_i(\xi)$ при различных значениях ξ, можно построить кусочно-полиномиальную аппроксимацию $\boldsymbol{u}_i$.

На рис. 5.4 представлено решение краевой задачи со случайными коэффициентами, линии 1 — функции плотности вероятности решений краевой задачи в узлах сетки разностной схемы, линии 2 — границы носителя решения краевой задачи.

Численные эксперименты при $K, L = 10$ показали хорошее согласование с методом Монте-Карло с числом бросаний 10^6. В этом случае ВВА эффективней метода Монте-Карло примерно в 10^4 раз.

Как было отмечено выше, в этом примере основные вычислительные затраты — построение массива u_{ikl}, что соответствует числу операций $\sim O(KLN)$. Вычислительные затраты на определение значения функции плотности вероятности $\boldsymbol{u}_i(\xi)$ составляют примерно $\sim O(m)$. Таким образом, построив массив u_{ikl}, можно относительно быстро находить функции плотности вероятности решения $\boldsymbol{u}_i$ при различных $\boldsymbol{p}, \boldsymbol{q}$.

5.7. Надежные оценки эмпирических распределений

В параграфе рассматриваются надежные оценки эмпирических распределений. Для этих целей используются порядковые статистики и вероятностные расширения интерполяционных полиномов и сплайнов.

Задачу интерполяции можно сформулировать следующим образом: пусть относительно значений $f_i = f(x_i)$ в точках $a = x_0 < x_1 < x_2 ... < x_n = b$, некоторой функции f, известно, что они являются случайными величинами $\boldsymbol{f}_i$ и задана их совместная функция плотности вероятности $p(f_0, f_1, \ldots, f_n)$. Возникает задача приближенного восстановления всех реализаций функции f в произвольной точке x. Далее эта задача будет решена в рамках численного вероятностного анализа с использованием понятия вероятностного расширения.

Для этих целей будут построены вероятностные расширения интерполяционных полиномов Лагранжа, кусочно-линейных функций, кубических сплайнов.

Случайные полиномы Лагранжа

Задачу интерполяции сформулируем следующим образом: пусть для некоторой функции f в точках $a = x_0 < x_1 < x_2 ... < x_n = b$ известны аппроксимации $f(x_i)$ случайными величинами $\boldsymbol{f}_i$ и задана их совместная функция плотности вероятности $p(f_0, f_1, \ldots, f_n)$. Необходимо построить случайный интерполяционный полином $\boldsymbol{l}_n(x)$: $\boldsymbol{l}_n(x_i) = \boldsymbol{f}_i$.

Рассмотрим интерполяционные полиномы Лагранжа для случая линейной интерполяции. Пусть для некоторой функции f в точках x_1, x_2 известны значения f_1, f_2.

В случае линейной интерполяции имеем точное равенство

$$f(x) = l_1(x) + \frac{(x - x_1)(x - x_2)}{2} f''(\xi),$$

где l_1 — полином Лагранжа первой степени,

$$l_1(x) = f_1 \frac{x_2 - x}{x_2 - x_1} + f_2 \frac{x - x_1}{x_2 - x_1};$$

$\xi \in [x_1, x_2]$.

В случае, когда $f_1 \in \boldsymbol{f}_1$, $f_2 \in \boldsymbol{f}_2$ известны неточно, необходимо построить линейную случайную функцию $\boldsymbol{l}(x)$ такую, что выполнены условия интерполяции $\boldsymbol{l}(x_1) = \boldsymbol{f}_1$ и $\boldsymbol{l}(x_2) = \boldsymbol{f}_2$. Таким образом, используя естественные вероятностные расширения, построим случайный полином Лагранжа первой степени

$$\boldsymbol{l}(x) = \boldsymbol{f}_1 \frac{x_2 - x}{x_2 - x_1} + \boldsymbol{f}_2 \frac{x - x_1}{x_2 - x_1}.$$

Заметим, что условия интерполяции выполнены и $\boldsymbol{l}(x)$ принимает соответствующие значения в узлах интерполяции. Отметим, что здесь сужение случайной линейной функции по константам $\boldsymbol{f}_i$ будет вещественной линейной функцией.

Далее, если необходимо построить случайную функцию $\boldsymbol{l}$, для которой выполнено включение $f \in \boldsymbol{l}$, $\forall x \in [x_1, x_2]$, то необходимо знать априорную информацию о плотности вероятности $f'' \in \boldsymbol{f}''$ на отрезке $[x_1, x_2]$. Заметим, что

$$f(x) = l(x) + \frac{(x - x_1)(x - x_2)}{2} f''(\xi).$$

Получаем оценку

$$f(x) \in \boldsymbol{l}(x) + \frac{(x - x_1)(x - x_2)}{2} \boldsymbol{f}''.$$

Рассмотрим интерполяционный полином Лагранжа в общем случае. Справедливо представление

$$\boldsymbol{l}_n(x) = \sum_{i=0}^{n} \boldsymbol{f}_i \omega_i(x),$$

где $\omega_i(x)$ — базисные функции,

$$\omega_i(x) = \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)}.$$

Таким образом, вычисление интерполяционного полинома Лагранжа в произвольной точке сводится к вычислению суммы $\boldsymbol{f}_i$ с весами. Наиболее просто это осуществляется в ситуации независимости случайных величин $\boldsymbol{f}_i$, поскольку попадает под условия теоремы 8.

Случайная кусочно-линейная интерполяция

При значениях $n > 5$ использование интерполяционных полиномов Лагранжа не эффективно. В этом случае можно использовать кусочно-линейную интерполяцию, при которой на каждом отрезке x_i, x_{i+1} реализуется полином Лагранжа первой степени:

$$\boldsymbol{l}_1(x) = \boldsymbol{f}_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + \boldsymbol{f}_{i+1} \frac{x - x_i}{x_{i+1} - x_i}.$$

Оценим математическое ожидание интерполяционных полиномов Лагранжа. В силу линейности математическое ожидание интерполяционного полинома будет линейной комбинацией математических ожиданий значений функции и совпадать с интерполяционным полиномом Лагранжа, проведенным через математические ожидания значений функции:

$$\mathsf{M}[\boldsymbol{l}_1(x)] = \mathsf{M}[\boldsymbol{f}_i] \frac{x_{i+1} - x}{x_{i+1} - x_i} + \mathsf{M}[\boldsymbol{f}_{i+1}] \frac{x - x_i}{x_{i+1} - x_i}.$$

В тех случаях, когда для математического ожидания случайной функции $\boldsymbol{f}$ известны оценки второй производной $\max_{x \in [a,b]} |M[\boldsymbol{f}^{(2)}]|$, справедлива теорема.

Теорема 10. *Пусть $\boldsymbol{l}_1$ — кусочно-линейная интерполяция случайной функции $\boldsymbol{f}$. Тогда справедлива оценка*

$$|\mathsf{M}[\boldsymbol{l}_1] - \mathsf{M}[\boldsymbol{f}]| \leq K h^2 \max_{x \in [a,b]} |\mathsf{M}[\boldsymbol{f}^{(2)}]|,$$

где K — константа, не зависящая от h.

Доказательство повторяет аналогичное доказательство для вещественных интерполяционных полиномов.

Рассмотрим свойства дисперсии кусочно-линейной интерполяции $\boldsymbol{l}_1$ случайной функции $\boldsymbol{f}$.

Теорема 11. *Дисперсия кусочно-линейной интерполяции $\boldsymbol{l}_1$ удовлетворяет следующей оценке:*

$$\mathsf{D}[\boldsymbol{L}_1] \le \max_{i=0}^{n-1}\{\max\{\mathsf{D}[\boldsymbol{F}_i], \mathsf{D}[\boldsymbol{F}_{i+1}], K_{i,i+1}\}\}.$$

Доказательство. Дисперсия линейной функции l_1 на отрезке $[x_i, x_{i+1}]$ выражается формулой [6]

$$\mathsf{D}[\boldsymbol{f}_i a_i + \boldsymbol{f}_{i+1} a_{i+1}] = \mathsf{D}[\boldsymbol{f}_i] a_i^2 + \mathsf{D}[\boldsymbol{f}_{i+1}] a_{i+1}^2 + 2 a_i a_{i+1} K_{i,i+1} \le$$

$$\le \max\{D[\boldsymbol{f}_i], \mathsf{D}[\boldsymbol{f}_{i+1}], K_{i,i+1}\}(a_i^2 + a_{i+1}^2 + 2 a_i a_{i+1}),$$

где $a_i = \frac{x_{i+1}-x}{x_{i+1}-x_i}$, $a_{i+1} = \frac{x-x_i}{x_{i+1}-x_i}$, $K_{i,i+1}$ — корреляционный момент $\boldsymbol{f}_i$, $\boldsymbol{f}_{i+1}$. Поскольку $a_i + a_{i+1} = 1, a_i \in [0,1]$, то $a_i^2 + a_{i+1}^2 + 2 a_i a_{i+1} = (a_i + a_{i+1})^2 = 1$.

Случайные сплайны

Рассмотрим вопросы построения случайных сплайнов. Они вводятся как вероятностные расширения по определенным параметрам соответствующих вещественных сплайнов. При небольшой априорной информации об интерполируемой функции строятся ее вероятностные приближения.

Приведем необходимые для дальнейшего изложения элементы теории сплайнов [2]. Пусть на отрезке $[a, b]$ задана сетка

$$\omega = \{x_i | a = x_0 < x_1 < ... < x_N = b\}$$

с целым $N \ge 2$ и шагами $h_i = x_{i+1} - x_i, \quad h = \max_{i=0}^{N-1} h_i$.

Функцию s называют сплайном степени n дефекта k (k — целое, $1 \le k \le n$) с узлами на ω, если:

$$1) \qquad s(x) = \sum_{i=0}^{n} a_{ij}(x - x_j)^i \qquad [x_j, x_{j+1}], \quad j = 0, ..., N-1; \tag{5.13}$$

$$2) \qquad s \in C^{n-k}[a, b]. \tag{5.14}$$

Множество сплайнов, удовлетворяющих этим условиям, обозначим через S_n^k.

Обратимся к вопросу интерполирования заданных функций сплайнами нечетных степеней. Положим целое $m = (n-1)/2$ и будем считать дефект $k \le m+1$.

Пусть $f \in C^{n-k}[a,b]$. Поставим задачу определения сплайна $s \in S_n^k$, интерполирующего функцию f в следующем смысле:

$$s^{(j)}(x_i) = f^{(j)}(x_i),\ i = 1, ..., N-1,\ j = 0, ..., k-1. \tag{5.15}$$

Дополнительно задается еще по $m+1$ краевому условию на $[a,b]$. Часто они берутся в виде

$$s^{(j)}(a) = f^{(j)}(a), \quad s^{(j)}(b) = f^{(j)}(b), \quad j = 0, ..., m. \tag{5.16}$$

Условия (5.15), (5.16) приводят, соответственно, к $2(N-1)k$ и $(n+1)$ — уравнениям для коэффициентов a_{ij}. Еще $(n-2k+1)(N-1)$ уравнений вытекают из условия непрерывности производных в соответствии с (5.14). В итоге получается система линейных алгебраических уравнений с квадратной матрицей $A \in R^{N(n+1)\times N(n+1)}$, вектором неизвестных $a \in R^{N(n+1)}$, известной правой частью $b \in R^{N(n+1)}$. Не исследуя общего случая, в дальнейшем выпишем эту систему для конкретных примеров и исследуем ее разрешимость.

Теперь перейдем к определению случайных сплайнов вещественного аргумента. Необходимость их использования возникает, когда вместо точных значений функции f и ее производных известны только случайные переменные, которые их определяют. Пусть известны случайные константы:

$$f^{(j)}(x_i) \in \boldsymbol{f}_i^j, \quad i = 1, ..., N-1, \quad j = 0, ..., k-1; \tag{5.17}$$

$$f^{(j)}(a) \in \boldsymbol{f}_0^j, \quad f^{(j)}(b) \in \boldsymbol{f}_N^j, \quad j = 0, ..., m. \tag{5.18}$$

Случайным сплайном назовем функцию $\boldsymbol{s} : [a,b] \to \boldsymbol{R}$ с значениями:

$$\boldsymbol{s}(x) = \{s(x) | s \in S_n^k, \quad s^{(j)}(x_i) \in \boldsymbol{f}_i^j, \quad i = 1, \ldots, N-1,$$
$$j = 0, ..., k-1; \quad s^{(j)}(a) \in \boldsymbol{f}_0^j, \quad s^{(j)}(b) \in \boldsymbol{f}_N^j, \quad j = 0, ..., m\}. \tag{5.19}$$

Эта случайная функция является вероятностным расширением соответствующего вещественного сплайна.

Для сплайнов известны следующие оценки [2].

Теорема 12. *Пусть $f \in W_\infty^p[a,b]$, $1 \le p \le n+1$ и сплайн $s \in S_n^k$ интерполирует f в смысле (5.15),(5.16). Тогда*

$$\|\partial^j (f-s)\|_\infty \le K_j h^{p-j} \|f\|_{p,\infty}, \tag{5.20}$$

где K_j — константы, не зависящие от f и h.

Перейдем к кубическим сплайнам. Рассмотрим следующую случайную функцию $\boldsymbol{s} : [a, b] \to \boldsymbol{R}$ со значениями

$$\boldsymbol{s}(x) = \{s(x) | s \in S_3^1, \quad s(x_i) \in \boldsymbol{f}_i^0, \quad i = 0, ..., N,$$

$$s''(a) \in \boldsymbol{f}_0^2, \quad s''(b) \in \boldsymbol{f}_N^2\}.$$

В итоге кубический сплайн на отрезках $[x_{j-1}, x_j], j = 1, ..., N$ имеет два представления [2]:

$$s(x) = M_{j-1}(x_j - x)^3/(6h_j) + M_j(x - x_{j-1})^3/(6h_j) +$$

$$+(f_{j-1} - M_{j-1}h_j^2/6)(x_j - x)/h_j + (f_j - M_j h_j^2/6)(x - x_{j-1})/h_j, \quad (5.21)$$

или

$$s(x) = m_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 -$$
$$m_j(x - x_{j-1})^2(x_j - x)/h_j^2 +$$
$$+f_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 +$$
$$+f_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3, \quad (5.22)$$

где $M_j = s''(x_j),\ m_j = s'(x_j),\ f_j = f(x_j)$.

Заменяя M_j, m_j, f_j на $\boldsymbol{M}_j$, $\boldsymbol{m}_j$, $\boldsymbol{f}_j \equiv \boldsymbol{f}_j^0$, мы получаем соответствующее представление случайного кубического сплайна. Выпишем случайные системы линейных алгебраических уравнений:

для $\boldsymbol{M}_j$

$$\mu_j \boldsymbol{M}_{j-1} + 2\boldsymbol{M}_j + \lambda_j \boldsymbol{M}_{j+1} = \boldsymbol{D}_j, \quad (5.23)$$

$$\boldsymbol{M}_0 = \boldsymbol{f}_0^2, \quad \boldsymbol{M}_N = \boldsymbol{f}_N^2,$$

$$\boldsymbol{D}_j = 6((\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)/h_{j+1} - (\boldsymbol{f}_j - \boldsymbol{f}_{j-1})/h_j)/(h_j + h_{j+1}),$$

$$\lambda_j = h_{j+1}/(h_j + h_{j+1}), \quad \mu_j = 1 - \lambda_j;$$

для $\boldsymbol{m}_j$

$$\lambda_j \boldsymbol{m}_{j-1} + 2\boldsymbol{m}_j + \mu_j \boldsymbol{m}_{j+1} = \boldsymbol{d}_j, \quad (5.24)$$

$$2\boldsymbol{m}_0 + \boldsymbol{m}_1 = 3(\boldsymbol{f}_1 - \boldsymbol{f}_0)/h_1 - h_1 \boldsymbol{f}_0^2/2,$$

$$2\boldsymbol{m}_N + \boldsymbol{m}_{N-1} = 3(\boldsymbol{f}_N - \boldsymbol{f}_{N-1})/h_N + h_N \boldsymbol{f}_N^2/2,$$

$$\boldsymbol{d}_j = 3\lambda_j(\boldsymbol{f}_j - \boldsymbol{f}_{j-1})/h_j + 3\mu_j(\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)/h_{j+1}, \quad j = 1, ..., N-1.$$

Матрицы систем (5.23), (5.24) имеют вещественные элементы, а правые части содержат случайные переменные. Кроме того, матрицы трехдиагональные и строго диагонально преобладают.

Таким образом, в силу вещественности матриц (5.23), (5.24), решения $\boldsymbol{M}$ и $\boldsymbol{m}$ можно представить как линейные комбинации от правых частей

$$\boldsymbol{M} = B_1\boldsymbol{f} \text{ или } \boldsymbol{m} = B_2\boldsymbol{f},$$

где B_1, B_2 — обратные матрицы к (5.23), (5.24).

С помощью найденных решений, используя вероятностные расширения, можно построить случайные функции $\boldsymbol{s}(x) : [a,b] \to \boldsymbol{R}$ со значениями

$$\begin{gathered}\boldsymbol{s}(x) = \boldsymbol{M}_{j-1}\left((x_j - x)^3/(6h_j) - (x_j - x)h_j/6\right) + \\ +\boldsymbol{M}_j((x - x_{j-1})^3/6h_j) - (x - x_{j-1})h_j/6) + \\ +\boldsymbol{f}_{j-1}(x_j - x)/h_j + \boldsymbol{f}_j(x - x_{j-1})/h_j,\end{gathered} \tag{5.25}$$

или

$$\begin{gathered}\boldsymbol{s}(x) = \boldsymbol{m}_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 - \\ -\boldsymbol{m}_j(x - x_{j-1})^2(x_j - x)/h_j^2 \\ +\boldsymbol{f}_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 + \\ +\boldsymbol{f}_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3,\end{gathered} \tag{5.26}$$

при $x \in [x_{j-1}, x_j], j = 1, ..., N$.

Заметим, что если в (5.25) брать сужение $\boldsymbol{s}$ по случайным константам $\boldsymbol{M}_j$, то в общем случае мы не получим кубического сплайна, так как $\boldsymbol{M}_j$ могут не удовлетворять системе (5.23) и, как следствие этого, $s' \notin C[a,b]$. Если брать сужение по $\boldsymbol{m}_j$ в (5.26), то $s' \in C[a,b]$, но $s'' \notin C[a,b]$.

Заметим, что сплайн на каждом отрезке $[x_{j-1}, x_j]$ можно представить как линейную комбинацию

$$\boldsymbol{s}(x) = \sum_i \boldsymbol{f}_i\psi_i(x),$$

где $\psi_i(x)$ — некоторые вещественные функции. В этом представлении сужение $\boldsymbol{s}$ по случайным константам $\boldsymbol{f}_i$ не нарушает гладкости сплайна.

Рассмотрим случайные эрмитовы кубические сплайны $\boldsymbol{s}_1 : [a,b] \to \boldsymbol{R}$. На каждом интервале $[x_{j-1}, x_j], j = 1, ..., N$, согласно формуле (5.22), эти сплайны представимы в виде

$$\begin{gathered}\boldsymbol{s}_1(x) = \boldsymbol{f}_{j-1}v((x - x_{j-1})/h_j) + \boldsymbol{f}^1_{j-1}w((x - x_{j-1})/h_j) + \\ +\boldsymbol{f}_j v((x - x_j)/h_j) + \boldsymbol{f}^1_j w((x - x_j)/h_j),\end{gathered}$$

где $v(x) = (|x| - 1)^2(2|x| + 1); w(x) = x(|x| - 1)^2$.

Кроме того, сужение $\boldsymbol{s}_1$ по константам $\boldsymbol{f}_j, \boldsymbol{f}^1_j$ будет эрмитовым кубическим сплайном. Следовательно, при аппроксимации функций эрмитовыми сплайнами наличие ошибок в данных не снижает гладкости сплайна.

Надежные оценки функции распределения

В разделе рассматривается построение надежных оценок эмпирической функции распределения.

Пусть $(\xi_1, \ldots, \xi_n)$ — выборка случайной величины X с функцией распределения $F(t), t \in [a, b]$. Эмпирическая функция распределения определяется следующим образом:

$$F_n(t) = \frac{m_t}{n}, \tag{5.27}$$

где m_t — число элементов $\xi_i < t$.

Пусть $z_i = F(x_i), i = 1, \ldots, n$. Заметим, что $z_i, i = 1, \ldots, n$ — равномерно распределенные случайные величины на $[0, 1]$. Если $z_1 \leq z_2 \leq \ldots \leq z_n$, тогда z_k — k-я порядковая статистика и математическое ожидание $\mathrm{M}[z_k] = k/(n+1)$ [23]. Далее будем использовать точки $(x_i, i/(n+1))$ для построения аппроксимации функции распределения $F(t)$.

Пусть $\omega = \{a = \xi_0 < \xi_1 < \xi_2 < \ldots < \xi_n < b = x_{n+1}\}$ — сетка. Построим на ω кусочно-линейную функцию s:

$$s(x_i) = i/(n+1), \quad i = 1, \ldots, n, s(a) = 0, s(b) = 1.$$

Заметим, если бы мы могли вместо математических ожиданий $i/(n+1)$ использовать точные значения z_i, тогда погрешность кусочно-линейной функции $s(x)$ на сетке $\{x_i\}$ с $h = \max_{i=0}^{n-1}(x_{i+1} - x_i), i = 0, \ldots, n$ удовлетворяла бы оценке

$$||F - s|| \leq Kh^2||F^{(2)}||.$$

Таким образом, даже при относительно небольших n, построенные оценки достаточно хорошо аппроксимируют функцию распределения F. Относительно z_i известно, что они образуют порядковые статистики.

Плотность вероятности k-й порядковой статистики

$$p_k(z) = \frac{n!}{(n-k)!(k-1)!} z^{k-1}(1-z)^{n-k}, \quad z \in [0, 1].$$

Совместная плотность вероятности вектора (z_j, z_k) выражается следующим образом:

$$p_{j,k}(z_j, z_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} z_j^{j-1}(z_k - z_j)^{k-j-1}(1-z_k)^{n-k},$$

$$j < k, \quad 0 \leq z_j \leq z_k \leq 1.$$

Каждому случайному вектору $(z_1, z_2, ..., z_n)$ соответствует кусочно-линейная функция s. Перебирая все возможные случайные векторы $(z_1, z_2, ..., z_n)$, получаем все множество кусочно-линейных функций $\{s\}$. Заметим, что $\{s\}$ содержит интерполянт функции распределения F. Таким образом, используя для значения в узле ξ_k плотность вероятности k-й порядковой статистики, множество $\{s\}$ можно представить в виде случайной кусочно-линейной функции $\boldsymbol{L}$. Соответственно, $\boldsymbol{L}$ — надежная оценка эмпирической функции распределения.

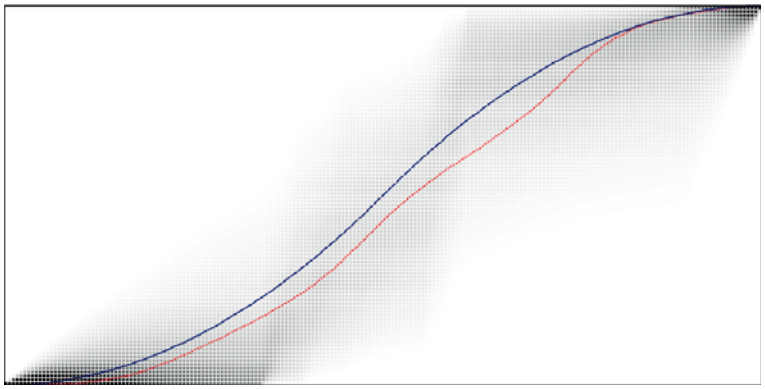


Рис. 5.5. Надежная оценка функции распределения

Рассмотрим аппроксимацию случайной кусочно-линейной функции $\boldsymbol{L}$ в виде гистограммной функции распределения (ГФР) [19].

Построим две сетки $\omega_y = \{y_i = i/N_y, i = 0, 1, \ldots, N_y\}$ и $\omega_t = \{0 = t_0 < t_1 < \ldots < t_{N_t}\}$. Не ограничивая общности, будем считать $\omega \subset \omega_t$.

На сетке ω построим гистограммную кусочно-линейную функцию. Для значения в узле ξ_k будем использовать плотность вероятности k-й порядковой статистики. Поскольку известны совместные плотности вероятности для порядковых статистик, то, используя численные операции над плотностями случайных величин, построим на сетке ω_y гистограммные значения h_t в узлах сетки ω_t.

Рассмотрим сетку $\Omega_h = \omega_t \times \omega_y$. Далее, используя гистограммы h_t, несложно в центре каждой ячейки сетки Ω_h восстановить среднее значение. Таким образом, на сетке Ω_h получаем кусочно-постоянную аппроксимацию, где (H_{ij}) — матрица значений в ячейках сетки.

Перепишем матрицу (H_{ij}) в виде списка гистограммной функции распределения (h_j, p_j), где h_j — гистограммы, образованные значениями H_{ij} при фиксированном j.

На рис. 5.5 представлена кусочно-постоянная аппроксимация надежной оценки эмпирической функции распределения, размерность выборки

— 7. Оттенками серого представлены плотности вероятности, сплошной линией показана точная функция распределения, точечной линией — математические ожидания гистограмм h_j.

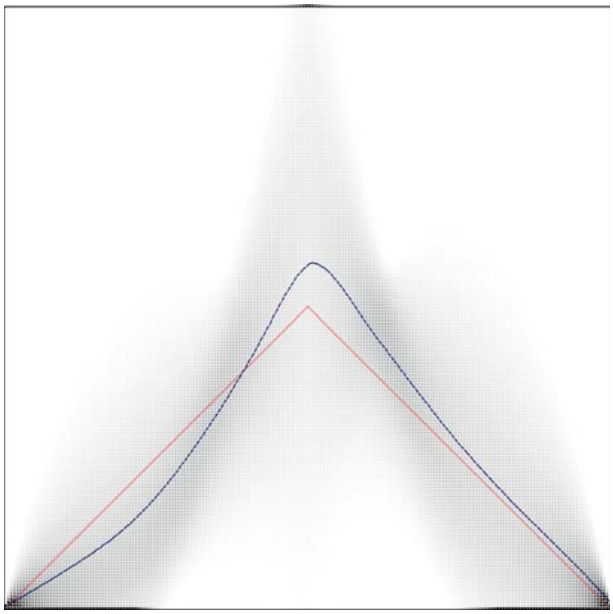


Рис. 5.6. Надежная оценка функции плотности вероятности

Надежные оценки функции плотности вероятности

Основным недостатком кусочно-линейных функций является невозможность вычислить от них производные и получить оценки функции плотности вероятности. Рассмотрим использование эрмитовых сглаживающих сплайнов для построения аппроксимаций функций распределения.

Пусть $(\xi_1, \ldots, \xi_n)$ — выборка случайной величины X с функцией распределения $F(t), t \in [a, b]$, сетка $\{x_i, i = 0, 1, \ldots, N\}$, $x_0 = a$, $x_N = b$. Далее, φ_l — базис в пространстве сплайнов Эрмита. Будем искать сплайн в виде

$$s(x) = \sum_{l=1}^{m} s_l \varphi(x),$$

с граничными условиями

$$s'(a) = 0, s(a) = 0, s'(b) = 0, s(b) = 1.$$

Для нахождения неизвестных параметров s_l будем использовать метод

наименьших квадратов

$$\sum_{i=1}^{n}(s(\xi_i) - z_i)^2 + \alpha||s''||^2 \to \min,$$

где $\alpha \geq 0$ — параметр. Заметим, что нахождение сплайна сводится к решению систем линейных алгебраических уравнений с трехдиагональной матрицей.

Таким образом, подобно описанному выше, для точного вектора (z_1, $z_2, ..., z_n$) получаем оценку точности (5.20). Продифференцировав сплайн s, получаем аппроксимацию функции плотности вероятности и, используя плотности вероятности порядковых статистик, строим гистограмму второго порядка [17].

На рис. 5.6 представлена гистограмма второго порядка, приближающая треугольную плотность вероятности (точечная линия). Исходная выборка имеет размерность 7. Сплошной линией показано математическое ожидание гистограммы второго порядка.

В параграфе показано использование случайных интерполяционных многочленов и сплайнов для построения надежных оценок эмпирических функций распределения и функций плотности вероятности.

Глава 6

Алгебраические задачи с неопределенностями

Решение систем линейных и нелинейных уравнений — одна из самых востребованных задач вычислительной математики. Они используются практически во всех разделах математического моделирования. Заметим, что при решении реальных практических задач коэффициенты матриц и правых частей редко известны точно. Один из подходов решения задач с неточными данными — метод Монте-Карло [27]. При всех его положительных качествах этот метод обладает рядом недостатков. Один из самых существенных — низкая скорость сходимости и большой объем вычислений, что представляет дополнительные сложности при работе с данными большого объема. С середины 60-х годов прошлого века стал развиваться альтернативный подход решения задач с неточными данными — интервальный анализ. Для его численной реализации необходимо знать интервалы изменений случайных величин. Соответственно, интервальный анализ дает только границы множеств решений исходных задач. В тех случаях, когда известны не только границы, но и функции плотности вероятности случайных величин, возможно применение вычисленного вероятностного анализа. Одна из первых работ этого направления [30].

6.1. Интервальные СЛАУ

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b, \tag{6.1}$$

где $b = (b_i), i = 1, \ldots, n$ — известный вектор; $A = (a_{i,j}), i, j = 1, \ldots, n$ — невырожденная матрица.

Тогда вектор $x = A^{-1}b$ — решение системы (6.1). Предположим, что A и b содержат ошибки и известно, что их элементы принадлежат соответствующим интервальным числам

$$a_{i,j} \in \boldsymbol{a}_{i,j},$$

$$b_i \in \boldsymbol{b}_i, i, j = 1, \ldots, n.$$

Пусть также все матрицы из множества $\boldsymbol{A}$ не вырождены. Множество векторов

$$\mathcal{X} = \{x | Ax = b, A \in \boldsymbol{A},\ b \in \boldsymbol{b}\}$$

назовем *множеством решений системы*

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{6.2}$$

Пример 5. Пусть необходимо решить систему интервальных линейных алгебраических уравнений

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

с матрицей $\boldsymbol{A}$ и правой частью $\boldsymbol{b}$.

$$\boldsymbol{A} = \begin{pmatrix} [2,4] & [-2,1] \\ [-1,2] & [2,4] \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} [-2,2] \\ [-2,2] \end{pmatrix}. \tag{6.3}$$

Множество векторов $\mathcal{X}$ для этой задачи изображено на рис. 6.1.

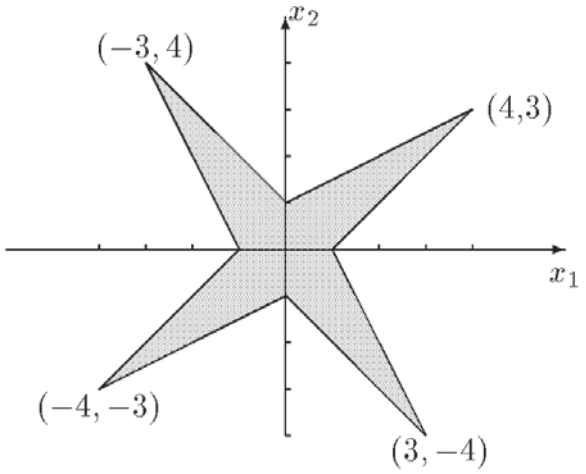


Рис. 6.1. Множество решений системы (6.3)

Минимальным интервальным вектором, содержащим множество ее решений, является интервальный вектор $\boldsymbol{x} = ([-4, 4], [-4, 4])^T$.

Множество $\mathcal{X}$ может быть описано следующим образом:

$$\mathcal{X} = \{x | x \in R^n, \boldsymbol{A}x \cap \boldsymbol{b} \neq \emptyset\} \tag{6.4}$$

или

$$\{x | x \in R^n, 0 \in \boldsymbol{A}x - \boldsymbol{b}\}.$$

Справедливо следующее утверждение:

$$x \in \mathcal{X} \Leftrightarrow |\operatorname{mid} Ax - \operatorname{mid} b| \leq \operatorname{rad}(\boldsymbol{b}).$$

Поставим задачу найти интервальный вектор $\boldsymbol{x} \in \boldsymbol{R}^n$, содержащий множество $\mathcal{X}$.

Самым простым способом для нашего примера будет метод Крамера. Действительно, для СЛАУ $A = (a_{ij})$, $b = (b_i)$, $i, j = 1, 2$. Решение находится по формулам

$$\boldsymbol{x}_1 = \frac{\Delta_1}{\Delta}, \boldsymbol{x}_2 = \frac{\Delta_2}{\Delta},$$

где

$$\Delta = a_{11}a_{22} - a_{12}a_{21},$$
$$\Delta_1 = b_1a_{22} - b_2a_{21},$$
$$\Delta_2 = a_{11}b_2 - a_{12}b_1.$$

Для решения СЛАУ с интервальными коэффициентами из примера (6.3) построим интервальное расширение формул Крамера. Непосредственными вычислениями получаем $\boldsymbol{x}_1 = [-6, 6]$, $\boldsymbol{x}_2 = [-6, 6]$.

Как видим, ответ несколько шире оптимального. Этот факт легко объясняется, поскольку рациональные выражения для $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ не удовлетворяют условию теоремы о естественных интервальных расширениях, т. е. содержат переменные более одного раза.

Применять формулы Крамера для решения больших систем крайне нерационально. В вычислительной математике для этих целей используют прямые методы: метод Гаусса, LU- и QR-разложения и т. д.

Рассмотрим еще ряд примеров.

Пример 6. Пусть необходимо решить систему интервальных линейных алгебраических уравнений

$$\boldsymbol{A} = \begin{pmatrix} [2, 4] & [-1, 1] \\ [-1, 1] & [2, 4] \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} [-3, 3] \\ [0, 0] \end{pmatrix}. \tag{6.5}$$

Множество векторов $\mathcal{X}$ этой задачи

$$\mathcal{X} = \{x | x \in R^2, 2|\boldsymbol{x}_2| \leq |\boldsymbol{x}_1|, 2|\boldsymbol{x}_1| \leq 3 + |\boldsymbol{x}_2|\}$$

и изображено на рис. 6.2.

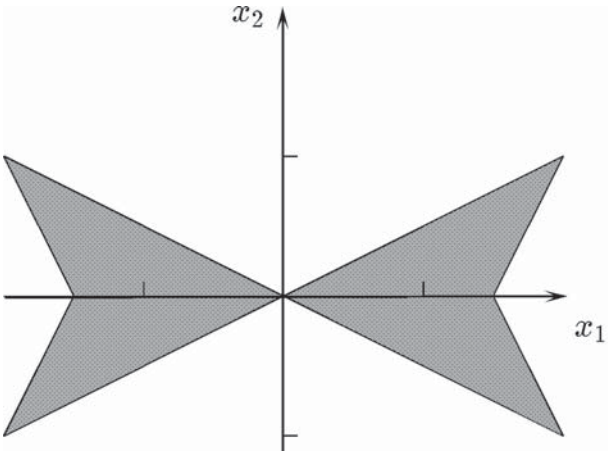


Рис. 6.2. Множество решений системы (6.5)

Пример 7. Пусть

$$\boldsymbol{A} = \begin{pmatrix} 2 & \alpha \\ \beta & 2 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 1.2 \\ -1.2 \end{pmatrix}, \quad \alpha, \beta \in [0, 1]. \tag{6.6}$$

Заметим, что по правилу Крамера

$$\boldsymbol{x}_1 = 1.2(2 - \alpha)/(4 - \alpha\beta),$$

$$\boldsymbol{x}_2 = -1.2(2 - \beta)/(4 - \alpha\beta).$$

Следовательно,

$$\square\mathcal{X} = \begin{pmatrix} [0.3, 0.6] \\ [-0.6, -0.3] \end{pmatrix}.$$

Далее $\square$ — символ интервальной оболочки [10].

Применяя формально правило Крамера к системе (6.6), мы получаем интервальный вектор $([0.3, 1.2], [-0.8, 1.2])$.

Найдем обратную матрицу

$$A^{-1} = \square \left\{ \frac{1}{4 - \alpha\beta} \begin{pmatrix} 2 & \alpha \\ \beta & 2 \end{pmatrix} \;\middle|\; \alpha, \beta \in [0, 1] \right\} =$$

$$= \begin{pmatrix} [1/2, 2/3] & [0, 1/3] \\ [0, 1/3] & [1/2, 2/3] \end{pmatrix}.$$

Несложно убедиться, что

$$A^{-1}b \neq \square\mathcal{X}.$$

Пусть A — симметричная матрица.

$$A = \begin{pmatrix} 2 & \alpha \\ \alpha & 2 \end{pmatrix}, \alpha \in [0, 1].$$

Множество решений системы линейных алгебраических уравнений с симметричной матрицей представляет собой отрезок с концами $[0.4, -0.4]$, $[0.6, -0.6]$ (рис. 6.3). Интервальная оболочка множества решений с симметричной матрицей

$$\square\mathcal{X}_{\text{sym}} = \begin{pmatrix} [0.4, 0.6] \\ [-0.6, -0.4] \end{pmatrix} \neq \square\mathcal{X}.$$

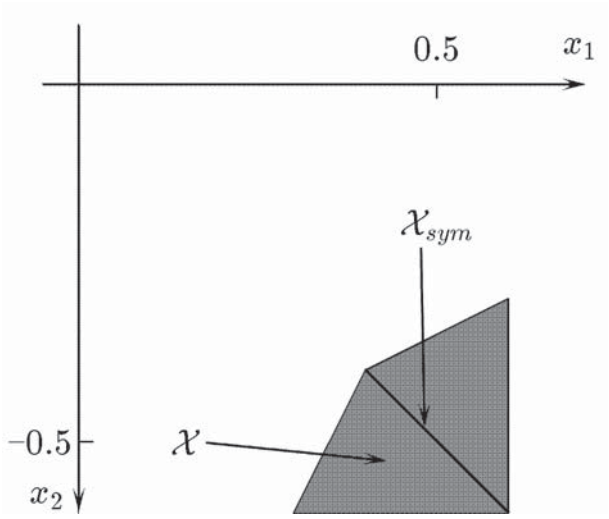


Рис. 6.3. Множество решений с симметричной матрицей

6.2. Системы линейных алгебраических уравнений со случайными коэффициентами

В главе рассматриваются задачи и методы численного решения систем линейных алгебраических уравнений с различными видами неопределенностей в данных. Актуальность данной тематики связана с тем,

что многие практические задачи, такие как задачи определения оптимальных параметров сложных систем, анализа рисковых ситуаций, обработки и анализа изображений и сигналов, задачи цифровой экономики, управления сложными системами и другие при адекватном математическом описании сводятся к решению соответствующих систем линейных и нелинейных уравнений.

Несмотря на то, что сегодня для проведения научно-технических расчетов имеется широкий выбор программно-аналитических систем и библиотек алгоритмов и программ для решения разнообразных задач линейной алгебры, в основе которых лежат строгие математические обоснования и предположения, имеется необходимость учитывать степень влияния неопределенностей в исходных данных на результат получаемых решений, поскольку даже незначительные погрешности случайного характера могут привести к тому, что полученные решения будут достаточно далеки от реальности. Ниже предлагаются методы исследования и численного решения систем линейных не линейных уравнений в условиях случайной неопределенности с использованием ВВА.

Системы линейных уравнений со случайными коэффициентами. Рассмотрим системы линейных уравнений

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, i = 1...n,$$

где $\boldsymbol{x} \in \boldsymbol{R}^n$ — случайный вектор решения; $\boldsymbol{A} = (\boldsymbol{a}_{ij})$, $\boldsymbol{b} = (\boldsymbol{b}_i)$ — случайная матрица и вектор правой части.

Предположим, что случайная матрица $\boldsymbol{a}_{ij}$ и вектор $\boldsymbol{b}_i$ имеют независимые компоненты с плотностями вероятности pa_{ij}, pb_i соответственно.

Пример 8. Пусть матрица A имеет вид

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Предположим, что случайный вектор $\boldsymbol{b}$ состоит из независимых компонент $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, каждая из которых — случайная величина, равномерно распределенная на отрезке $[0, 1]$. Тогда носитель плотности вероятности вектора $\boldsymbol{b}$ — квадрат $[0, 1]^2$.

Построим функцию распределения случайного вектора $\boldsymbol{x}$. Заметим вероятность того, что $x_1 < r_1$ и $x_2 < r_2$ равна площади, отсекаемой от квадрата $[0, 1]^2$ прямыми, проходящими через точку $b_0 = Ar$, где $r = (r_1, r_2)^T$ с направляющими векторами $l_1 = (2, -1)$; $l_2 = (-1, 2)$.

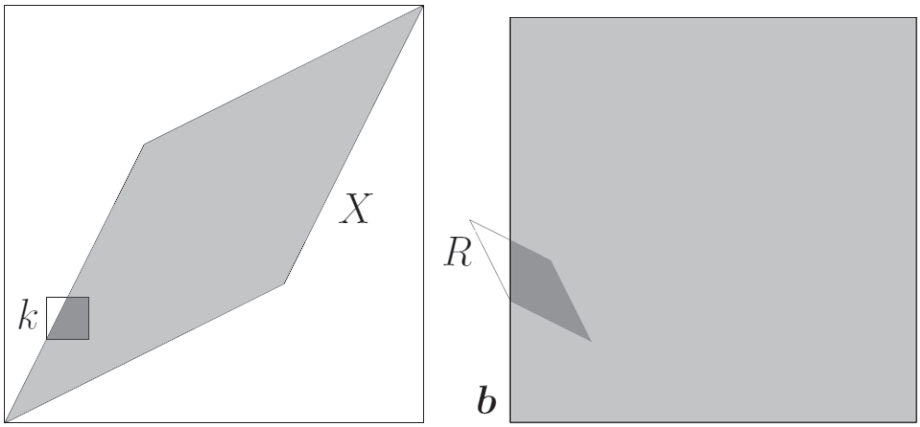


Рис. 6.4. Вычисление плотности вероятности вектора решений

На рис. 6.4 в левой части серым цветом закрашено X — множество решений системы. Для вычисления вероятности попадания в квадрат k отобразим его с помощью матрицы A на вектор правой части $\boldsymbol{b}$ (правая часть рисунка). Квадрат k отображается в ромб R. Таким образом, вероятность попадания решения в квадрат k равняется интегралу от плотности вероятности вектора $\boldsymbol{b}$ по области пересечения ромба R и вектора правой части $\boldsymbol{b}$.

На рис. 6.5 приведено кусочно-постоянное с шагом $h = 0{,}1$ приближение совместной плотности вероятности вектора $\boldsymbol{x}$. Сплошной линией проведена граница множества решений исходной системы. Величина вероятности в точности пропорциональна площади пересечения элементарного квадрата и множества решений.

Заметим, что при $h \to 0$ приближение совместной плотности вероятности стремится к точному P_x — равномерному с носителем, совпадающим с границами множества решений исходной системы. Непосредственными вычислениями легко убедиться, что

$$A\boldsymbol{x} = \boldsymbol{b}.$$

Перейдем теперь к случаю, когда $\boldsymbol{A} = (\boldsymbol{a}_{ij})$ — случайная матрица. Каждому $x \in X$ можно сопоставить подмножество коэффициентов $A_x \subset \boldsymbol{A}, b_x \subset \boldsymbol{b}$

$$\Omega_x = \{A, b | Ax = b, A \in \boldsymbol{A}, b \in \boldsymbol{b}\}.$$

Заметим, что для фиксированного x коэффициенты матрицы и вектора правой части связаны соотношением

$$\sum_{j=1}^{n} a_{ij}x_j - b_i = 0, i = 1, ..., n;$$

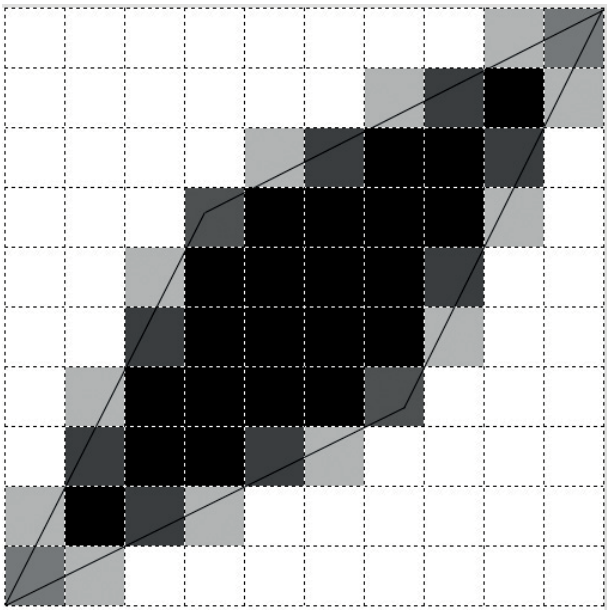


Рис. 6.5. Кусочно-постоянное приближение совместной плотности вероятности

следовательно,

$$\Omega_x = \{A, b | \sum_{j=1}^{n} a_{ij} x_j - b_i = 0, i = 1, ..., n\}.$$

Пусть необходимо найти вероятность $P(X_0)$ попадания решения x в некоторое подмножество $X_0 \subset X$. Сопоставим X_0 множество $\Omega_0 = \{\Omega_x | x \in X_0\}$.

Тогда

$$P(X_0) = \int_{\Omega_0} \Pi_{i=1}^{n} \Pi_{j=1}^{n} p a_{ij} \Pi_{i=1}^{n} p b_i d\Omega.$$

Пример 9. Рассмотрим случайную матрицу

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}_{21} & \boldsymbol{a}_{22} \end{pmatrix}.$$

Элементы матрицы a_{ij} состоят из независимых равномерно распределенных компонент, с носителями a_{11}, a_{22} на отрезке $[2, 4]$, a_{21}, a_{12} — $[-1, 0]$. Вектор $\boldsymbol{b}$ состоит из независимых компонент $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, каждая из которых — случайная величина, равномерно распределенная на отрезке $[0, 1]$.

На рис. 6.6 приведено кусочно-постоянное приближение совместной плотности вероятности вектора решения системы случайных линейных алгебраических уравнений $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$.

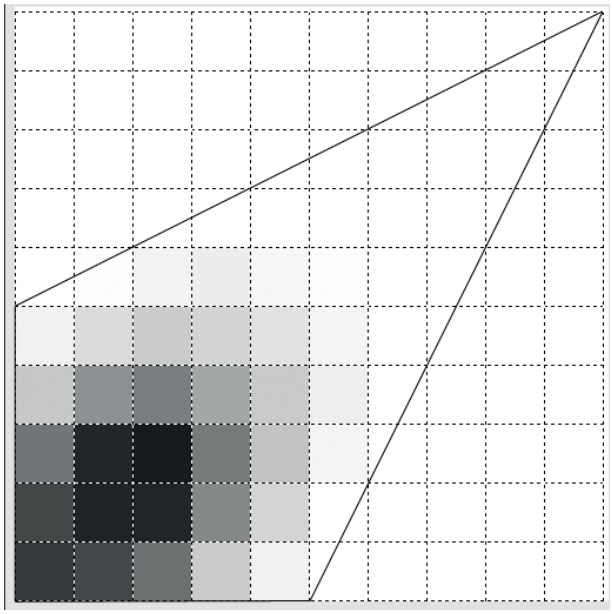


Рис. 6.6. Приближение совместной плотности вероятности вектора x

Вычисление элементов обратной матрицы. Пусть $C = A^{-1}$, $C = (c_{ij})$, тогда элемент c_{11} можно представить в виде

$$c_{11} = \frac{a_{22}}{a_{11}a_{22} - a_{12}a_{21}} = \frac{1}{a_{11} - a_{12}a_{21}/a_{22}}.$$

Заметим, что в последнем выражении каждая переменная встречается только один раз, и следовательно c_{11} можно вычислить, используя арифметику над плотностями вероятности.

6.3. Использование вероятностных расширений

В качестве одного из примеров численного моделирования рассмотрим решение системы линейных алгебраических уравнений

$$Ax = b, \tag{6.7}$$

где $A = (a_{ij})$ — случайная матрица и $b = (\boldsymbol{b}_i)$ — случайный вектор соответственно. Предположим, что случайная матрица A и вектор b имеют независимые компоненты с плотностями вероятности $\boldsymbol{A} = (\boldsymbol{a}_{ij})$, $\boldsymbol{b} = (\boldsymbol{b}_i)$, соответственно

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_{11} & \boldsymbol{a}_{12} & \dots & \boldsymbol{a}_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_{n1} & \boldsymbol{a}_{n2} & \dots & \boldsymbol{a}_{nn} \end{pmatrix}.$$

Носитель множества решений может быть представлен в виде [10]

$$\mathcal{X} = \{x | Ax = b, A \in \mathrm{supp}(\boldsymbol{A}), b \in \mathrm{supp}(\boldsymbol{b})\}.$$

Построим вероятностное расширение вектора решения $\boldsymbol{x}(\cdot, \boldsymbol{A}, \boldsymbol{b})$:

$$\boldsymbol{x}_1(\cdot, \boldsymbol{A}, \boldsymbol{b}) = \frac{\begin{vmatrix} \boldsymbol{b}_1 & \boldsymbol{a}_{12} & \dots & \boldsymbol{a}_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{b}_n & \boldsymbol{a}_{n2} & \dots & \boldsymbol{a}_{nn} \end{vmatrix}}{\begin{vmatrix} \boldsymbol{a}_{11} & \boldsymbol{a}_{12} & \dots & \boldsymbol{a}_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_{n1} & \boldsymbol{a}_{n2} & \dots & \boldsymbol{a}_{nn} \end{vmatrix}}$$

или

$$\boldsymbol{x}_1(\xi, \boldsymbol{A}, \boldsymbol{b}) = \iint \boldsymbol{a}_{12}(t_{12}) \dots \boldsymbol{a}_{nn}(t_{nn}) \frac{\sum \boldsymbol{b}_i \Delta_i(t_{12}, \dots, t_{nn})}{\sum \boldsymbol{a}_{1i} \Delta_i(t_{12}, \dots, t_{nn})}(\xi) dt_{12} \dots dt_{nn}, \tag{6.8}$$

где $\Delta_i(t_{12}, \dots, t_{nn}) \in R$ — миноры из метода Крамера для решения СЛАУ, $t_{ij} \in \mathrm{supp}(\boldsymbol{a}_{ij})$. Выражение

$$\left(\frac{\sum \boldsymbol{b}_i \Delta_i(t_{12}, \dots, t_{nn})}{\sum \boldsymbol{a}_{1i} \Delta_i(t_{12}, \dots, t_{nn})} \right)(\xi)$$

вычисляется, используя вероятностные арифметики.

Изучим вопрос построения совместной функции распределения p_x вектора $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. Рассмотрение начнем со случая, когда матрица A является детерминированной:

$$A\boldsymbol{x} = \boldsymbol{b}.$$

Тогда

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}.$$

Пусть для определенности совместная функция распределения $p(b_1, \dots, b_n)$ вектора $(\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ задана на $\Omega = [\underline{b}_1, \overline{b}_1] \times \dots \times [\underline{b}_n, \overline{b}_n]$. Ω — параллелепипед с ребрами $e_i, i = 1, \dots, n$. Возьмем $e_i, i = 1, \dots, n$ в качестве базисных, начало координат выберем в точке $O = (\underline{b}_1, \dots, \underline{b}_n)$. Тогда матрица A^{-1} как аффинное преобразование трансформирует Ω в Ω', $e_i, i = 1, \dots, n$ в $e'_i, i = 1, \dots, n$ и O в O'. Таким образом, значение плотности вероятности $p(b_1, \dots, b_n)$ в базисе $\{O, e_i, i = 1, \dots, n\}$ преобразуется в значение $\alpha p(b_1, \dots, b_n)$ в базисе $\{O', e'_i, i = 1, \dots, n\}$, где

$$\alpha(A) = \frac{|\Omega|}{|\hat{\Omega}|}.$$

Для вычисления плотности вероятности в некоторой точке $x = (x_1, \ldots, x_n)$ необходимо найти ее координаты в базисе $\{O'e'_i, i = 1, \ldots, n\}$: $x' = (x'_1, \ldots, \hat{x}'_n)$. Заметим, что

$$x'(A) = A(x - O').$$

Следовательно,

$$p_x(x_1, \ldots, x_n) = \alpha(A)p(x'(A)).$$

Окончательно получаем

$$p_x(x_1, \ldots, x_n) = \iint \boldsymbol{a}_{11}(t_{11}) \ldots \boldsymbol{a}_{nn}(t_{nn})\alpha(A(t))p(x'(A(t)))dt_{11} \ldots dt_{nn}. \tag{6.9}$$

Пример 10. Рассмотрим систему линейных алгебраических уравнений

$$Ax = b. \tag{6.10}$$

Пусть $A = (a_{ij})$ — случайная матрица $n = 2$. Элементы матрицы $\boldsymbol{A}$ независимы и распределены по треугольному закону, $\boldsymbol{a}_{11}, \boldsymbol{a}_{22}$ распределены на интервале $[2, 4]$, a_{21}, a_{12} распределены на отрезке $[-1, 1]$. Вектор $\boldsymbol{b}$ состоит из независимых компонент $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, распределенных по треугольному закону на отрезке $[0, 2]$.

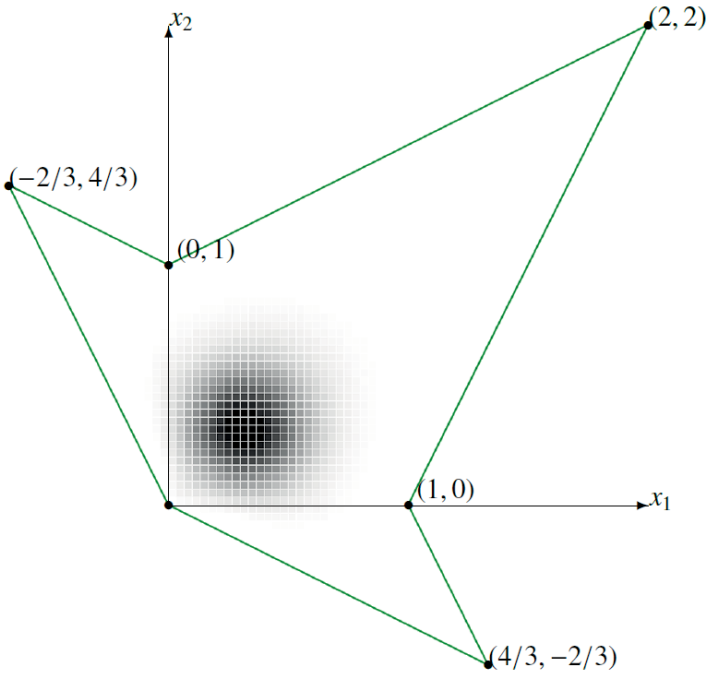


Рис. 6.7. Граница множества решений и совместная функция плотности вероятностей (x_1, x_2)

Вектор (x_1, x_2): решение (6.10):

$$x_1 = \frac{a_{22}b_1 - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}},$$

$$x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{12}a_{21}}.$$

На рис. 6.7 показана совместная плотность вероятности вектора (x_1, x_2). Значение вероятности представлено оттенками серого. Сплошная линия — граница множества решений.

В этом случае для вычисления x_1 заменим a_{22}, a_{12} на t_{22}, t_{12}. Вычислим интеграл численно для разных ξ

$$\boldsymbol{x}_1(\xi) = \iint \boldsymbol{a}_{22}(t_{22})\boldsymbol{a}_{12}(t_{12}) \left(\frac{t_{22}\boldsymbol{b}_1 - t_{12}\boldsymbol{b}_2}{\boldsymbol{a}_{11}t_{22} - t_{12}\boldsymbol{a}_{21}} \right)(\xi) dt_{22} dt_{12}$$

и получим функцию плотности вероятности 1-й компоненты вектора решений x_1.

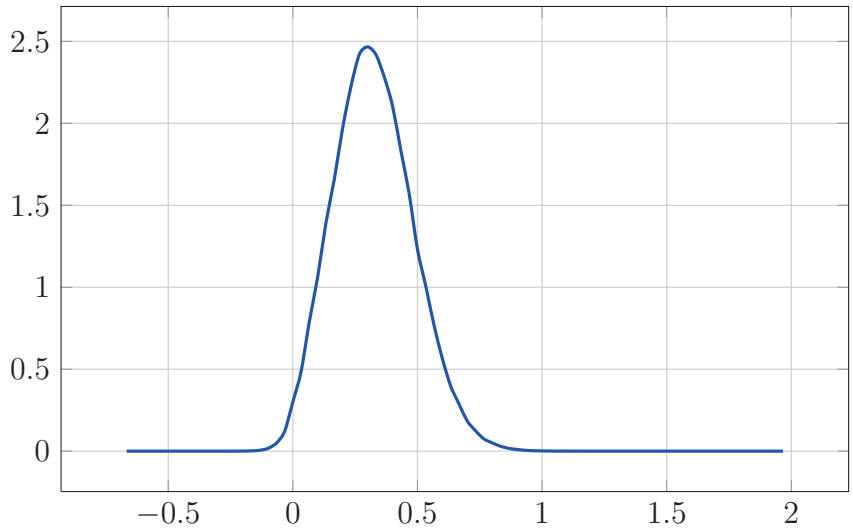


Рис. 6.8. Функция плотности вероятности x_1

На рис. 6.8 показана функция плотности вероятности первой компоненты вектора решений x_1. Плотность вероятности случайной величины x_1 имеет носитель $[-2/3, 2]$. Тем не менее вне интервала $[0, 1]$ плотность вероятности пренебрежимо мала.

6.4. Совместное использование метода Монте-Карло и вычислительного вероятностного анализа

Как известно, при переходе к сложным многомерным задачам эффективность метода Монте-Карло [27] возрастает. Принято считать, что для размерностей $D \leq 3$ предпочтительней использовать квадратурные

формулы. Для размерностей больше 10 метод Монте-Карло не имеет конкурентов.

Заметим, что размерность D задачи (6.7) $D = (n+1)n$. Использование вычислительного вероятностного анализа понижает размерность на $2n$ до $(n-1)n$.

В тех случаях, когда численное вычисление интегралов (6.8), (6.9) затруднительно, воспользуемся методом Монте-Карло совместно с вычислительным вероятностным анализом. Для этих целей, например при вычислении $\boldsymbol{x}_1(\boldsymbol{A}, \boldsymbol{b})(\xi)$ (6.8), реализуем N выборочных значений

$$\zeta_i = (t^i_{12}, \ldots, t^i_{nn}), i = 1, \ldots, N$$

случайного вектора $\zeta = (a_{12}, \ldots, a_{nn})$ согласно плотностей $\boldsymbol{a}_{12}, \ldots, \boldsymbol{a}_{nn}$. Для каждого вектора ζ_i вычислим

$$I_i = t^i_{12} \cdot \ldots \cdot t^i_{nn} \frac{\sum \boldsymbol{b}_i \Delta_i(t_{12}, \ldots, t_{nn})}{\sum \boldsymbol{a}_{1i} \Delta_i(t_{12}, \ldots, t_{nn})}(\xi), i = 1, \ldots, N. \tag{6.11}$$

Следовательно, согласно методу Монте-Карло [27]

$$\boldsymbol{x}_1(\xi, \boldsymbol{A}, \boldsymbol{b}) \approx \bar{I} = \frac{1}{N} \sum_i^N I_i.$$

Таким образом, совместное использование Монте-Карло и ВВА позволит уменьшить время расчетов и поднять точность.

Предложенный подход позволяет решить задачу вычисления функции плотности вероятности в процессах моделирования со случайными входными данными. Для этих целей мы предлагаем использовать параллельно-рекурсивную организацию вычислительного процесса. Таким образом, важная проблема вычисления вероятностных расширений может быть решена в рамках параллельного рекурсивного программирования. Это открывает множество возможностей для изучения различных моделей со случайными входными данными. Быстрые и точные вычисления основаны на свойствах числовых арифметических процедур над кусочно-полиномиальными моделями, разработанными в рамках вычислительного вероятностного анализа.

6.5. Решения нелинейных уравнений

Одномерные задачи. Рассмотрим системы нелинейных уравнений

$$f_i(x, k) = 0, i = 1...n,$$

где $x \in R^n$ — случайный вектор решения; $k \in R^m$ — случайный вектор коэффициентов, $\boldsymbol{p}_k(\xi_1\xi_2...\xi_m)$ — функция совместной плотности вероятности вектора k и K — носитель $\boldsymbol{p}_k$. Множество решений X имеет вид

$$X = \{x | f(x,k) = 0, k \in K\}.$$

Каждому $x \in X$ можно сопоставить подмножество коэффициентов $K_x \subset K$:

$$K_x = \{k | f(x,k) = 0\}.$$

Пусть необходимо найти вероятность $P(X_0)$ попадания решения x в некоторое подмножество $X_0 \subset X$. Сопоставим X_0 множество коэффициентов $K_0 = \{K_x | x \in X_0\}$.

Тогда вероятность

$$P(X_0) = \int_{K_0} p_k(\xi_1, \xi_2, ..., \xi_m) d\xi_1 d\xi_2 ... d\xi_m.$$

Нахождение подмножеств $K_x \subset K$ в общем случае — непростая задача. Но для ряда частных случаев вполне реализуемая.

Рассмотрим задачу нахождения корня одномерного уравнения

$$f(x,k) = 0.$$

Предположим, что корень локализован на отрезке $[a,b]$. Заметим, что x будет представлять собой случайную величину, и необходимо найти плотность распределения $\boldsymbol{x}$ этой случайной величины. Далее мы будем предполагать, что можно с достаточной точностью для любого $z \in [a,b]$ вычислять плотность распределения ϕ_z случайной величины $f(z, \boldsymbol{k})$.

Тогда $P(z)$ есть вероятность, что корень лежит левее (правее) точки z:

$$P(z) = \int_{-\infty}^{0} \phi_z(\xi) d\xi.$$

Действительно, перепишем уравнение $f(x,k) = 0$ в виде $\phi(k) - x = 0$. Тогда $\boldsymbol{x} = \phi(\boldsymbol{k})$, и ϕ_z есть плотность распределения $\phi(\boldsymbol{k})$ или, что то же самое, плотность распределения $\boldsymbol{x}$, сдвинутая на z.

Таким образом, плотность распределения корня $\boldsymbol{x} = P'(z)$. В тех случаях, когда $P(z)$ — не аналитическая функция, и явное вычисление производной затруднено, можно вычислять производную численным способом. Для построения гистограммы g с узлами $\{z_i\}$, приближающей

плотность распределения $\boldsymbol{x}$, достаточно вычислить $P(z_i)$. Тогда значение гистограммы g_i на отрезке $[z_{i-1}, z_i]$ определяется соотношением

$$g_i = \frac{P(z_i) - P(z_{i-1})}{z_i - z_{i-1}}.$$

В качестве примера рассмотрим решение простейшего квадратного уравнения $\boldsymbol{a}x^2 - \boldsymbol{b} = 0$, где a, b — случайные величины с равномерным законом распределения, заданные, соответственно, на отрезках $[1, 2]$, $[2, 4]$. Несложно убедиться, что $\boldsymbol{x}$ задана на отрезке $[1, 2]$. Таким образом, для $\forall x \in [1, 2]$: $\boldsymbol{f}(x, \boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}x^2 - \boldsymbol{b}$.

Заметим, что ax^2 — равномерная случайная величина, заданная на отрезке $[x^2, 2x^2]$, тогда для вычисления $\boldsymbol{f}$ необходимо найти разность двух независимых равномерных случайных величин. Для построения $\boldsymbol{x}$ вычислим функцию распределения P_x в точках сетки $\{x_i = 1 + ih, h = 1/n, i = 0, 1, ..., n\}$. Далее по ней построим функцию плотности вероятности для корня квадратного уравнения. На рис. 6.9 приведена функция плотности вероятности $\boldsymbol{x}$ корня квадратного уравнения.

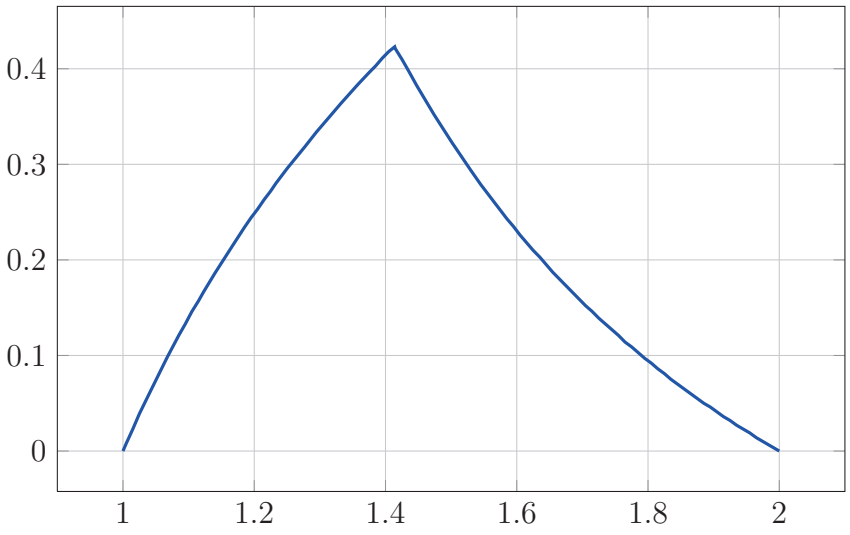


Рис. 6.9. Функция плотности вероятности корня квадратного уравнения

6.6. Системы нелинейных уравнений

Рассмотрим задачу нахождения множества решений системы нелинейных уравнений

$$f_i(x, k) = 0, i = 1, \ldots, n,$$

где $x \in R^n$ — вектор решений, $k \in R^m$ — вектор параметров. Относительно $k \in R^m$ будем предполагать, что известны функции плотности вероятности. Будем рассматривать случай $m \geq n$. Заметим, что случай строгого неравенства $m > n$ можно свести к случаю $m = n$ использованием результатов вероятностных расширений.

Рассмотрим частный случай $m = n$. Продифференцировав исходную систему нелинейных уравнений, получаем

$$F'_x(x,k)dx + F'_k(x,k)dk = 0,$$

$$dx = -(F'_x(x,k))^{-1}F'_k(x,k)dk.$$

Таким образом, зная решение x при некоторых значениях параметров k, можно получить связь dx и dk. Последнее позволяет вычислить значение совместной функции плотности вероятности решения.

Рассмотрим пример системы нелинейных уравнений

$$x^2 + y^2 - \boldsymbol{r}^2 = 0,$$

$$xy - \boldsymbol{c} = 0,$$

где r, c — равномерные случайные величины, плотности вероятности которых имеют носители $[1, 1.1]$, $[0.4, 0.5]$.

Пусть при некоторых значениях r_0, c_0 решение системы x_0, y_0, тогда справедливо

$$\begin{pmatrix} 2x_0 & 2y_0 \\ y_0 & x_0 \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} 2r_0 dr \\ dc \end{pmatrix}.$$

При этом прямоугольник S_0 со сторонами dr, dc переходит в четырехугольник S_1 со сторонами dx, dy. Решив систему, получаем значения dx, dy. Плотность вероятности множества решений в точке (x_0, y_0) пропорциональна отношению площадей $|S_0|/|S_1|$:

$$p(x_0, y_0) = p_1(r_0, c_0)|S_0|/|S_1|.$$

Таким образом, для того чтобы построить совместную функцию плотности случайных величин x, y, построим сетки в области носителей случайных величин r, c: $\{r_i, i := 0, \ldots, m$, $\{c_i, i := 0, \ldots, m$. Решим $(m+1)^2$ систем нелинейных уравнений

$$x^2 + y^2 - r_i^2 = 0,$$

$$xy - c_j = 0, i, j = 0, \ldots, m.$$

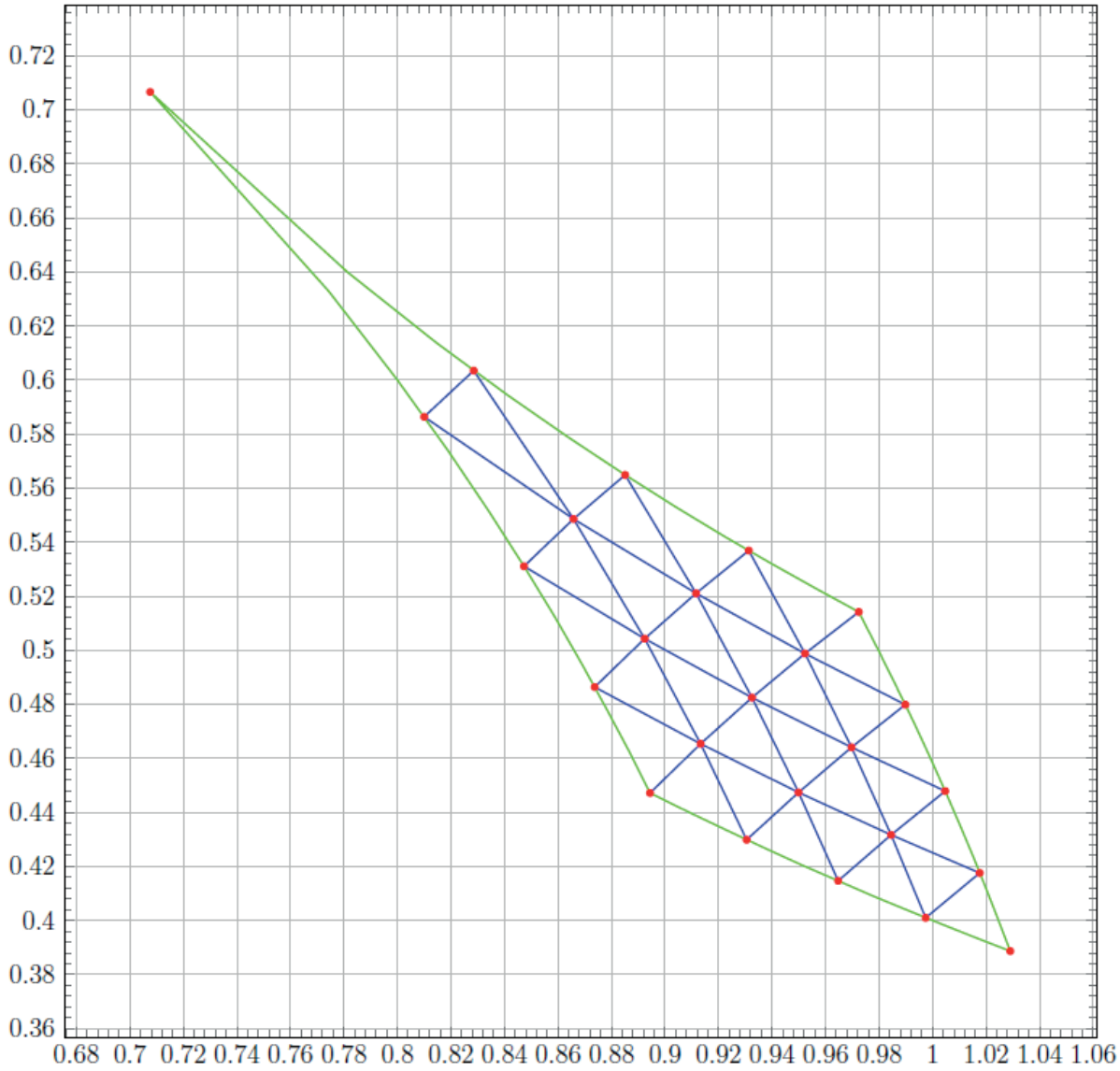


Рис. 6.10. Триангуляция области решения системы нелинейных уравнений

На рис. 6.10 представлена триангуляция области решения системы нелинейных уравнений. Вершины треугольников — точки решений с вычисленными значениями плотности совместной функции вероятности решения $p(x_i, y_j)$. Таким образом, используя линейную интерполяцию, на каждом треугольнике можно построить значение совместной функции плотности вероятности. На рис. 6.11 оттенками серого представлена совместная функция плотности решения системы нелинейных уравнений.

В данном примере для построения совместной функции плотности решения необходимо решить всего 25 раз систему нелинейных уравнений, метод Монте-Карло потребует порядка 10^6 решений. Общий случай

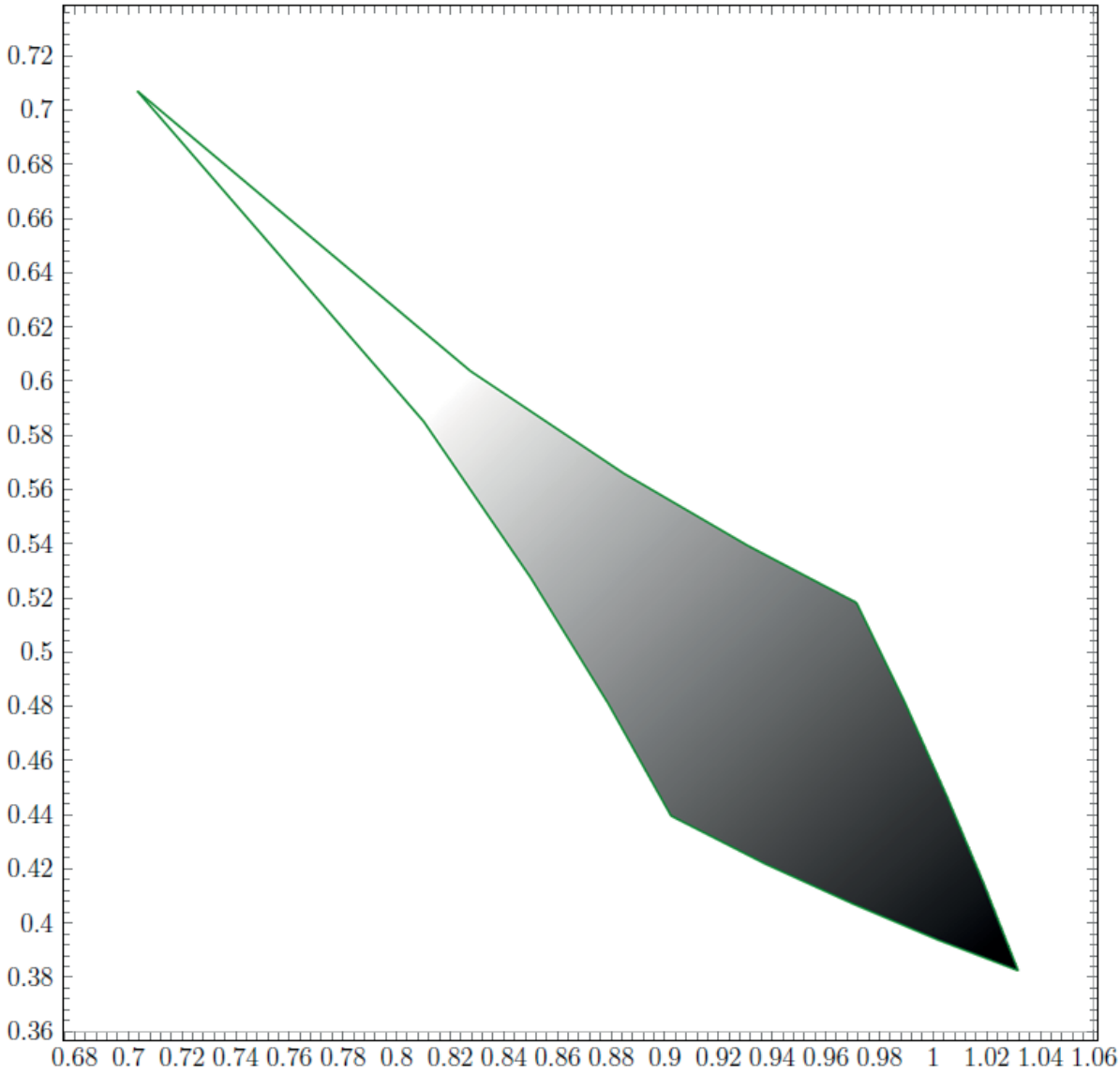


Рис. 6.11. Совместная функция плотности решения системы нелинейных уравнений

построения плотностей вероятности может быть реализован, используя построение вероятностных расширений.

Таким образом, вычислительный вероятностный анализ позволяет решать задачи стохастического моделирования с относительно небольшим числом операций.

Глава 7

Временные ряды распределений

Временные ряды — это особый способ представления данных, характеризующих изменение некоторого показателя (показателей) во времени.

В экономике это ежедневные цены на акции, курсы валют, еженедельные и месячные объемы продаж, годовые объемы производства и т. п. В метеорологии типичными временными рядами являются ежедневная температура, месячные объемы осадков, в гидрологии — периодически измеряемые уровни воды в реках. В технике временные ряды возникают при измерении значений приборов и параметров технологических процессов в последовательные моменты времени.

Многообразие систем и процессов, протекающих в них с течением времени, определяет различные виды временных рядов, а также методы и подходы к их исследованию.

Как правило, временной ряд — это последовательность чисел; его элементы — значения некоторого процесса в определенные моменты времени t_i, обычно через равные промежутки; элементы временного ряда x_i нумеруют в соответствии с номером момента времени, к которому они относятся. Порядок следования элементов временного ряда весьма существен [41].

Понятие временного ряда часто толкуют расширительно. Например, одновременно могут регистрироваться несколько характеристик процесса. В этом случае говорят о многомерных временных рядах. Если измерения производятся непрерывно, говорят о временных рядах с непрерывным временем, или о случайных процессах. Наконец, текущая переменная может иметь не временной, а какой-нибудь иной характер, например пространственный (тогда говорят о случайных полях). Особенностью измерения элементов временных рядов x_i является присутствие случайных помех, случайных ошибок и т. д.

Временной ряд распределений (ВРР) описывает ситуации, когда в течение каждого момента времени известны кусочно-полиномиальные функции, аппроксимирующие функции плотности некоторых случайных величин. Подобные ситуации возникают, когда необходима агрегация большого числа данных в некоторые моменты времени. Во многих случаях аппроксимации функции плотности вероятности более информативны, чем, например, среднее значение. Области, где ВРР полезны, включают экономику, мониторинг окружающей среды.

Обычно временной ряд — значения во времени каких-либо параметров (в простейшем случае одного) исследуемого процесса. Однако подобные ряды не описывают явления, когда реализация наблюдаемой величины доступна для каждого момента времени в виде некоторого множества.

Вот две типичных ситуации, когда это происходит:

1. Если измеряется некая переменная во времени для группы людей. Но исследование заключается не в каждом отдельном человеке, но в группе в целом. В этом случае временной ряд представляет выборочное среднее наблюдаемой величины с момента времени.

2. Когда переменная наблюдается, например, раз в секунду или в минуту, но должна быть проанализирована на более низкой частоте, скажем за день. В этом случае среднее значение и интервальный анализ многую информацию не учитывают.

Эти две ситуации описывают распределенная и временная агрегации соответственно. В каждом случае временной ряд функций плотности вероятности предложил бы более информативное представление, чем другие его формы.

Таким образом, чтобы использовать временные ряды функций плотностей вероятности, нужно определить, как представлять наблюдаемые распределения. Распределения могут быть оценены любым параметрическим или непараметрическим методом.

В этой главе рассмотрено представление функций плотности вероятности с использованием кусочно-полиномиальных функций, в том числе гистограмм, частотных полигонов и сплайнов.

Такие ряды, естественно, возникают во многих приложениях, включая экономику, финансы, метеорологию и т. д. [56]. На рис. 7.1 приведен пример ВВР глобального распределения температуры.

В символическом анализе данных и Data Mining [59] гистограммы используются для исследования множества различных процессов и при-

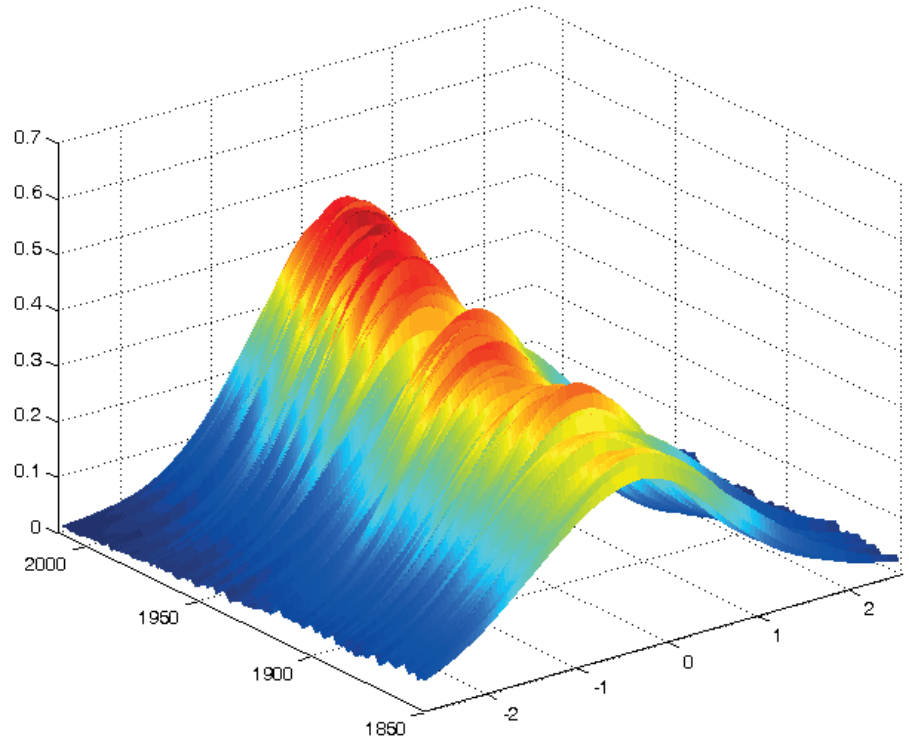


Рис. 7.1. Глобальное распределение температуры

меняются для описания изменчивости количественных признаков.

Использование во временных рядах кусочно-полиномиальных функций, аппроксимирующих плотности вероятности, порождает BPP. Важно отметить, что в анализе BPP цель состоит в том, чтобы сделать прогноз BPP в виде функций плотности вероятности.

BPP подходит для представления агрегированных данных. Очевидно, что если интерес заключается в исходных данных, агрегирование не должно быть рассмотрено. Заметим, что BPP сохраняют больше информации, чем среднее или интервал.

Для прогнозирования BPP в ряде работ использовались адаптации известных методов. В работе [82] предложены методы сглаживания, основанные на гистограммной арифметике. В работе [56] адаптирован алгоритм $k - NN$ для прогноза BPP. Способность к прогнозу и простота $k - NN$ делают свою адаптацию к BPP подходящей. Другая сила алгоритма $k - NN$ — своя многосторонность: это может быть применено к оценке плотности, классификации, приближению функции и также к прогнозированию временного ряда.

7.1. Основы временных рядов распределений

Определения. Определим временной ряд распределений как последовательность плотностей вероятности, представленных в виде кусочно-полиномиальных функций $\boldsymbol{P}_i$.

В символьном анализе данных [58, 59] интервальные, гистограммные и кусочно-полиномиальные переменные применены, чтобы описать изменчивость количественных признаков понятий.

Почему кусочно-полиномиальные переменные?

Использование кусочно-полиномиальных переменных обусловлено прежде всего тем, что они позволяют достаточно точно представлять произвольные распределения. Вторая причина — развитая арифметика для работы с кусочно-полиномиальными переменными.

Важно отметить, что кусочно-полиномиальные функции охватывают все возможные интервалы оценки плотности вероятности. Наиболее популярными из них являются гистограммы с фиксированной шириной столбцов, которые в большом объеме используются на практике, частотные полигоны и сплайны.

Причины для использования кусочно-полиномиальных функций могут быть сформулированы следующим образом:

- можно использовать их для любой исходной плотности вероятности;
- они могут описывать данные с достаточной степенью точности;
- простая и гибкая структура упрощает их использование.

7.2. Оценка погрешности для временных рядов распределений

В классических временных рядах точность оценки основана на разнице между наблюдаемыми и прогнозируемыми ценностями, т. е. между наблюдением и прогнозируемым значением. Однако из-за сложностей временных рядов распределений (ВРР) требуется другой подход к количественной оценке различий между функциями плотности вероятности.

Для этого рассмотрим метрику Вассерштейна и Mallows [97] для обратных функций распределения. Пусть $f(x)$ и $g(x)$ — функции плотности вероятности, тогда представим метрики

$$\rho_W(f,g) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt, \tag{7.1}$$

$$\rho_M(f,g) = (\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt)^{1/2}, \tag{7.2}$$

где $F^{-1}(t)$, $G^{-1}(t)$ — обратные функции к функции распределения.

Причинами выбора этой метрики было успешное применение в ряде работ [56]. Более того, эта метрика эквивалентна Earth Mover's Distance (EMD), предложенной Rubner, Tomasi и Guibas. EMD — известное компьютерное расстояние, которое используется для того, чтобы измерить несходства между гистограммами текстуры и цвета. EMD между двумя гистограммами — наименьшее количество объема работы, чтобы преобразовать одну гистограмму в другую.

Если $h(x)$ — некоторая гистограмма, тогда функцию распределения H, соответствующую этой гистограмме, можно представить в виде

$$H(x) = \int_{-\infty}^{x} h(\xi) d\xi.$$

В силу того, что гистограмма — кусочно-постоянная функция, вычисление интеграла от нее не представляет особого труда. В результате функции распределения H будет кусочно-линейная функция. Таким образом, метрики (9.2),(9.1) можно интерпретировать как площадь между функциями распределения.

В работе [82] была предложена мера оценки погрешности ГВР. Пусть $\{h_i\}$, $i = 1, 2, ..., n$ — наблюдаемые значения, $\{\hat{h}_i\}$, $i = 1, 2, ..., n$ — прогноз. Среднее расстояние ошибки (The Mean Distance Error)

$$MDE(\{h_i\}, \{\hat{h}_i\}) = \frac{1}{n} \sum_{i=1}^{n} \rho(h_i, \hat{h}_i).$$

В работе [82] была предложена средняя масштабируемая оценка погрешности ГВР (Mean Scaled Distance Error)

$$MSDE(\{h_i\}, \{\hat{h}_i\}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\rho(h_i, \hat{h}_i)}{MDE_m},$$

$$MDE_m = \frac{1}{m-1} \sum_{i=2}^{m} \rho(h_i, h_{i-1}).$$

7.3. Прогноз временных рядов распределений

Опишем использование метода $k - NN$ для прогноза временных рядов.

1. Временной ряд $\{X_t\}, t = 1, \ldots, n$ представим в виде последовательности векторов размерности d:

$$X_t^{d,\tau} = \left(X_t, X_{t-\tau}, \ldots, X_{t-(d-1)\tau}\right), \tag{7.3}$$

где $d, \tau \in N$ с d — числом задержек и параметра задержки τ.

Если $\tau = 1$, как принято во многих случаях, временной ряд векторов обозначен $\left\{X_t^d\right\}, t = 1, \ldots, n$:

$$X_t^d = \left(X_t, X_{t-1}, \ldots, X_{t-(d-1)}\right), \tag{7.4}$$

где d — вектор последовательных наблюдений, которые могут быть представлены как точка в пространстве размерности d.

2. Вычисляется расстояние между последним вектором

$$X_n^d = (X_n, X_{n-1}, \ldots, X_{n-d+1})$$

и каждым вектором $\left\{X_t^d\right\}$, где $t = d, \ldots, n-1$. Далее отбираем k ближайших к X_n^d. Обозначим их $X_{T_1}^d, X_{T_2}^d, \ldots, X_{T_k}^d$. Обычно на этом шаге используется евклидово расстояние.

3. Полученные k ближайших векторов $X_{T_1}^d, X_{T_2}^d, \ldots, X_{T_k}^d$, их последующие значения $X_{T_1+1}, X_{T_2+1}, \ldots, X_{T_k+1}$ усредняем, чтобы получить прогноз X_{n+1}.

Это самый простой $k-NN$, но могут быть предложены и более сложные версии. Например, Meade (2002) использует геометрически взвешенное евклидово расстояние, чтобы поместить больший акцент в подобие между более свежими наблюдениями, и получает прогноз как взвешенное среднее значение, которое назначает больше веса самым близким соседям. Кроме того, больше внимания можно уделить выбору значений k, d.

Адаптация метода $k - NN$

Адаптация метода $k - NN$: чтобы иметь дело с BPP, необходимо выбрать метрики для пространства кусочно-полиномиальных функций. Эти метрики будут использоваться для измерения несходства между кусочно-полиномиальными функциями, для построения прогнозов и измерения ошибок прогноза. Поскольку эти три проблемы близко связаны, рекомендуется использовать то же самое расстояние для всех случаев.

Ранее была рассмотрена метрика, подходящая для измерения ошибок в BPP и т. п.

Рассмотрим ГВР $\{h_t\}$ с $t = 1, \ldots, n$ и построим ряд векторов гистограмм размерности d:

$$h_t^d = (h_t, h_{t-1}, \ldots, h_{t-d+1}) \tag{7.5}$$

с $t = d, \ldots, n$. Расстояние между последним вектором h_n^d и всеми векторами h_t^d, с $t = d, \ldots, n-1$ вычислим как

$$D\left(h_n^d, h_t^d\right) = \frac{1}{d} \sum_{i=1}^{d} D\left(h_{n-i+1}, h_{t-i+1}\right). \tag{7.6}$$

Далее находим k самых близких элементов к h_n^d и обозначим их $h_{T_p}^d$ с $p = 1, \ldots, k$.

Построение прогноза

В методе $k - NN$ прогнозы обычно вычисляются как среднее число k соседних последовательностей. Чтобы приспособить метод $k - NN$ к ВРР, процедуру усреднения можно заменить построением центра тяжести гистограмм. Прогноз $\hat{h}_{n+1}$ будет решением задачи

$$\arg\min_{\hat{h}_{n+1}} \sum_{p=1}^{k} w_p D\left(\hat{h}_{n+1}, h_{T_p+1}\right). \tag{7.7}$$

7.4. Методы сглаживания для временных рядов распределений

В разделе представлены методы сглаживания для временных рядов распределений. Основные идеи изложены в работе [55], они базируются на понятии барицентрической гистограммы, которая представляет операцию осреднения в пространстве гистограмм.

С появлением очень больших наборов данных существует острая необходимость в разработке новых статистических подходов. Символический анализ данных, проведенный L. Billard и E. Diday [59], дает новые методики в этом направлении. Основная идея заключается в том, что данные могут быть организованы в объекты, которые обеспечивают информацию за пределами простых скалярных переменных. Интервальные данные, гистограммы, частотные полигоны, сплайны являются примерами символических данных; они обеспечивают графическое представление

для описания размаха, как в случае интервалов, или более информативное представление в случае гистограммы: дисперсия и форма распределения частот реализаций случайной величины.

Гистограмма представлена как символический объект. Это означает, что некоторый случайный эксперимент производит реализацию случайной величины, которая наблюдается как частотные распределения, или, другими словами, гистограмма. Многие авторы адаптировали методы для анализа множеств гистограмм, например анализ главных компонент [122] и кластерный анализ [137].

С точки зрения временного ряда можно определить гистограммные временные ряды (Histogram-Valued Time Series, (HTS)) как совокупность гистограмм, упорядоченных по времени.

Классические подходы в моделировании потребуют оценки зависимости данных и поиск наиболее подходящей модели. К сожалению, состояние моделей для ГВР еще не достигло завершающей стадии. Тем не менее, отсутствие совершенных моделей для построения прогноза не должно ограничивать исследования.

В работе [55] прогнозирование основано на определении средних значений информации временного ряда. В работе [56] адаптирован (k – NN)-алгоритм, и в [54] применили этот подход для прогнозирования финансовых показателей. В этом параграфе представлен анализ методов сглаживания для прогнозирования ГВР.

При работе с функциями плотности вероятности, представленными кусочно-полиномиальными моделями (гистограммы, частотные полигоны, сплайны), основной вопрос любого механизма сглаживания — арифметические операции. Учитывая, что вероятностные арифметики в их прямом применении часто не дают нужного результата, в работе [55] предлагается альтернативный подход, основанный на понятии барицентрической гистограммы, которая является «центром тяжести» гистограмм. Барицентр минимизирует сумму расстояний между некоторой гистограммой и всеми остальными в наборе. Для этих целей используют метрики: Mallows и Wasserstein. Обе метрики оценивают расстояния между квантилями любых двух гистограмм, но Mallows использует норму L_2, а Wasserstein использует норму L_1. Основываясь на барицентрическом подходе, можно представить методы сглаживания либо как скользящие средние, либо как экспоненциальное сглаживание. Показано, что только барицентр Mallows сохраняет естественную гладкость, которая является сутью любого механизма сглаживания.

Поскольку наш интерес заключается в прогнозировании, понятие стохастического процесса и гистограммного временного ряда определены в [54].

Стохастический процесс в пространстве гистограмм представляет собой совокупность гистограммых случайных переменных, которые индексируются по времени, т. е. h_{Xt} для $t \in T \subset R$ с каждым h_{Xt}.

Гистограммный временной ряд (HTS)— реализация стохастического гистограммого процесса, и он будет эквивалентно обозначен как $\{h_t, t = 1, ..., T\}$.

Методы сглаживания

Для заданного временного ряда гистограммы (ГВР) наша цель состоит в формировании прогноза h_{t+1} на один шаг вперед по доступной информации до времени T. Рассмотрим адаптацию скользящего среднего и экспоненциальное сглаживание для ГВР [55].

Прогноз ГВР на один шаг вперед на основе скользящего среднего q-го порядка — взвешенное (или невзвешенное) среднее из q прошлых наблюдений гистограмм [55]

$$\hat{h}_{t+1} = \omega_1 h_t + \ldots + \omega_q h_{t-q}, \tag{7.8}$$

где ω_i — веса.

Экспоненциальное сглаживание

$$\hat{h}_{t+1} = \alpha h_t + (1 - \alpha)\hat{h}_t, \tag{7.9}$$

где $\alpha \in [0, 1]$.

В (7.8), (7.9) ключевая операция «сумма» функций плотности вероятности, которая необходима для нахождения «средней функции плотности вероятности», которая является основой любой процедуры сглаживания.

Без ограничения общности давайте рассмотрим простое среднее двух функций плотности вероятности, такое как

$$h_{\bar{X}} = \frac{h_{X1} + h_{X2}}{2}. \tag{7.10}$$

Рассмотрим среднее двух случайных величин, имеющих треугольные функции плотности вероятности (тонкие линии) с носителями [0,1] и вершинами в точках (0,2) и (1,2) (рис. 7.2).

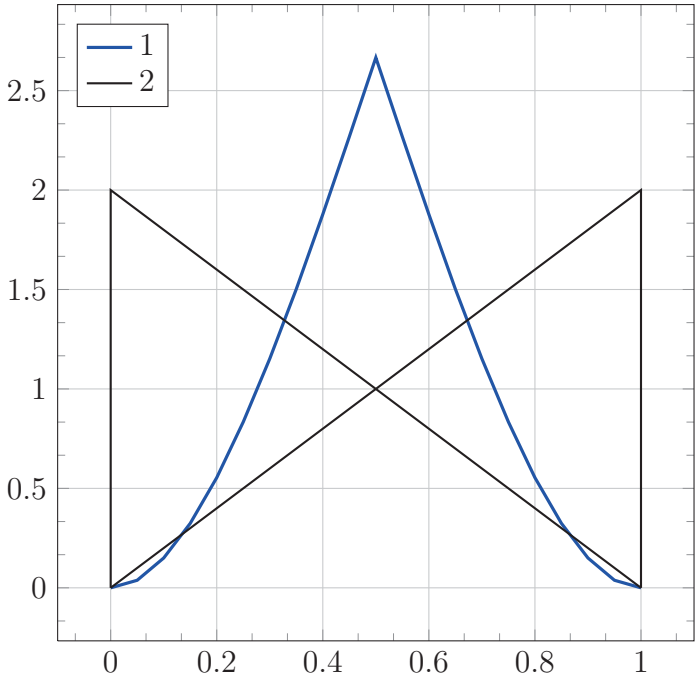


Рис. 7.2. Среднее двух случайных величин

На рис. 7.2 представлены две функции плотности вероятности и их средняя $h_{\bar{X}}$. Средняя функция плотности вероятности имеет большую массу в центре. По этим причинам нам нужно искать альтернативные подходы к вычислению среднего функций плотности вероятности.

Барицентрическая гистограмма

Барицентр набора гистограмм был предложен Irpino и Verde в работе [137] в контексте кластеризации, где барицентр гистограммы представляет центроид скопления гистограммных данных.

Барицентрическая гистограмма h_{XB} определяется как гистограмма, которая минимизирует расстояние между собой и набором гистограмм h_{Xi} для $i = 1, ..., n$,

$$\min_{h_{XB}} \left(\sum_{i=1}^{n} \omega_i D^p(h_{xi}, h_{xB}) \right)^{1/p},$$

где ω_i — вес, связанный с гистограммой h_{Xi}, такой, что $\omega_i \geq 0$ и $\sum \omega_i = 1$; p — положительное целое число, а $D(h_{Xi}, h_{XB})$ — метрика. Барицентрическая гистограмма hX_B также понимается как выпуклая рассматриваемая комбинация n гистограмм hX_i.

Verde and Irpino в работе [138] проанализировали различные меры

расстояния. Только расстояние Mallows ($p = 2$) подходит для построения барицентров. В работе [56] предложен барицентр с использованием расстояния Wasserstein ($p = 1$) для построения прогноза гистограммы на основе k-NN алгоритма.

Расстояние между двумя гистограммами h_1 и h_2 определяется также как расстояния Wasserstein и Mallows для обратных функций распределения. Пусть $f(x)$ и $g(x)$ — функции плотности вероятности, тогда представим метрики

$$\rho_W(f, g) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt, \tag{7.11}$$

$$\rho_M(f, g) = (\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt)^{1/2}, \tag{7.12}$$

где $F^{-1}(t)$, $G^{-1}(t)$ — обратные функции к функции распределения. Тогда расстояние Mallows $D_M(h_1, h_2) = \rho_M(h_1, h_2)$.

Теорема 13 ([55]). *Пусть $D^p(h_x, h_{xB})$ — расстояние Mallows, а $p = 2$. Барицентрическая гистограмма h_{xB}, решающая задачу оптимизации*

$$\min_{h_{xB}} \left(\sum_{i=1}^{n} \omega_i \int_0^1 (H_i^{-1}(t) - H_B^{-1}(t))^2 dt \right)^{1/2} \tag{7.13}$$

имеет квантильную функцию, которая удовлетворяет

$$H_B^{-1}(t) = \sum_{i=1}^{n} \omega_i H_i^{-1}(t), \ t \in [0, 1]. \tag{7.14}$$

На рис. 7.3 показано построение барицентра в пространстве функций распределения. Тонкие линии 2 — функции распределения, построенные по треугольным функциям плотности вероятности (см. рис. 7.2). Линия 1 — функция распределения барицентра Mallows. На рис. 7.4 показана плотность вероятности барицентра Mallows.

Барицентр и скользящее среднее

Учитывая связь между понятиями среднего и барицентра, прогноз ГВР $\hat{h}_{X_{t+1}}$ на основе скользящего среднего, как в формуле (7.8), может

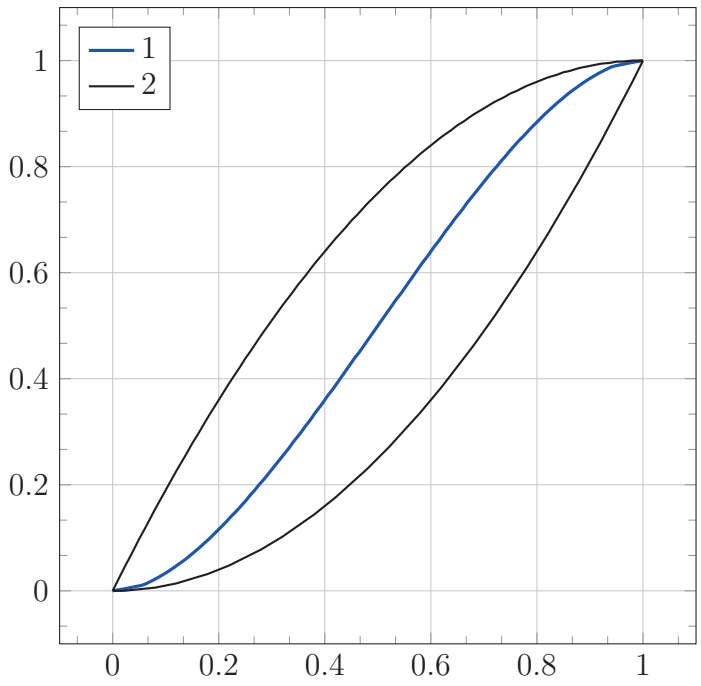


Рис. 7.3. Построение барицентра в пространстве функций распределения

быть переформулирован как барицентрическая гистограмма, которая решает следующую задачу оптимизации:

$$\min_{\hat{h}_{X_{t+1}}} \sum_{i=1}^{q} \left(\omega_i D^p(\hat{h}_{X_{t+1}}, h_{X_{t-i+1}}) \right)^{1/p},$$

где q — порядок скользящего среднего, D — подходящее расстояние и ω_i — веса, связанные с $h_{X_{t-i+1}}$.

Если D — расстояние Mallows и $p = 2$, то $\hat{h}_{X_{t+1}}$ будет таким, что

$$\hat{H}^{-1}_{t+1}(l) = \sum_{i=1}^{q} \omega_i H^{-1}_{t-i+1}(l), l \in [0, 1].$$

Барицентр и экспоненциальное сглаживание

Следуя тем же аргументам, что и в предыдущем разделе, прогноз $\hat{h}_{t+1}$ на основе фильтра экспоненциального сглаживания, как в формуле (7.9), может быть представлен как следующая задача оптимизации

$$\min_{\hat{h}_{t+1}} \left(\alpha D^p(\hat{h}_{t+1}, h_t) + (1 - \alpha) D^p(\hat{h}_{t+1}, \hat{h}_t) \right).$$

Метрика Mallows и $p = 2$ в этом случае также предпочтительней.

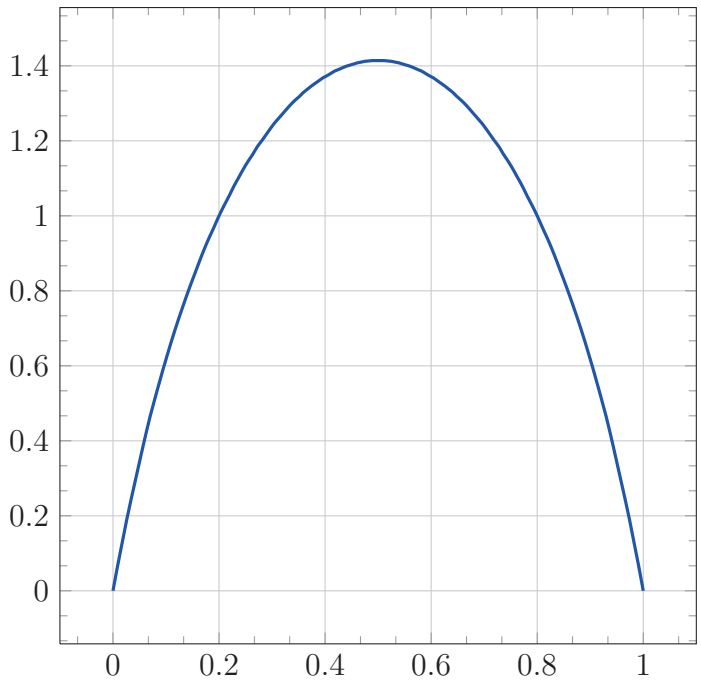


Рис. 7.4. Функция плотности вероятности барицентра Mallows

В разделе показана возможность использования метрик Mallows и $p = 2$ для прогноза гистограммных временных рядов как в методе скользящего среднего, так и экспоненциального сглаживания.

Операции среднего над функциями плотности вероятности

Опишем операции над функциями плотности вероятности на примере построения среднего двух треугольных функций плотности вероятности. Заметим, что вершина первой треугольной функции плотности вероятности f_1 имеет координаты (0,1), второй f_2 — (1,1). Естественно положить, что результат операции осреднения

$$f_s = tf_1 + (1-t)f_2, \quad t \in (0,1)$$

— функция плотности вероятности f_s должна иметь треугольный вид с вершиной в точке $t \cdot 0 + (1-t) \cdot 1$. Результаты операций осреднения для $t = 0.1, 0.2, \ldots, 0.9$ показаны на рис. 7.5 линиями 1, линии 2 — границы носителей и координаты вершин.

7.5. Метод расщепления

Далее рассмотрим подход построения прогноза временного ряда распределений (ВВР), основанного на расщеплении ряда распределений на

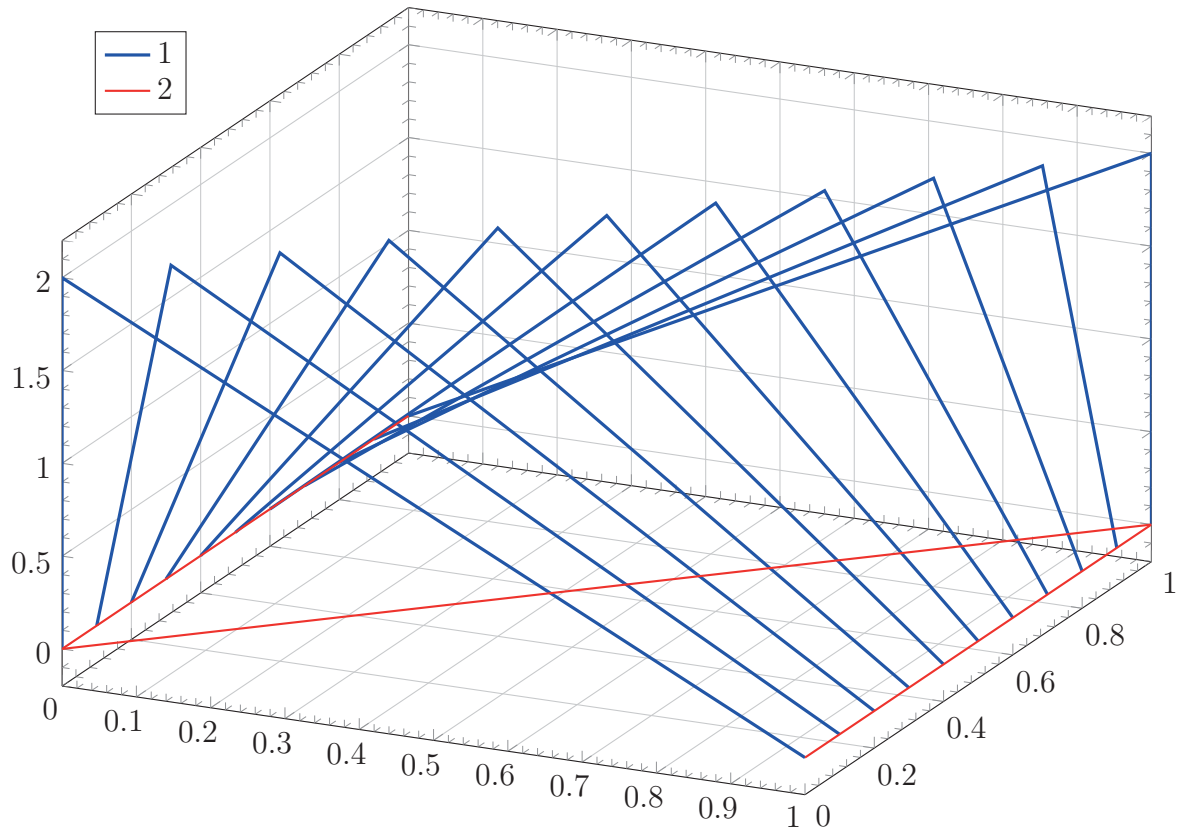


Рис. 7.5. Операции осреднения для треугольных функций плотности вероятности

подряды. Функции распределений будем представлять в виде кусочно-полиномиальных функций: кусочно-постоянных (гистограмм), кусочно-линейных (частотных полигонов) и сплайнов. Рассмотрим ВВР h_i, $i = 1, ..., N$. Далее пусть каждая кусочно-полиномиальная функция определяется сеткой $\omega^i = \{z_i^j, j = 0, .., k\}$ и значениями $p_i^j,\ j = 1, .., k$. Рассмотрим вспомогательные временные ряды

$$z_i^0, z_i^k, p_i^j, j = 1, ..., k, i = 1, ..., N. \tag{7.15}$$

Для построения прогноза $\hat{h}_{N+1}$ будем использовать последние d значений рядов (7.15). Таким образом, используя один из методов построения прогноза, построим значения $\hat{z}_{N+1}^0, \hat{z}_{N+1}^k, \hat{p}_{N+1}^j$. Для кусочно-полиномиальной функции $\hat{h}_{N+1}$ определим равномерную сетку $\omega^{N+1} = \{\hat{z}_{N+1}^j, j = 0, .., k\}$ и положим на ней значения $\hat{p}_{N+1}^j$. Последним шагом в построении прогноза будет нормировка кусочно-полиномиальной функции $\hat{h}_{N+1}$

$$\hat{h}_{N+1}^n = \frac{1}{\beta}\hat{h}_{N+1},$$

$$\beta = \int_{z_{N+1}^0}^{z_{N+1}^k} \hat{h}_{N+1}^{nor}(\xi)d\xi,$$

таким образом, чтобы выполнялось соотношение

$$\int_{z_{N+1}^0}^{z_{N+1}^k} \hat{h}_{N+1}^{nor}(\xi)d\xi = 1.$$

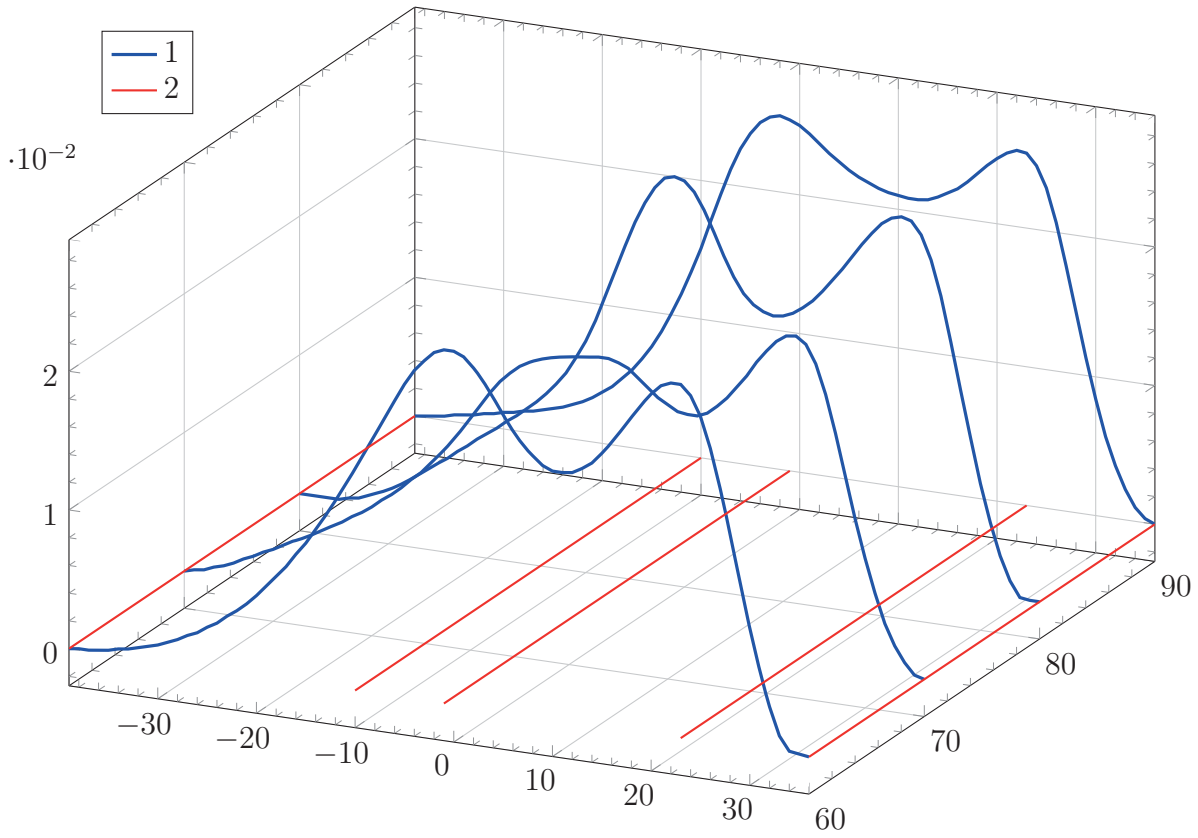


Рис. 7.6. Распределения температуры с 1960 по 1990 год

7.6. Численный пример

На рис. 7.6 приведены кусочно-полиномиальные аппроксимации функций плотности вероятности максимальной температуры за 1960, 1970, 1980 и 1990 годы. Для аппроксимаций использовались эрмитовы сплайны пятой степени. Эти сплайны $s(x)$ определяются в узлах сетки x_i значениями f_i, f_i', f_i''

$$s(x) = \sum_i f_i\varphi((x-x_i)/h_i) + h_i f_i'\varphi_1((x-x_i)/h_i) + h_i^2 f_i''\varphi_2((x-x_i)/h_i),$$

$h_i = x_{i+1} - x_i,$

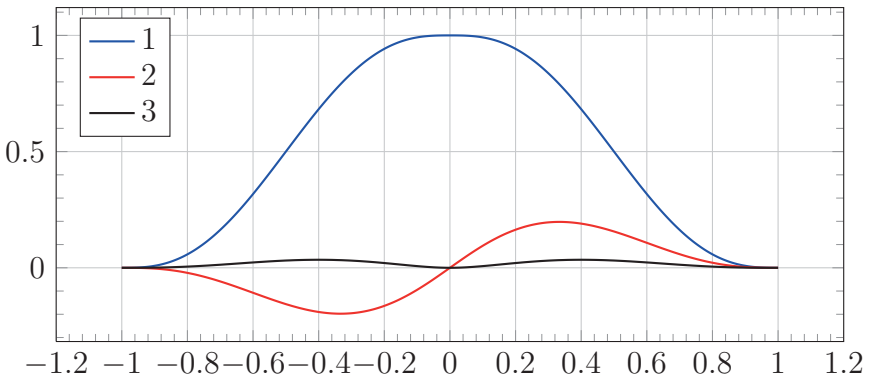


Рис. 7.7. Базисные функции эрмитового сплайна пятой степени

где $\varphi_0, \varphi_1, \varphi_2$ — базисные функции, такие что $\varphi_i^{(\nu)}(x) = 0, \nu \in \{0,1,2\}$, $x \in \{-1,0,1\}$. Исключение составляют

$$\varphi_0(0) = 1; \varphi_1'(0) = 1; \varphi_2''(0) = 1.$$

Таким образом, решая в пространстве эрмитовых полиномов пятой степени, получаем

$$\varphi_0(x) = \begin{cases} -(x-1)^3(6x^2+3x+1), & x \geq 0; \\ (x+1)^3(6x^2-3x+1), & x \leq 0. \end{cases}$$

$$\varphi_1(x) = \begin{cases} -x(x-1)^3(3x+1), & x \geq 0; \\ -x(x+1)^3(3x-1), & x \leq 0. \end{cases}$$

$$\varphi_2(x) = \begin{cases} -x^2(x-1)^3, & x \geq 0; \\ x^2(x+1)^3, & x \leq 0. \end{cases}$$

На рис. 7.7 показаны базисные функции эрмитового сплайна пятой степени.

Для построения сплайновой аппроксимации функции плотности вероятности f на первом этапе использовались ядерные оценки с параметром $h = 1$. Приближение $\hat{f}_l$ функции плотности вероятности строилось в узлах сетки $\{x_l | l = 0, \ldots, N\}$. Не ограничивая общности, будем считать, что носитель $\mathrm{supp}(f) = (x_0, x_N)$. Краевые условия $s(x_0) = 0$, $s'(x_0) = 0$, $s(x_N) = 0$, $s'(x_N) = 0$. Для построения сплайна зададимся сеткой $\{\xi_i | \xi_i \in (x_0, x_N), i = 0, \ldots, n\}$, неизвестные значения f_i, f_i', f_i'' найдем, используя метод наименьших квадратов

$$\sum_l^N (s(x_l) - \hat{f}_l)^2 \to \min.$$

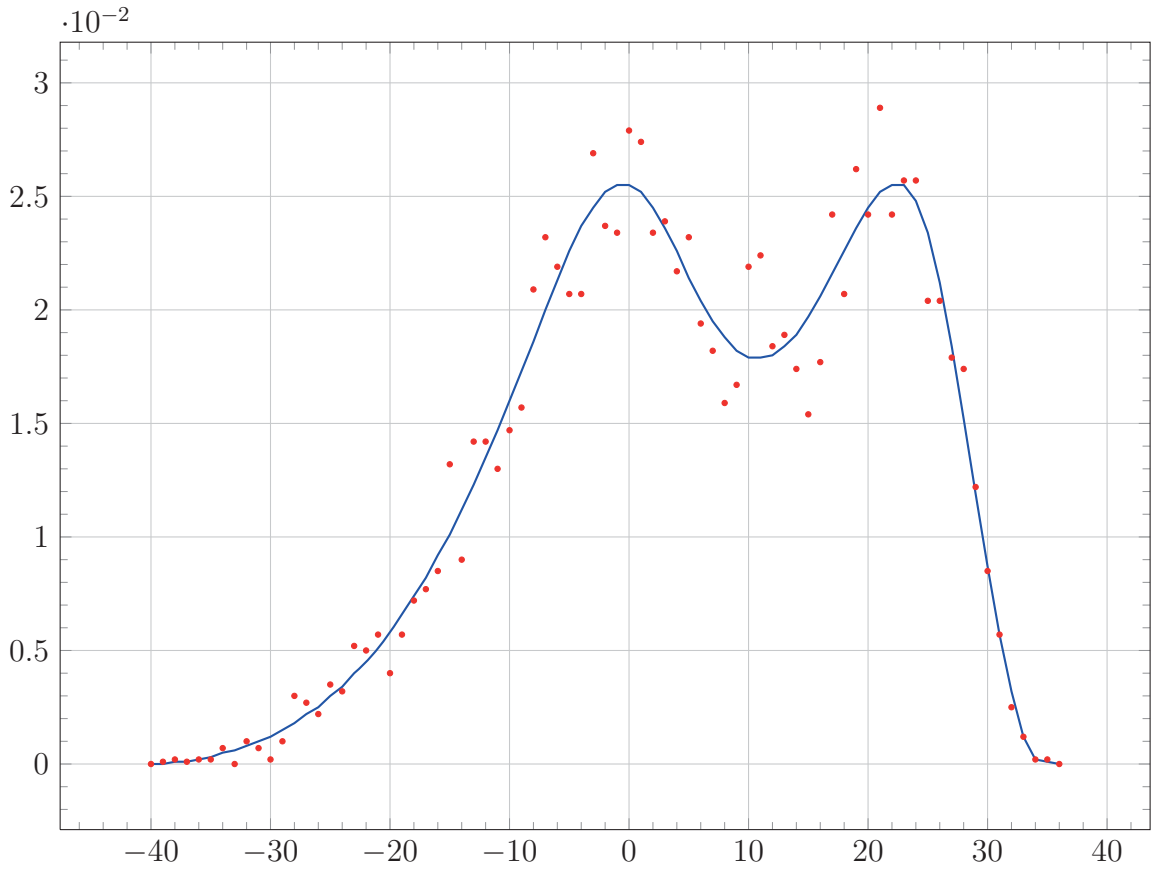


Рис. 7.8. Построение сплайновой аппроксимации

На рис. 7.8 показано построение сплайновой аппроксимации функции плотности вероятности температуры за 1960 год. Точки — ядерные оценки функции плотности вероятности, сплошная линия — эрмитов сплайн пятой степени s, узлы сетки сплайна $\omega = \{-40, -10, -1, 23, 36\}$, краевые условия

$$s(-40) = 0, s(36) = 0,$$
$$s'(-40) = 0, s'(36) = 0.$$

Таким образом, зная значения f_i, f'_i, f''_i построенных сплайнов в узлах сетки $\omega = \{-40, -10, -1, 23, 36\}$, используя метод расщепления, можно построить прогноз плотностей вероятности.

Приведенные исследования использования временных рядов распределений показали перспективность развития данного направления. Временные ряды распределений в дальнейшем предполагается использовать в тех областях, где обычные временные ряды недостаточно полно описывают происходящие процессы, не эффективны и требуют больших вычислительных затрат. Такие процессы возникают при обработке информации зондирования Земли, в прогнозах гидрологических рядов, эконо-

метрике.

Дальнейшее перспективное направление — исследование в областях прогнозирования плотности (density forecast).

Глава 8

Случайное программирование

Многие практические проблемы для своего решения требуют применения оптимизационного подхода. Задачи оптимизации в конечномерных пространствах могут быть охарактеризованы некоторым числом фиксированных входных параметров, которые определяют структуру решаемой задачи [29]. Например, такими фиксированными параметрами являются коэффициенты целевой функции, матрицы ограничений и правых частей ограничений. Решение подобных оптимизационных задач состоит в поиске оптимального решения для заданных входных фиксированных параметров в условиях имеющихся ограничений.

Отметим, что для многих практических задач имеет место ситуация, когда параметры моделей оптимального программирования точно не известны, имеют случайный характер или оцениваются на основе экспертных методов. Для целей моделирования часто берутся средние оценки коэффициентов, а далее получается некоторое решение, оптимальное в данной модели, но которое не всегда является оптимальным в исходной задаче.

Существующие подходы стохастического программирования помогают исследователям раскрыть природу неопределенности во входных данных и учесть ее в модели. В то же время имеется ряд сложностей. Например, связанных с численным преобразованием задачи линейного стохастического программирования в детерминированную задачу нелинейного программирования. Хорошо известен тот факт, что алгоритмы нелинейного программирования на практике применимы к задачам с относительно небольшой размерностью.

Важно заметить, что после представления имеющихся неопределенностей формируется проблема выбора метода, который позволит осуществить последующие расчеты таким образом, чтобы получить реальные

результаты, с тем, чтобы не получить дополнительные неопределенности [125].

В настоящее время с этой целью развивается математический аппарат неопределенного программирования. Неопределенное программирование представляет собой теоретические основы решения оптимизационных задач в условиях различных видов неопределенности. В работах [95] известный специалист в области теории и практики решения оптимизационных задач в условиях неопределенности выделяет три основных вида неопределенности: случайность, нечеткость и неточность. В данном направлении следует отметить разработку методов решения задач оптимизации с использованием интервального анализа [28, 52, 47]; гибридных алгоритмов, совмещающих в себе идеи и подходы статистического моделирования, нейронные сети, генетические алгоритмы, имитационный отжиг и табу-поиск [95].

Важно отметить, что в большинстве алгоритмов, представленных в рамках указанных направлений, применяется оператор математического ожидания, или проводятся процедуры усреднения, или требуется знание законов распределений.

В работах [81, 94, 115] ставились задачи учета влияния имеющейся неопределенности и оценки функций плотности вероятности решений и целевой функции. Определенный прогресс в этой области был получен в шестидесятые и семидесятые годы XX века и связан с именами J. B. Ewbank, P. Kall, A. Prékopa. Следует отметить что при этом накладывались существенные ограничения на вид входных данных.

В тех случаях, когда известны функции плотности вероятности для стохастических данных, можно с успехом использовать метод Монте-Карло [94].

В данной главе рассматривается новый подход к решению оптимизационных задач со случайными входными параметрами, который определяется как случайное программирование. Данный подход использует вычислительный вероятностный анализ и позволяет строить множество решений оптимизационной задачи на основе совместной функции плотности вероятности [113, 69].

В отличие от стохастического программирования [129], где оптимальным решением является фиксированное решение, этот подход позволяет нам построить целый набор решений задачи оптимизации, который описывается совместной функцией плотности вероятности.

В формулировке задач неопределенной оптимизации обычно пыта-

ются найти хороший компромисс между реалистичностью модели оптимизации и ее позволительной способностью в выборе и дальнейшем использовании соответствующего численного метода решения изучаемой проблемы. Эти две составляющие в совокупности обычно влияют на полезность и качество получаемых решений. В результате этих соображений существует большое количество различных подходов к постановке и решению задач оптимизации в условиях неопределенности. Остановимся на стохастическом подходе к решению оптимизационных задач.

Рассмотрим общую постановку задачи стохастического программирования (Stochastic Programming Problem (SPP)) [129]:

$$\max f(x, \xi),$$

$$g_i(x, \xi) \leq 0, \quad i = 1, ..., p,$$

где x — вектор решения; ξ — случайный вектор; $f(x, \xi)$ — целевая функция; $g_i(x, \xi)$ — случайные функции ограничений.

С целью применения соответствующих подходов к решению оптимизационных задач в условиях стохастической неопределенности общая задача SPP может быть сформулирована в M и P постановках по отношению к записи целевой функции и ограничений. M-постановка означает оптимизацию математического ожидания целевой функции и представляет собой первый тип постановки задачи стохастического программирования. В зарубежной литературе такие задачи называются моделями ожидаемого значения, или expected model value (EMV) [95]. M-постановка имеет вид

$$\max \mathsf{M}[f(x, \xi)], \tag{8.1}$$

$$\mathsf{M}[g_i(x, \xi)] \leq 0, \quad i = 1, ..., p. \tag{8.2}$$

Во многих случаях задача стохастической оптимизации может ставиться как многокритериальная. В этом случае имеет место многокритериальное стохастическое программирование.

Существуют два основных подхода к решению задач стохастического программирования:

1) непрямые методы, которые заключаются в нахождении функций $F(x)$, G_i и решении эквивалентной задачи нелинейного программирования вида (8.1), (8.2);

2) прямые методы стохастического программирования, основанные на информации о значении функций $f(x, \xi)$, $g_i(x, \xi)$, получаемой в результате проведения экспериментов.

Следует указать на относительно новые постановки оптимизационных задач в условиях интервальной неопределенности. Например, задача линейного программирования с интервальными данными формулируется следующим образом [47]:

$$(c, x) \to \min, \tag{8.3}$$

$$Ax = b, x \geq 0. \tag{8.4}$$

$$A \in \boldsymbol{A}, b \in \boldsymbol{b}, c \in \boldsymbol{c}, \tag{8.5}$$

где $\boldsymbol{A}$ — интервальная матрица; $\boldsymbol{b}$, $\boldsymbol{c}$ — интервальные векторы размерности n.

8.1. Постановка задачи

Сформулируем задачу случайного программирования в следующем виде:

$$f(x, \xi) \to \min, \tag{8.6}$$

$$g_i(x, \xi) \leq 0, \quad i = 1, ..., m, \tag{8.7}$$

где x — вектор решения; $f(x, \xi)$ — целевая функция; $g_i(x, \xi)$ — функции ограничений; ξ — случайный вектор параметров. Вектор ξ формализуем через пространство вероятностей, $(\Omega, \mathcal{F}, P)$, где Ω, $\mathcal{F}$, P — множество случайных событий, σ-алгебра подмножеств Ω и соответствующая вероятностная мера. Элементы этого вероятностного пространства появляются как параметры во входном случайном векторе $\xi(\omega), \omega \in \Omega$.

$$f(x^*, \xi(\omega)) = \inf_U f(x, \xi(\omega)),$$

где

$$U = \{x | g_i(x, \xi(\omega)) \leq 0, \quad \omega \in \Omega, i = 1, ..., m\}.$$

Множество решений (8.6)–(8.7) определим следующим образом:

$$\mathcal{X} = \{x | \min f(x, \xi(\omega)), g_i(x, \xi) \leq 0, \quad i = 1, ..., m, \omega \in \Omega\}.$$

Заметим, что x^* — случайный вектор, поэтому в отличие от детерминированной задачи для x^* необходимо определять функции плотности вероятности для каждой компоненты x_i^* как совместную плотность вероятности.

Задача линейного программирования со случайными данными формулируется следующим образом:

$$\min(c(\omega), x), \tag{8.8}$$

$$A(\omega)x = b(\omega), x \geq 0, \omega \in \Omega. \tag{8.9}$$

Точка x^* — решение задачи (8.8),(8.9), если

$$(c(\omega), x^*(\omega)) = \inf_{U(\omega)} (c(\omega), x),$$

где

$$U(\omega) = \{x | A(\omega)x = b(\omega), x \geq 0, \omega \in \Omega\}.$$

Множество решений (8.8),(8.9)

$$\mathcal{X} = \{x | \min(c(\omega), x), A(\omega)x = b(\omega), x \geq 0, \omega \in \Omega\}.$$

8.2. Случайное линейное программирование

Известно, что для задачи (8.8), (8.9) оптимальное решение x^* достигается в угловой точке множества U [5].

Теорема 14 ([5]). *Пусть множество U определено условиями (8.9). Для того чтобы точка $x = (x_1, ..., x_n) \in U$ была угловой необходимо и достаточно, чтобы существовали номера $j_1, ...j_r$:*

$$A_{j_1}x_{j_1} + ... + A_{j_r}x_{j_r} = b; x_j = 0, j \neq j_l, l = 1, ..., r,$$

причем столбцы $A_{j_1}, ..., A_{j_r}$ линейно независимы.

Пример 11. Пусть U определяется матрицей A и вектором b:

$$A = \begin{pmatrix} 1 & 1 & 3 & 1 \\ 1 & -1 & 1 & 2 \end{pmatrix}, b = \begin{pmatrix} 3 \\ 1 \end{pmatrix},$$

тогда столбцам матрицы A_1, A_2 соответствует угловая точка с координатами $(2, 1, 0, 0)$, A_1, A_3 — $(0, 0, 1, 0)$, A_2, A_4 — $(0, 5/7, 0, 4/3)$.

Заметим, что из n столбцов можно выбрать r линейно независимых не более чем C_n^r способами. Следовательно, число угловых точек множества U конечно.

Это значит, что каноническую задачу (8.8)–(8.9) можно попытаться решить следующим образом:

1) найти все угловые точки x множества U;

2) вычислить значение функции (c, x) в каждой из угловых точек и определить наименьшее из них.

Однако такой подход практически не применяется, так как даже в задачах не очень большой размерности число угловых точек может быть очень большим. Тем не менее идея перебора угловых точек множества оказалась весьма плодотворной и послужила основой ряда методов решения задач линейного программирования. Одним из таких методов является так называемый симплекс-метод.

Для задачи (8.8)–(8.9) построим совместную плотность вероятности вектора x^*. Для этой цели воспользуемся одним из способов решения детерминированных задач линейного программирования, например симплекс-методом.

Рассмотрим вспомогательную задачу. Пусть A_j, b_j, c_j — реализации $A_j = A(\omega_j)$, $b_j = b(\omega_j)$, $c_j = c(\omega_j)$, $\omega_j \in \Omega$.

$$\min (c_j, x), \tag{8.10}$$

$$A_j x = b_j, x \geq 0. \tag{8.11}$$

Найдем решение x_t^* и соответствующую ему угловую точку с номерами $j_1, ...j_r$.

Решим систему линейных алгебраических уравнений

$$(\boldsymbol{A}_{j_1} ... \boldsymbol{A}_{j_r})\boldsymbol{x} = \boldsymbol{b}.$$

Для этих целей мы можем воспользоваться методами, описанными в главе 5, в частности, использованием вероятностных расширений и численных операций над кусочно-полиномиальными функциями.

Совместная плотность вероятности найденного решения будет соответствовать $\boldsymbol{x}_j^*$. Если носители входных параметров достаточно малы, то в силу непрерывности $\boldsymbol{x}_j^*$ будет совпадать с $\boldsymbol{x}^*$.

В случае произвольных носителей входных параметров процедуру выбора $A(\omega_j), b(\omega_j), c(\omega_j)$, следует повторить, используя подходы метода Монте-Карло или генетических алгоритмов. Если при этом будут получены разные решения $\boldsymbol{x}_j^*$, то их можно сравнить, вычисляя вероятностные расширения $\boldsymbol{f}_j = (\boldsymbol{c}, \boldsymbol{x}_j^*)$.

Заметим, что

$$\boldsymbol{x}^* = (\boldsymbol{A}_j)^{-1}\boldsymbol{b}$$

и выражение

$$\boldsymbol{f}_j = \boldsymbol{c}^T(\boldsymbol{A}_j)^{-1}\boldsymbol{b}$$

можно оценить, используя вероятностные расширения.

Численный пример. Рассмотрим следующую задачу:

$$(c, x) \to \min, \tag{8.12}$$

$$Ax = b, x \geq 0. \tag{8.13}$$

$$A \in \boldsymbol{A}, b \in \boldsymbol{b}, c \in \boldsymbol{c}, \tag{8.14}$$

где $\boldsymbol{A} = (\boldsymbol{a}_{ij})$ — равномерная случайная матрица, каждый элемент — равномерная случайная величина с носителем $[\underline{a}_{ij}, \overline{a}_{ij}]$, аналогично $\boldsymbol{b}, \boldsymbol{c}$ — случайные векторы с элементами в виде равномерных случайных величин.

Носители заданы следующим образом:

$$A = \begin{pmatrix} [1-r, 1+r] & [1-r, 1+r] \\ [1-r, 1+r] & [-1-r, -1+r] \end{pmatrix}$$

$$\begin{matrix} [3-r, 3+r] & [1-r, 1+r] \\ [1-r, 1+r] & [2-r, 2+r] \end{matrix} \Bigg),$$

$$b = \begin{pmatrix} [3-r, 3+r] \\ [1-r, 1+r] \end{pmatrix},$$

$$c = (-1, -1, 0, 0).$$

При $r = 0$, что соответствует детерминированному случаю, решение $x^* = (2, 1, 0, 0)$, столбцы матрицы A_1, A_2 соответствуют угловой точке.

На рис. 8.1 приведена совместная плотность вектора $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ при $r = 0.1$, компоненты $x_3 = 0, x_4 = 0$. Сплошная линия — граница множества решений на плоскости (x_1, x_2). Множество решений $\mathcal{X}$ — четырехугольник с вершинами (2.0,0.636), (2.444,1.0), (2.0,1.444), (1.636,1.0). Как видно из рис. 8.1, плотность вероятности распределена крайне неравномерно, самая большая в центре, в окрестности точки (2.0, 1.0). Подробнее методы построения совместной функции распределения решения случайных систем линейных алгебраических уравнений представлены в главе 6.

Рассмотрим оценку функции плотности вероятности целевой функции

$$\boldsymbol{f} = c^T (\boldsymbol{A})^{-1} \boldsymbol{b}.$$

Пусть

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}_{21} & \boldsymbol{a}_{22} \end{pmatrix},$$

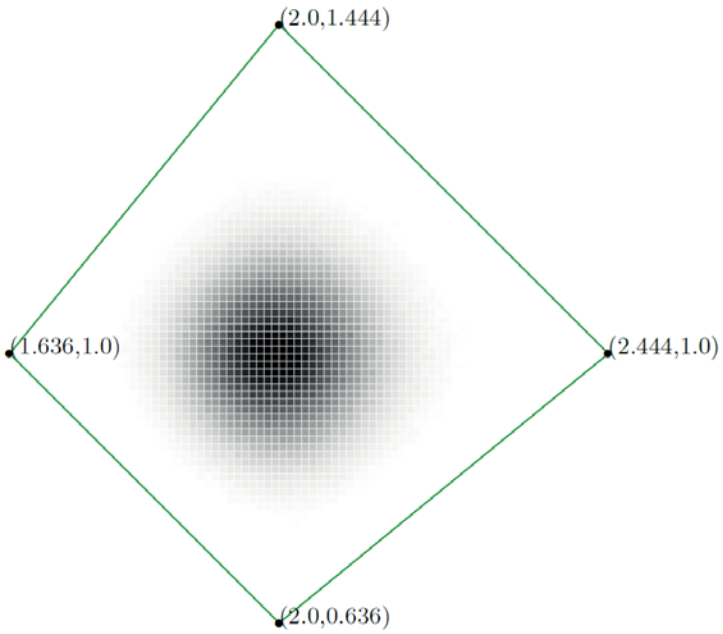


Рис. 8.1. Совместная плотность вектора $\boldsymbol{x}_1, \boldsymbol{x}_2$

тогда

$$\boldsymbol{A}^{-1} = \frac{1}{\boldsymbol{\Delta}} \begin{pmatrix} \boldsymbol{a}_{22} & -\boldsymbol{a}_{12} \\ -\boldsymbol{a}_{21} & \boldsymbol{a}_{11} \end{pmatrix},$$

где $\boldsymbol{\Delta} = \boldsymbol{a}_{11}\boldsymbol{a}_{22} - \boldsymbol{a}_{12}\boldsymbol{a}_{21}$. Целевая функция

$$\boldsymbol{f} = -\frac{1}{\boldsymbol{\Delta}} \left((\boldsymbol{a}_{22} - \boldsymbol{a}_{21})\boldsymbol{b}_1 + (\boldsymbol{a}_{11} - \boldsymbol{a}_{12})\boldsymbol{b}_2 \right)$$

и вероятностное расширение

$$\boldsymbol{f}(\xi) = -\int \boldsymbol{a}_{12}(t_{12}) \dots \boldsymbol{a}_{22}(t_{22}) \times$$

$$\times \frac{1}{\Delta(t_{11}, ..., t_{22})} \left((t_{22} - t_{21})\boldsymbol{b}_1 + (t_{11} - t_{12})\boldsymbol{b}_2 \right)(\xi) dt_{12} \dots dt_{22},$$

где $\Delta(t_{11}, ..., t_{22}) = t_{11}t_{22} - t_{12}t_{21}$. Выражение

$$\left(\frac{1}{\Delta(t_{11}, ..., t_{22})} ((t_{22} - t_{21})\boldsymbol{b}_1 + (t_{11} - t_{12})\boldsymbol{b}_2) \right)(\xi)$$

будем вычислять, используя численные вероятностные арифметики над кусочно-полиномиальными функциями.

На рис. 8.2 приведена аппроксимация сплайнами функции плотности вероятности целевой функции $c_1x_1 + c_2x_2$.

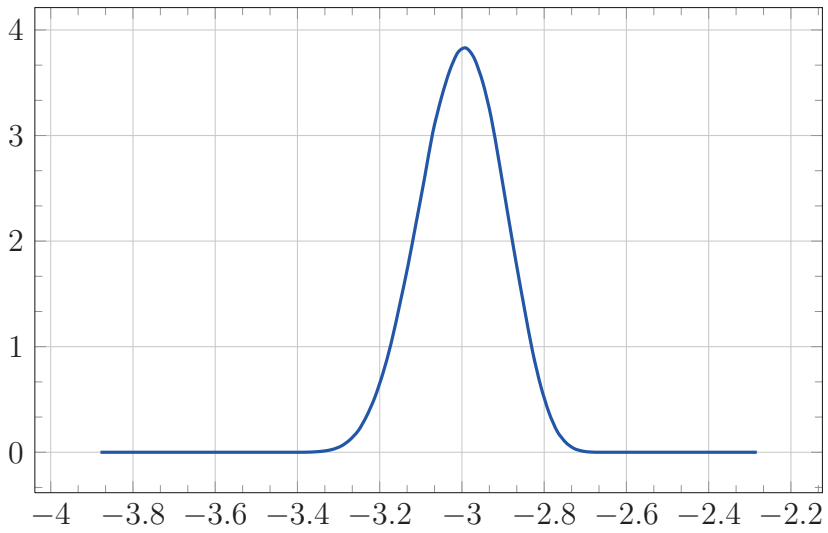


Рис. 8.2. Функция плотности вероятности случайной целевой функции (c, x^*)

Площадь $\mathcal{X}$ сильно зависит от r, с увеличением r она растет и уже при $r = 1$ становится бесконечной. Это определено тем, что среди матриц

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in \begin{pmatrix} [0,2] & [0,2] \\ [0,2] & [-2,0] \end{pmatrix}$$

есть линейно зависимые столбцы, матрица системы становится вырожденной и компоненты решения уходят на бесконечность.

8.3. Случайное нелинейное программирование

Рассмотрим задачу случайного нелинейного программирования без ограничений в следующем виде:

$$\frac{1}{2}(Ax, x) - (b, x) \to \min\,. \tag{8.15}$$

$$A \in \boldsymbol{A}, b \in \boldsymbol{b}, \tag{8.16}$$

где $\boldsymbol{A}$ — случайная матрица; $\boldsymbol{b}$ — случайный вектор.

Задача (8.15),(8.16) в случае симметричных положительно определенных матриц A сводится к решению случайной системы линейных алгебраических уравнений

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{8.17}$$

Для решения случайных систем линейных алгебраических уравнений вида (8.17) можно использовать вычислительный вероятностный анализ совместно с методом Монте-Карло.

В общем случае задачу случайного нелинейного программирования (8.6),(8.7) можно свести к решению случайной системы нелинейных уравнений

$$F(x,k)=0, \quad k\in \boldsymbol{k}, \tag{8.18}$$

где $\boldsymbol{k}$ — множество случайных векторов параметров. Как и в случае систем линейных алгебраических уравнений, следуя результатам главы 6, можно построить как совместную плотность вероятности решения, так и плотности вероятности отдельных компонент решения. В результате для задачи (8.17) или (8.18) получаем совместную плотность вероятности решения $\boldsymbol{x}$.

Численные примеры. Пусть в задаче (8.15) $\boldsymbol{A}$ — равномерная случайная матрица

$$\boldsymbol{A}=\begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \\ \boldsymbol{a}_2 & \boldsymbol{a}_1 \end{pmatrix},$$

$\boldsymbol{b}$ — равномерный случайный вектор. Носители $\boldsymbol{a}_1=[2,4]$, $\boldsymbol{a}_2=[-1,0]$, $\boldsymbol{b}_1=\boldsymbol{b}_2=[0.5,1]$.

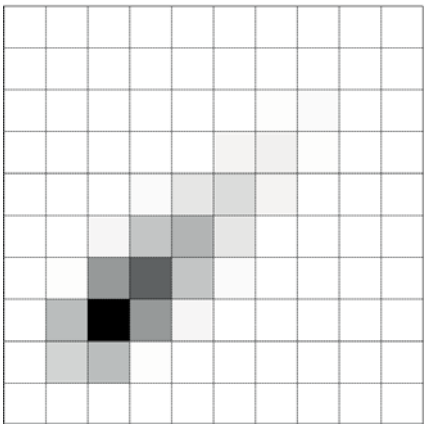


Рис. 8.3. Совместная плотность вектора $\boldsymbol{x}$

На рис. 8.3 для задачи (8.15),(8.16) приведена кусочно-постоянная аппроксимация совместной плотности вероятности вектора $\boldsymbol{x}$. Для сравнения на рис. 8.4 приведены частные решения системы (8.17), аналогичные квази Монте-Карло методу [109]. В силу определенной симметрии матрицы A частные решения системы (8.17) образуют некоторый порядок.

Добавим к задаче (8.15) ограничение в виде

$$x_1+x_2=a, \tag{8.19}$$

где $a\in \boldsymbol{a}$, $\boldsymbol{a}$ — равномерная случайная величина с носителем [0.9, 1.0]. В этом случае решение оптимизационной задачи выписывается в явном

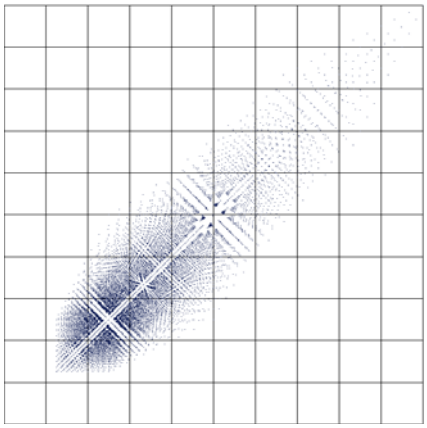


Рис. 8.4. Частные решения системы (8.17)

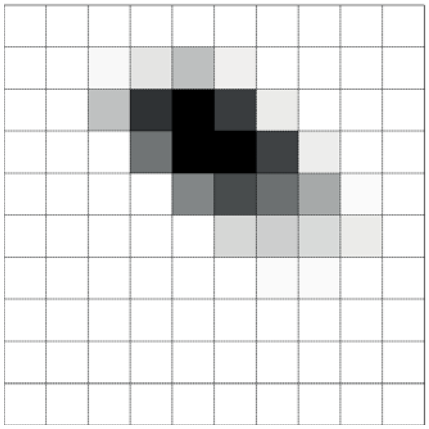


Рис. 8.5. Решение задачи с ограничениями

виде

$$x_1 = -(b_2 - b_1 - 2aa_2 - 2aa_1)/(8a_1),$$

$$x_2 := a - x_1.$$

Используя вычислительный вероятностный анализ [69], построим решение $\boldsymbol{x}$. На рис. 8.5 на квадрате $[0.7, 1.0] \times [0, 0.3]$ приведена совместная плотность вероятности вектора решения задачи (8.15) с ограничениями (8.19).

Рассмотренные методы решения приведенных задач линейной и нелинейной оптимизации позволяют представить *случайное программирование* как эффективный метод решения оптимизационных задач в условиях неопределенности входных параметров. В дальнейшем планируется разработать алгоритмы выбора наилучших оптимальных решений из построенного множества решений.

Глава 9

Регрессионный анализ

Регрессионное моделирование представляет собой способ исследования объектов на основе использования информационного подхода для выявления факта существования различных зависимостей между входными и выходными данными.

Регрессионный анализ может рассматриваться как метод моделирования измеряемых данных и исследования их свойств. Данные представляются в виде пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Регрессионный анализ предназначен для изучения по выборочным данным статистической зависимости ряда величин, некоторые из которых являются случайными. При статистической зависимости величины не связаны функционально, но как случайные величины заданы совместным распределением вероятностей.

Исследование зависимости случайных величин приводит к моделям регрессии и регрессионному анализу на базе выборочных данных. Теория вероятностей и математическая статистика представляют лишь инструмент для изучения статистической зависимости, но не ставят своей целью установление причинно-следственной связи. Числовые данные обычно имеют между собой явные (известные) или неявные (скрытые) связи. Поэтому важной задачей для исследователя является с помощью различных методов выявить скрытые зависимости и закономерности, содержащиеся в данных, и выразить их в виде формул, т. е. математически смоделировать явления или процессы. Регрессионный анализ называют основным методом современной математической статистики для выявления неявных и завуалированных связей между данными наблюдений.

В регрессионном анализе имеют место следующие допущения: коли-

чество наблюдений достаточно для проявления статистических закономерностей относительно факторов и их взаимосвязей; обрабатываемые данные содержат некоторые ошибки (помехи), обусловленные погрешностями измерений, воздействием неучтенных случайных факторов; матрица результатов наблюдений является единственной информацией об изучаемом объекте, имеющейся в распоряжении перед началом исследования.

В условиях «больших» данных предполагается целесообразным имеющиеся наблюдения или «сырые данные» подвергнуть процедуре обработки, в частности агрегации.

Построение регрессионных моделей во многом зависит от свойств эмпирической информации. В данной главе рассматривается новый подход к регрессионному моделированию на основе методов ВВА. Для представления эмпирических данных на этапе предобработки предлагается использовать кусочно-полиномиальные модели, в том числе сплайн-функции. Данное преобразование данных может рассматриваться как процедура агрегации исходной информации для подготовки данных к численному моделированию, в том числе как входные и выходные переменные регрессионной модели. В этом случае, соответственно, необходимо адаптировать понятия и методы классической регрессии к новым видам переменных.

Регрессионные модели для этого типа данных обязательно более сложны, чем простое обобщение классической регрессионной модели. Функциональные линейные соотношения между кусочно-полиномиальными переменными не могут быть простой адаптацией в рамках классического регрессионного анализа.

В этой главе предложены новые модели и методы построения и исследования линейной регрессии для подобных данных. Предлагаемый подход можно интерпретировать как построение функций распределения, а регрессионные модели на таких данных можно рассматривать как модели на эмпирических распределениях.

9.1. Регрессионные модели над эмпирическими распределениями

Регрессионный анализ данных в настоящее время является областью активных исследований, как в общем теоретическом плане, так приме-

нительно к различным областям его практического применения.

Проблема повышения точности результатов численного моделирования, снижения уровня неопределенности с учетом всех особенностей эмпирической информации является актуальной задачей и занимает важное место, в том числе и для повышения уровня доверия лица, принимающего решения, к результатам моделирования. Многие исследователи отмечают важность и необходимость анализа эмпирической информации уже на этапе предобработки данных. Обоснованно подобранные модели представления данных на этапе предобработки позволяют определить вид входных переменных и осуществить выбор соответствующих процедур и арифметик для последующего моделирования в соответствии с видом неопределенности и объемом имеющейся информации.

Идея преобразования эмпирических данных на основе применения математических моделей на этапе представления данных и последующего их использования в виде входных и выходных факторов для моделирования способствовала появлению особого вида переменных. Например, использование кусочно-полиномиальных моделей данных в виде входных переменных для регрессионного моделирования способствовало появлению нового понятия «распределенно-значные переменные», которые представляют собой особый вид переменных, где каждому такому объекту (признаку) соответствует распределение, которое может быть представлено в виде кусочно-полиномиальной функции. Такие переменные изучаются, например, в символьном анализе [58].

В последнее время наблюдается растущий интерес к моделированию и анализу интервально-значных и гистограммно-значных переменных [58, 59]. Однако анализ публикаций по данной теме исследований показал, что существующие методы и подходы к регрессионному моделированию на гистограммно-значных переменных встречают ряд трудностей. Например, для линейных моделей регрессии для этого типа данных отмечается, что ее параметры не могут быть отрицательными. Для определения параметров этой модели необходимо решить квадратичную задачу оптимизации, при условии неотрицательности ограничений на неизвестных. Определенную проблему составляет задача выбора и вычисления меры погрешности между предсказанными и наблюдаемыми распределениями.

В главе для описания случайной неопределенности во входных и выходных переменных на этапе преобразования данных предлагается использовать кусочно-полиномиальные переменные, которые представля-

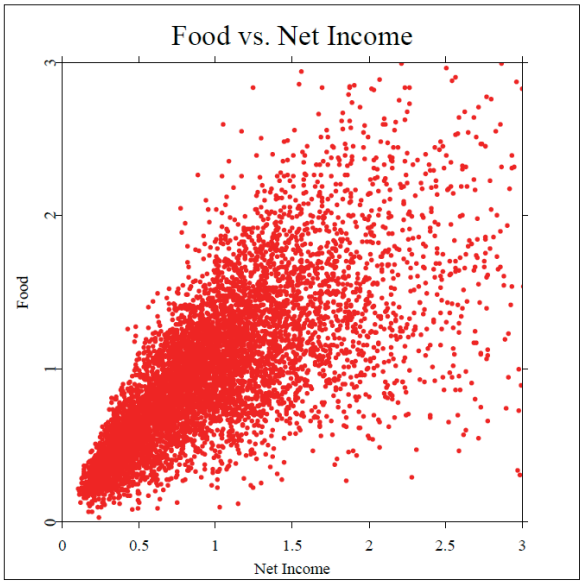


Рис. 9.1. Зависимость расходов на питание от чистого дохода. Множество точек зависимости расходов на питание Y от чистого дохода X

ют собой функции плотности вероятности соответствующих переменных, построенные по эмпирическим данным в классе кусочно-полиномиальных моделей. Для вычисления неизвестных параметров модели предлагается использовать вычислительный вероятностный анализ, в котором имеются соответствующие арифметики и процедуры.

В рамках применения данного подхода рассматриваются новые подходы моделирования функциональных зависимостей на основе аппроксимаций сплайнами [36]. Для исследования точности вычислений используется метод построения апостериорных оценок [35, 71].

Численная реализация модельных примеров регрессивного моделирования показала хорошую сходимость предложенного подхода. Использование регрессионного моделирования на основе кусочно-полиномиальных моделей открывает новые возможности в прогнозировании состояний сложных систем, дистанционного зондирования Земли, оценок надежности ответственного оборудования, оценки гидрологических, инвестиционных рисков [70].

На рис. 9.1 изображен набор данных расходов на питание Y и дохода X. Это графическое представление полного набора данных не выглядит достаточно ясным, особенно в левом нижнем углу [48]. Как отмечал W. Hardle: «Желательно иметь какой-либо метод, позволяющий увидеть места скопления данных» [48]. Иллюстрацией такого метода является так называемый *«цветочный график»*. На рис. 9.2 показан пример зависимо-

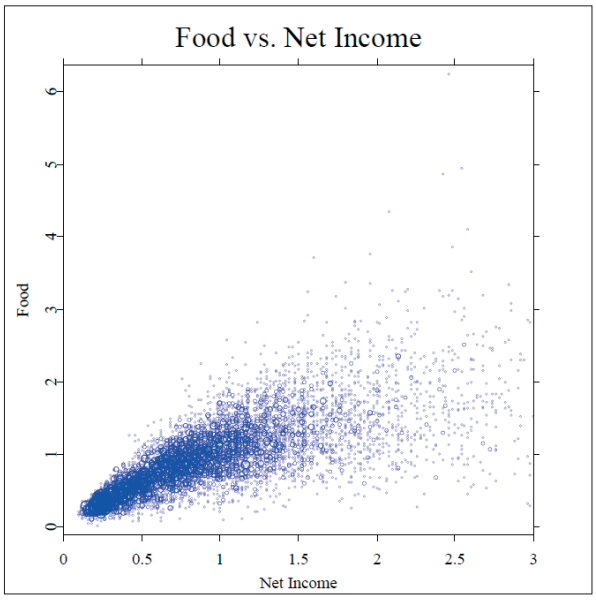


Рис. 9.2. «Цветочный график» расходов на питание Y от чистого дохода X

сти питания от чистого дохода.

Цветочный график строится посредством определения сети квадратов, покрывающих плоскость (X, Y), и подсчета числа наблюдений, попадающих в отдельные квадратики. Число «лепестков цветка» соответствует числу наблюдений в квадрате этого «цветка», представляя эмпирическое распределение данных. Такой график зависимости расходов на питание от чистого дохода указывает на сосредоточение данных вокруг увеличивающейся группы плотно упакованных «цветков». Форма этой группы позволяет предположить гладкую зависимость кривой среднего отклика от x.

9.2. Агрегация данных

В данном разделе рассматривается процедура агрегирования данных как метод предобработки для последующего регрессионного моделирования.

Суть процедуры агрегирования составляют методы, позволяющие первоначальный набор данных свести к наборам данных меньшего объема, сохраняя и обнаруживая при этом полезные знания в соответствии с возможностями используемых методов. Другими словами, агрегация представляет собой процедуру сжатия информации.

С другой стороны, агрегация может рассматриваться как процесс преобразования данных с высокой степенью детализации к более обобщен-

ному их представлению за счет вычисления так называемых агрегатов — значений, получаемых в результате применения данного преобразования к некоторому набору фактов, связанных с определенным измерением. Примером таких процедур является простое суммирование, вычисление среднего, медианы, моды или выбор максимального или минимального значений. Применение процедуры агрегирования имеет свои достоинства и недостатки. В качестве положительных моментов укажем, например, на то, что детализированные данные часто оказываются очень изменчивыми из-за воздействия различных случайных факторов, что затрудняет обнаружение общих тенденций и закономерностей исследуемого процесса. Важно иметь в виду, что применение таких процедур, как усреднение, исключение экстремальных значений (выбросов), процедур сглаживания, может привести к потере важной и значительной части информации об объекте исследования.

Существуют различные способы агрегирования данных. Рассмотрим подход агрегации данных, основанный на кусочно-полиномиальной аппроксимации. Он полезен по следующим причинам. В основе этого подхода лежит понятие кусочно-полиномиальной функции. Ее можно рассматривать как математический объект, который удобен для описания и вычисления математических процедур и операций, сохраняя суть частотного распределения данных. В рамках ВВА разработана численная арифметика над функциями плотностей вероятности, которая позволяет выполнять различные арифметические операции над случайными переменными, включая операции вычисления максимума и минимума, возведения в степень, процедуры сравнения. В настоящее время кусочно-полиномиальные функции, такие как гистограммы, активно применяются в различных областях. Например, гистограммы активно используются в базах данных, в задачах аппроксимирования и сжатия информации, распознавания образов и обработки изображений. В задачах обработки изображений гистограмма характеризует статистическое распределение количества пикселов изображения в зависимости от их значений. Для построения гистограммы достаточно вычислить количество пикселов в каждом интервале значений DN, а затем поделить его на общее число пикселов N.

Отметим важное свойство кусочно-полиномиальных функций, которое связано с понятиями агрегирования информации и данных. Кусочно-полиномиальные функции задаются сеткой, на каждом отрезке которой они являются некоторым полиномом. Использование кусочно-поли-

номиальных функций обусловлено прежде всего тем, что они позволяют достаточно точно представлять произвольные распределения случайных величин. Несмотря на свою простоту, они охватывают все возможные оценки функции плотности вероятности. Простая и гибкая структура кусочно-полиномиальных функций существенно упрощает их использование в численных расчетах и имеет наглядный визуальный образ, что удобно для аналитических выводов.

В работах [18] для агрегации данных временного ряда и данных, полученных на основе мониторинга дистанционного зондирования Земли, показана эффективность применения данного подхода к обработке данных с целью их агрегирования. В статье [32] для агрегации временных рядов водится понятие гистограммного временного ряда и на основе применения вычислительного вероятностного анализа строятся информационно-аналитические модели прогнозирования. В работе [18] рассматривается проблема изучения природных процессов на основе данных космического и наземного мониторинга. На основе вычислительного вероятностного анализа предлагается концептуально-гистограммный подход, который применяется для разработки процедур агрегирования информационных потоков, а также для численного моделирования и представления характеристик природных объектов. Показывается, что применение разработанных процедур снижает уровень неопределенности в данных и существенно повышает эффективность численных расчетов.

В качестве примера агрегаций могут служить следующие типичные ситуации: первая — если измеряется, например, температура по некоторой области, как обычно бывает при зондировании Земли из космоса. При этом в каждый момент времени измерения получается N значений температуры. Обычно в этих случаях используют среднее значение или интервал изменения. Понятно, что при этом значительная часть информации теряется. Агрегирование информации по пространственному признаку с помощью гистограмм позволяет более точно представить измеренную информацию. Такая агрегация называется *распределенной* [31].

Второй способ агрегации называется *«временной»* и возникает, когда переменная наблюдается, например, раз в секунду или в минуту, но должна быть проанализирована на более низкой частоте, скажем за день. В этом случае использование средних значений или интервального анализа приводит к потере информации.

Эти две ситуации описывают распределенную и временную агрегации соответственно. В каждом из этих случаев представление данных в ви-

де плотностей вероятности является более информативным, чем другие способы.

Рассмотрим пример временной агрегации. На рис. 9.1 изображен набор данных (x_l, y_l) расходов на питание (y_l) и дохода (x_l), число наблюдений равно 7125. Графическое представление полного набора данных визуально представляет область изменения и сосредоточения данных. Заметим, что по многим причинам работать с таким большим объемом информации напрямую неудобно. Для повышения эффективности анализа данных с целью извлечения полезной информации целесообразно эти данные агрегировать в гистограммы. Для этих целей построим в области изменения дохода X сетку $\{x_0 < x_1, ..., < x_n\}$. Для каждого отрезка $[x_{i-1}, x_i)$ построим множество $\mathcal{Y}_i = \{y_l | x_l \in [x_{i-1}, x_i)\}$. По множеству $\mathcal{Y}_i$ построим кусочно-полиномиальную функцию Y_i, $i = 1, ..., n$. Способ агрегирования исходного набора (x_l, y_l) в виде кусочно-полиномиальной функции $\{Y_i,\ i = 1,..,n\}$ будем называть «*временной*» агрегацией. Роль времени играет «доход».

9.3. Регрессионное моделирование на основе агрегированных данных

Регрессионное моделирование представляет собой способ исследования объектов на основе использования информационного подхода для выявления факта существования различных зависимостей между входными и выходными данными. Регрессионный анализ может рассматриваться как метод моделирования измеряемых данных и исследования их свойств. Данные представляются в виде пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Регрессионный анализ предназначен для изучения по выборочным данным статистической зависимости ряда величин, некоторые из которых являются случайными. При статистической зависимости величины не связаны функционально, но как случайные величины заданы совместным распределением вероятностей. Исследование зависимости случайных величин приводит к моделям регрессии и регрессионному анализу на базе выборочных данных. Теория вероятностей и математическая статистика представляют лишь инструмент для изучения статистической зависимости, но не

ставят своей целью установление причинно-следственной связи. Числовые данные обычно имеют между собой явные (известные) или неявные (скрытые) связи. Поэтому важной задачей для исследователя является с помощью различных методов выявить скрытые зависимости и закономерности, содержащиеся в данных, и выразить их в виде формул, т. е. математически смоделировать явления или процессы. Регрессионный анализ называют основным методом современной математической статистики для выявления неявных и завуалированных связей между данными наблюдений. В регрессионном анализе имеют место следующие допущения: количество наблюдений достаточно для проявления статистических закономерностей относительно факторов и их взаимосвязей; обрабатываемые данные содержат некоторые ошибки (помехи), обусловленные погрешностями измерений, воздействием неучтенных случайных факторов; матрица результатов наблюдений является единственной информацией об изучаемом объекте, имеющейся в распоряжении перед началом исследования.

В условиях «больших» данных предполагается целесообразным имеющиеся наблюдения или «сырые данные» подвергнуть процедуре обработки, в частности агрегации.

С этой целью применим вычислительный вероятностный анализ, и только после этого на агрегированных данных будем строить регрессионные модели. Такой подход уже на стадии подготовки данных к моделированию позволяет провести предварительный анализ данных с целью их сжатия, извлечения полезной информации и ориентировать вычислительный процесс на оптимизацию численных процедур, необходимых для построения регрессионных моделей. Ниже сформулируем «классическую» постановку задачи регрессионного моделирования и далее перейдем к постановке задачи для «гистограммной» регрессии.

9.4. Классическая параметрическая регрессия

Пусть входные данные $X = (x_1, ..., x_n)$ и целевая переменная Y являются числовыми. Тогда для каждой записи Y_i, X_i можно построить модель:

$$Y_i = f(X_i, a) + \epsilon_i, i = 1, ..., N,$$

где f — функция зависимости целевой переменной от входных данных и некоторых параметров, $a = (a_0, a_1, ..., a_n)$ — параметры регрессии, а ϵ_i

— ошибки. Необходимо найти наилучшую функцию f и наилучшие параметры a таким образом, чтобы ошибки ϵ были достаточно малы. Если накладывается условие: $||\epsilon|| \to \min$, то можно говорить об оптимизации функции

$$\Phi(a) = \sum_{i=1}^{N} ||Y_i - f(X_i, a)||^2.$$

Существует класс функций и норма $||\cdot||$, когда задача регрессии может быть сведена к решению систем линейных алгебраических уравнений. В случае линейных функций относительно $x_1, ..., x_n$ и евклидовой нормы получаем классическую линейную регрессионную модель

$$\Phi(a) = \sum_{i=1}^{N} (Y_i - a_0 - \sum_{j=1}^{n} a_j x_{ji})^2,$$

решение которой сводится к решению системы линейных алгебраических уравнений

$$Aa = b,$$

где матрица $A = X'X$, $b = X'Y$:

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ 1 & x_{21} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \dots & x_{Nn} \end{pmatrix}.$$

9.5. Метрики в пространстве распределений

В силу специфики кусочно-полиномиальных функций далее будем использовать специальные подходы к количественной оценке расстояний между функциями плотности вероятности.

Для этого рассмотрим метрику Вассерштейна и Mallows [97] для обратных функций распределения. Пусть $f(x)$ и $g(x)$ — функции плотности вероятности, тогда представим метрики

$$\rho_W(f, g) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt, \tag{9.1}$$

$$\rho_M(f, g) = (\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt)^{1/2}, \tag{9.2}$$

где $F^{-1}(t)$, $G^{-1}(t)$ — обратные функции к функции распределения.

Причинами выбора этой метрики было успешное применение в ряде работ [56]. Более того, эта метрика эквивалентна Earth Mover's Distance (EMD), предложенной Rubner, Tomasi и Guibas [124]. EMD — известное компьютерное расстояние, которое используется для того, чтобы измерить несходства между гистограммами текстуры и цвета. EMD между двумя гистограммами — наименьшее количество объема работы, чтобы преобразовать одну гистограмму в другую.

Если $h(x)$ — некоторая кусочно-полиномиальная функция, тогда функцию распределения H, соответствующую этой функции, можно представить в виде

$$H(x) = \int_{-\infty}^{x} h(\xi)d\xi.$$

В силу того, что $h(x)$ — кусочно-полиномиальная функция, вычисление интеграла от нее не представляет особого труда. В результате функцией распределения H будет кусочно-полиномиальная функция. Таким образом, метрики (9.1),(9.2) можно интерпретировать как площадь между функциями распределения.

9.6. Регрессия над эмпирическими распределениями

Пусть входные данные $X = (x_1, ..., x_n)$ и целевая переменная Y являются гистограммными переменными и для вектора $X = (x_1, ..., x_n)$ известна совместная плотность вероятности $p(x_1, ..., x_n)$.

Аналогично числовой регрессии для каждой пары (X_i, Y_i) можно записать

$$Y_i = f(X_i, a) + \epsilon_i, i = 1, ..., N,$$

или в случае линейной модели

$$Y_i = a_0 + \sum_{j=1}^{n} a_j x_{ij} + \epsilon_i, i = 1, ..., N.$$

Таким образом, для нахождения неизвестных параметров $a_0, a_1, ..., a_n$ можно записать задачу оптимизации

$$\Phi(a) = \sum_{i=1}^{N} \rho(Y_i, a_0 + \sum_{j=1}^{n} a_j x_{ji})^2 \to \min. \tag{9.3}$$

В силу нелинейности операции сложения для решения задачи (9.3) можно использовать, например, метод наискорейшего спуска [5].

Для вычисления градиента Φ' будем использовать разностные производные

$$\Phi'_i = \frac{\Phi(a_0, ..., a_i + h, ..., a_n) - \Phi(a_1, ..., a_i, ..., a_n)}{h},$$

где h — параметр.

Начальное приближение для вектора параметров $a_0, a_1, ..., a_n$ можно получить, решив задачу регрессии для математических ожиданий $(\mathsf{M}[X_i], \mathsf{M}[Y_i])$.

Пример 12. Рассмотрим линейную модель

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \varepsilon,$$

где X_1, X_2 — предикторные переменные, Y — зависимая переменная, ε — ошибка. По наблюдаемым значениям Y_i $X_{1,i}$ $X_{2,i}$ после агрегации плотности Y, X_1, X_2 представлены сплайнами: S_y, S_1, S_2.

Неизвестные параметры a_0, a_1, a_2 будем искать исходя из минимума функционала

$$\Phi(a_0, a_1, a_2) = \rho(S_y, (a_0 + a_1 S_1 + a_2 S_2)) \to \min.$$

Для численной реализации X_1, X_2 генерировались как суммы трех равномерных на $[0, 1]$ случайных величин, сдвинутых на 1 и 2 соответственно, ε с функцией плотности вероятности $(|2x| - 1)^2(2|2x| + 1)$ с носителем $[-0.5, 0.5]$ и $Y = X_1 + X_2 + \varepsilon$.

Минимизация функционала $\Phi(a_0, a_1, a_2)$ осуществлялась методом наискорейшего спуска. При значениях $a_0 = -0.089$, $a_1 = 1.031$, $a_2 = 1.029$ величина $\Phi(a_0, a_1, a_2)$ не превысила значения $0.3 \cdot 10^{-3}$.

На рис. 9.3 показано: линия 1 — результат регрессии над эмпирическими распределениями, линия 2 — ее относительная ошибка, умноженная на 1000.

9.7. Эмпирическая функциональная регрессия

Пусть известны значения (x_i, y_i) $i = 1, 2, \ldots, N$. Будем считать, что случайная величина y_i распределена по закону $\boldsymbol{y}(x_i)$, семейство функций плотности вероятности $\boldsymbol{y}(x)$ зависит непрерывно от x.

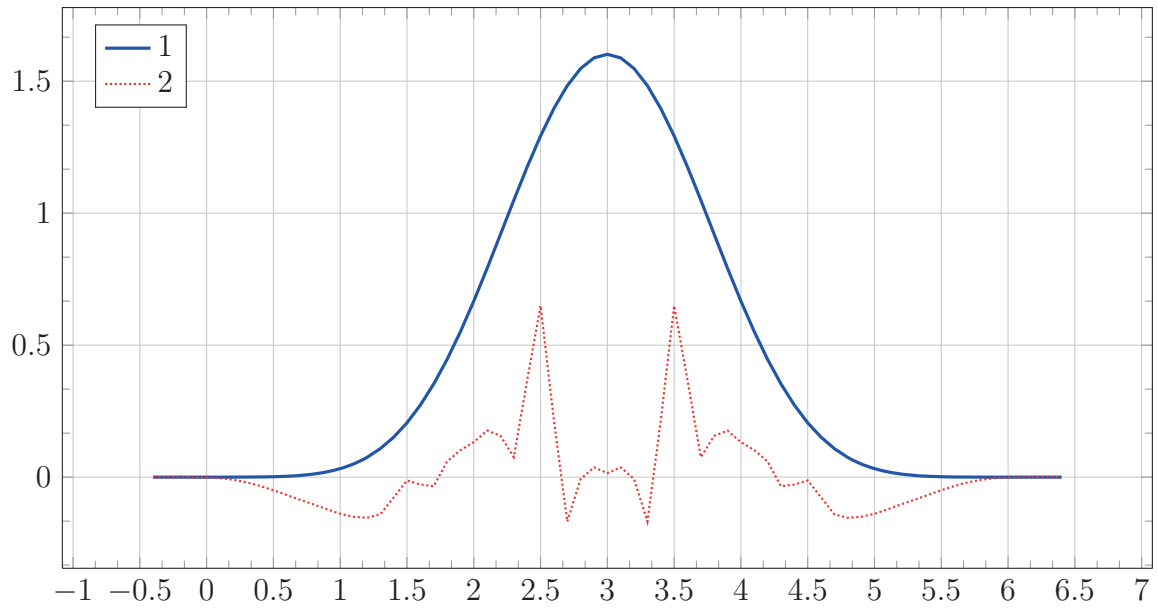


Рис. 9.3. Регрессия над эмпирическими распределениями

Необходимо по данным (x_i, y_i) $i = 1, 2, \ldots, N$ оценить семейство функций плотности вероятности $\boldsymbol{y}(x), x \in X$ и построить функциональную регрессию:

$$\boldsymbol{y}(x) = \sum \boldsymbol{a}_i \varphi_i(x) + \varepsilon(x).$$

Для построения оценки $\boldsymbol{y}$ в некоторой точке x_0 зададимся параметром $h > 0$ и построим на отрезке $[x_0 - h, x_0 + h]$ по данным $D_h = \{(y_i, x_i) | x_i \in [x_0 - h, x_0 + h]\}$ регрессию $r(x)$. Далее построим выборку $Z_h = \{z_i = y_i - r(x_i) | x_i \in [x_0 - h, x_0 + h]\}$. По Z_h и используя ядерные оценки, построим приближение $\hat{\boldsymbol{y}}^h(\cdot, x_0) \approx \boldsymbol{y}(\cdot, x_0)$. Заметим, что в данном случае мы строим оценку $\bar{\boldsymbol{y}}^h(\xi, x_0)$:

$$\bar{\boldsymbol{y}}^h(\xi, x_0) = \frac{1}{2h} \int_{x_0 - h}^{x_0 + h} \boldsymbol{y}(\xi, x) dx.$$

Несложно видеть, что

$$\boldsymbol{y}(\xi, x_0) = \bar{\boldsymbol{y}}^h(\xi, x_0) + Ch^2 + O(h^4),$$

где C — константа, не зависящая от h. В этом случае для повышения точности можно использовать экстраполяцию Ричардсона. Для этого построим оценки при h и $2h$: $\bar{\boldsymbol{y}}^h(\xi, x_0)$ и $\bar{\boldsymbol{y}}^{2h}(\xi, x_0)$. Далее

$$\boldsymbol{y}(\xi, x_0) = \frac{4}{3} \bar{\boldsymbol{y}}^h(\xi, x_0) - \frac{1}{3} \bar{\boldsymbol{y}}^{2h}(\xi, x_0) + O(h^4).$$

Будем стремиться построить оценки так, чтобы

$$\bar{\boldsymbol{y}}^h(\xi, x_0) - \hat{\boldsymbol{y}}^h(\xi, x_0) \approx O(h^4).$$

В этом случае будет справедлива оценка

$$\boldsymbol{y}(\xi, x_0) = \frac{4}{3}\hat{\boldsymbol{y}}^h(\xi, x_0) - \frac{1}{3}\hat{\boldsymbol{y}}^{2h}(\xi, x_0) + O(h^4).$$

Модельный пример. Сгенерируем данные следующим образом: построим сетку $\{x_i = ih, i = 0, 1. \ldots, N\}$. Для каждого x_i сгенерируем случайную величину y_i

$$y = ((4 - x)x/8 + 1/4) + (t_1 + t_2 + t_3 + t_4)/(4 + 4x),$$

где $N = 80\ 000$, $h = 1/N$. Результаты моделирования исходных данных приведены на рис. 9.4.

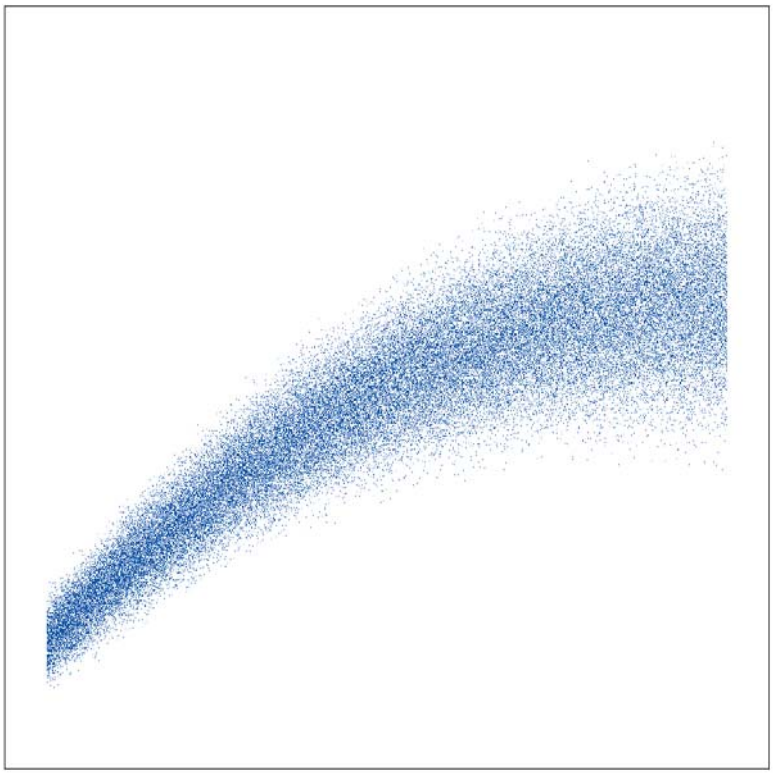


Рис. 9.4. Исходные данные

На рис. 9.5 оттенками серого приведена оценка семейства $\boldsymbol{y}$. Красные сплошные линии: верхняя граница носителей $\boldsymbol{y}(x)$, математическое ожидание $\boldsymbol{y}(x)$ и нижняя граница носителей $\boldsymbol{y}(x)$.

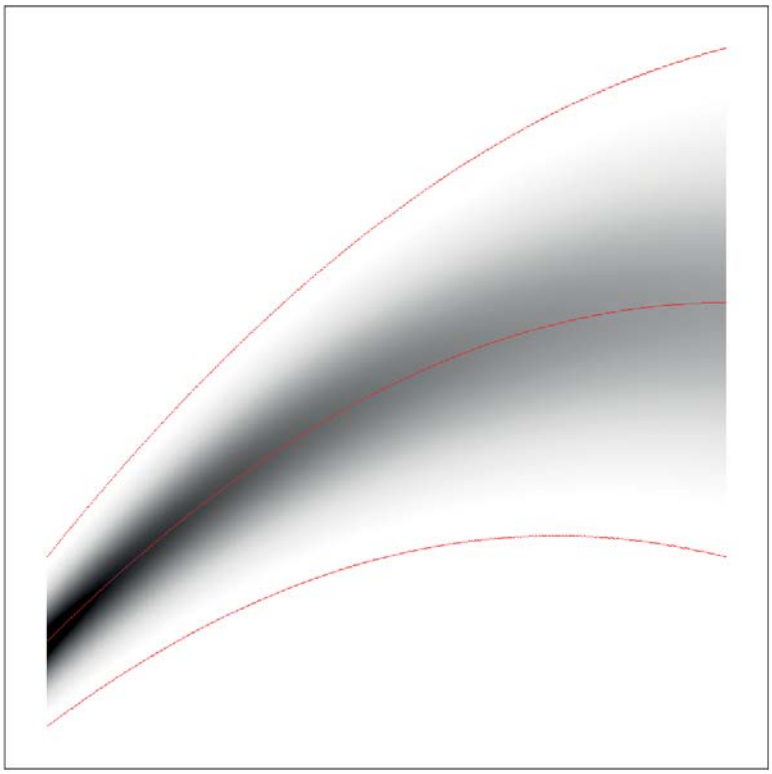


Рис. 9.5. Оценка семейства $\boldsymbol{y}$

9.8. Применение регрессионного подхода к функциональным временным рядам

Рассмотрим вопрос численного моделирования для агрегации функциональных временных рядов. Временной ряд хорошо подходит для моделирования многих практических ситуаций. Следует отметить, что во многих случаях временные ряды анализируются как данные большого объема. Для анализа связи между данными по времени будем использовать процедуры агрегации. Для начала отметим следующие моменты.

Известно, что временные ряды хорошо описывают эмпирические данные для многих практических и теоретических ситуаций. Есть исследования, показывающие, что временные ряды не дают достоверного описания явления, где наблюдаемая переменная имеет определенную степень изменчивости.

Это происходит в двух типичных ситуациях [55]: если некоторый фактор измеряется по времени для каждого человека в группе, но интерес не в отдельных людях, а в группе в целом. В этом случае временной ряд среднего значения наблюдаемого фактора даст слабое представление о

процессе. Другой случай, когда переменная наблюдается с заданной частотой (скажем, раз в минуту), но должна анализироваться с меньшей частотой (например, за весь день).

Эти две ситуации описывают распределенную и временную агрегацию соответственно. В каждом случае временной ряд распределений будет предлагать более информативное представление, чем другие формы агрегированных временых рядов. В качестве доказательства мы представляем, что Schweizer утверждает, что «распределения — это числа будущего» [126]. Таким образом, лучше предложить методы, которые имеют дело с распределениями напрямую. Рассмотрим модели представления распределений.

В нашем исследовании мы предлагаем представить их с помощью кусочно-полиномиальной функции агрегации, потому что она предлагает хороший компромисс между простотой и точностью.

Рассмотрим использование регрессионного подхода к сплайн-агрегированным временным рядам.

В этом параграфе мы сосредоточимся на ситуациях, когда мы хотим описать изменчивость данных как регрессионную модель.

В начале мы предлагаем изучить следующую модель:

$$Y = a_1\varphi_1(t) + a_2\varphi_2(t) + a_3\varphi_3(t),$$

где a_1, a_2, a_3 — постоянные.

Рассмотрим исторические данные о максимальной температуре в г. Красноярске за последние сто лет. Для каждого дня с 01 апреля по 01 октября данные агрегируются в виде сплайнов Y_i, $i = 1, 2, ..., 184$. На рис. 9.6 линии 1 — кусочно-полиномиальные аппроксимации функций плотности вероятности максимальной температуры за 15.04, 01.05, 15.05, 01.06, 15.06. Линии 2 — кусочно-полиномиальная регрессия функций плотности вероятности, линии 3 — регрессионные кривые границ носителей и максимумов функций плотности вероятности.

В этом случае регрессионная модель может быть представлена в виде

$$\hat{Y}_i = A_1\varphi_1(t_i) + A_2\varphi_2(t_i) + A_3\varphi_3(t_i), i = 1, 2, ..., 184,$$

где A_1, A_2, A_3 — функции плотности вероятности, φ_1, φ_2, φ_3 являются квадратичными функциями:

$$\varphi_1(t_1) = 1, \varphi_1(t_{92}) = 0, \varphi_1(t_{184}) = 0,$$

$$\varphi_2(t_1) = 0, \varphi_2(t_{92}) = 1, \varphi_2(t_{184}) = 0,$$

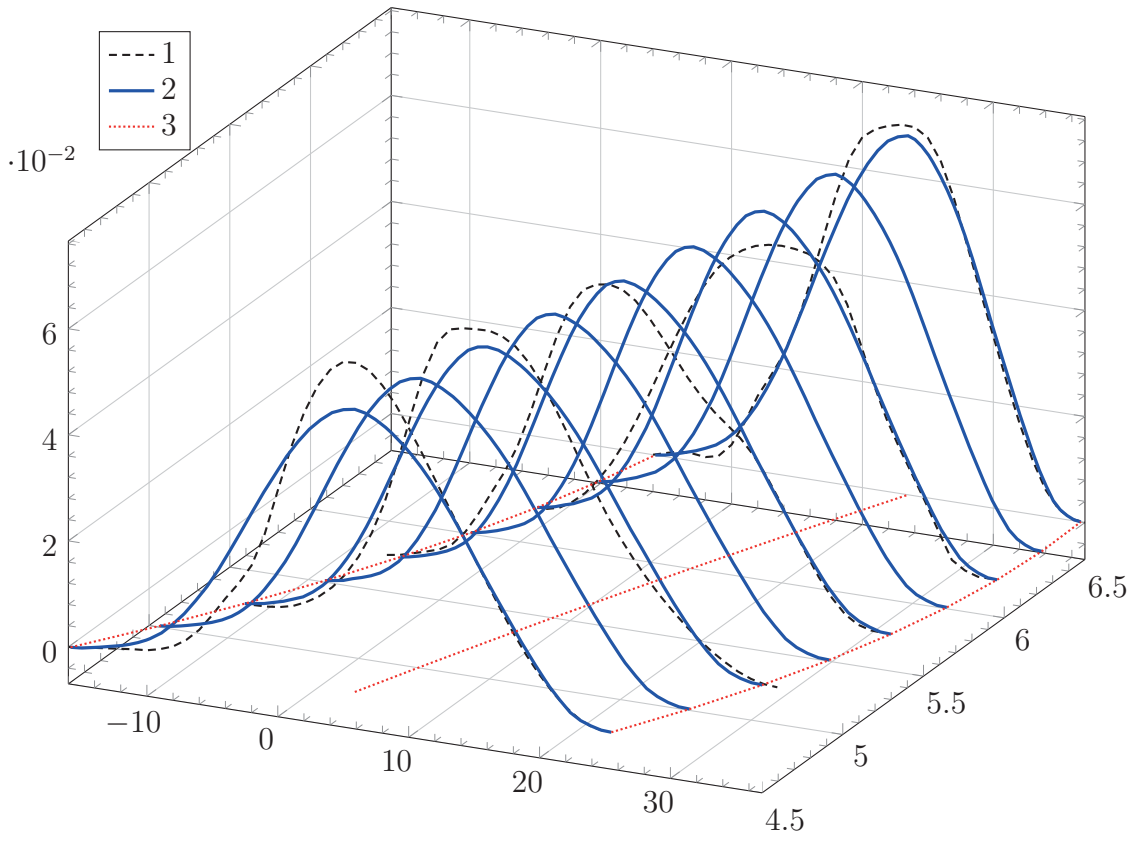


Рис. 9.6. Аппроксимации функций плотности вероятности максимальной температуры

$$\varphi_3(t_1) = 0, \varphi_3(t_{92}) = 0, \varphi_3(t_{184}) = 1.$$

Функции плотности $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$ представлены в форме эрмитовых сплайнов s_i.

Сплайны определяются сеткой $\{x_1^i, x_2^i, x_3^i\}$. Граничные условия: $s(x_1^i) = 0$, $s'(x_1^i) = 0$, $s(x_3^i) = 0$.

Переменные x_1^i, x_3^i расположены на регрессионных кривых минимальных и максимальных температур. Для x_2^i выбрана кривая регрессии средней температуры и

$$\Phi(\vec{a}) = \sum_{i=1}^{184} \rho^2(Y_i(\vec{a}), \hat{Y}_i),$$

$$\Phi(a^*) = \min_{\vec{a}} \Phi(\vec{a}).$$

На рис. 9.7 показаны функции плотности вероятности температурных данных за последние сто лет в городе Красноярске, с 01 апреля по 01 октября. Оттенки серого определяют значения функции плотности вероятности. Верхняя и нижняя линия представляет максимальную и

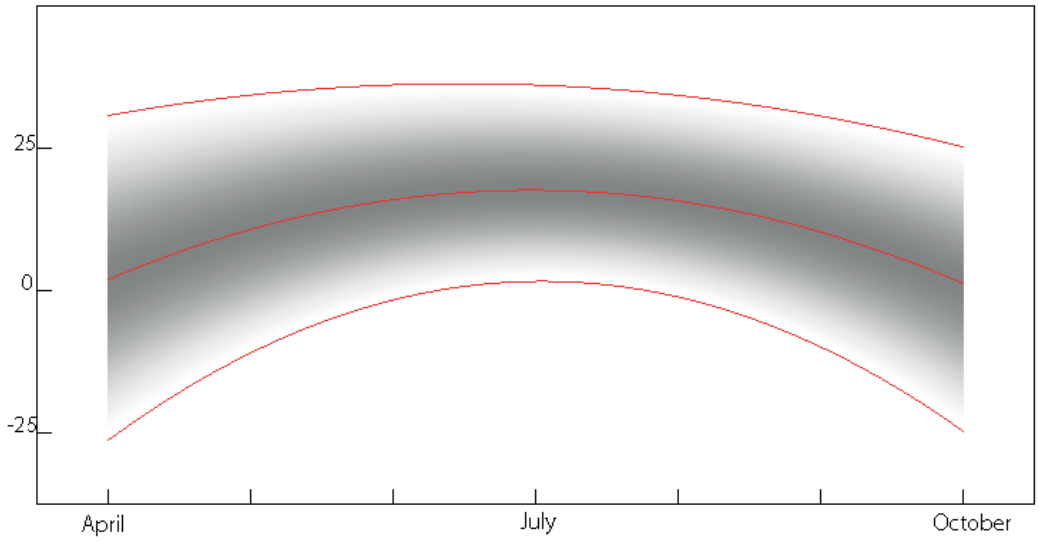


Рис. 9.7. Аппроксимации функции плотности вероятности температуры

минимальную температуру в каждый день за последние сто лет соответственно. Средняя линия обозначает среднее значение дневной температуры за последние сто лет. Каждое вертикальное сечение является приближением функции плотности вероятности температуры, соответствующей определенному дню года, в соответствии с наблюдениями дня за последние сто лет. На первом этапе данные были представлены функцией распределения для каждого дня в сплайне Эрмита пятой степени, как указано выше.

Данные регрессии представлены в виде производной эрмитова сплайна пятой степени. Таким образом, данные о температуре в течение ста лет с апреля по октябрь агрегируются с помощью эрмитовых сплайнов пятой степени. Визуальное представление показывает изменение в максимальной, минимальной и наиболее вероятной температуре. Оттенки серого показывают распределение плотности вероятности.

Разработанные на основе ВВА процедуры и методы кусочно-полиномиального представления функций плотности вероятности, обработки и численного регрессионного моделирования позволяют снизить уровень неопределенности в информационных потоках, существенно сократить время обработки и выполнения численных процедур. Данный подход также позволяет в режиме визуально-интерактивного моделирования представить необходимые данные для оперативного принятия решений.

Глава 10

Приложения ВВА

Вычислительный вероятностный анализ направлен на решение разнообразных практических задач, где присутствует необходимость обработки эмпирической информации, содержащей как эпистемическую, так и элиторную неопределенность.

Круг решаемых с помощью ВВА практических задач достаточно широк и может быть определен следующими ситуациями.

Первая ситуация характеризуется наличием некоторого вероятностного пространства, в котором определены и известны вероятностные распределения параметров, а переменные и функции интерпретируются как случайные. Эта ситуация приводит к стохастической постановке задачи.

Вторая ситуация определяется тем, что вероятностные распределения параметров модели неизвестны, однако в распоряжении у исследователя имеется достаточное количество накопленной информации о реализациях параметров, на основе которой можно построить статистически достоверные оценки вероятностного распределения.

Третья ситуация состоит в том, что вероятностные распределения параметров модели неизвестны, однако в распоряжении у исследователя имеется недостаточное количество информации о реализациях параметров (малые выборки), на основе которой можно построить статистически достоверные оценки вероятностного распределения [7].

Среди таких задач можно выделить задачи цифровой экономики и проблемы обработки данных больших объемов (Big Data) [20, 75], оценку показателей надежности систем ответственного назначения в условиях ограниченного объема информации (малые выборки) [38], обработку и численные методы анализа данных дистанционного зондирования Земли [72], задачи оценки экономических и других рисков.

10.1. Проблемы цифровой экономики

Раздел посвящен важным аспектам, характеризующим проблемы цифровой экономики [66, 75]. Первая проблема связана с большой обработкой данных и обнаружением знаний в базе данных [104]. Вторая проблема касается вопроса об уменьшении уровня неопределенности и изучении изменчивости данных в больших базах данных.

Цифровая экономика, основанная на больших данных, является прогностической по своему типу: здесь прогноз, план и факт имеют тенденцию к равенству. Его основным инструментом является прогностическая аналитика, основной вид производства персонализирован для нужд клиента, и конкуренция идет не столько за перераспределение существующих рынков, сколько с образованием новых, где больше конкурируют не товары и технологии, а цифровые системы управления на основе цифровых платформ [66].

Наш подход основан на технологиях Big Data [104], включая процедуры агрегации данных для входных и выходных параметров и вычислительного вероятностного анализа.

Важной особенностью исследования является демонстрация способа представления агрегированных данных. Предлагается использовать кусочно-полиномиальные модели, в том числе сплайн-агрегатные функции. Мы показываем, что предлагаемый подход к агрегации данных можно интерпретировать как распределение частоты. Для изучения его свойств используется понятие функции плотности. Обсуждаются различные типы математических моделей агрегации данных [73, 74].

Хотя существует множество способов агрегации данных, включая простое среднее, мы утверждаем, что использование кусочно-полиномиальных функций агрегации будет предлагать более информативное представление об изменчивости в больших данных, чем другие формы агрегации.

В настоящее время доминирующие парадигмы экономических теорий основаны на классической математике и представлены в терминах вероятностных и статистических методов. Следует подчеркнуть, что в приложениях вероятностные и статистические методы часто и успешно используются в синтезе с современными методами мягких вычислений. Теперь понятно, что в приложениях мы часто имеем дело со случайной и эпистемической неопределенностью [68].

В последние десятилетия существует много современных методов мо-

делирования неопределенности. Как правило, они не противоречат традиционному вероятностному подходу, поскольку касаются других (не вероятностных) типов неопределенностей.

Обработка неопределенности в анализе, проектировании и принятии решений проходит переход парадигмы от вероятностной основы к обобщенной структуре, которая включает как вероятностные, так и невероятностные методы.

Многие важные практические проблемы связаны с различными типами неопределенностей. На практике несколько источников неопределенности требуемой информации препятствуют оптимальному принятию решений в классическом смысле.

Когда доступна только неопределенная информация (что наиболее часто бывает), тогда принятие решения требует более сложных методов представления данных и их анализа.

Анализ подходов к оценке инвестиционных рисков

Важность учета влияния факторов неопределенности на оценку финансовых рисков, в том числе рисков инвестиционных проектов, сегодня признается всеми финансовыми аналитиками. Под неопределенностью в анализе инвестиционных проектов понимается возможность разных сценариев реализации проекта, которая возникает из-за неполноты, неточности информации об условиях реализации инвестиционного проекта; ошибок в расчетах параметров проекта, обусловленных упрощениями при формировании моделей сложных технических и организационно-экономических систем; колебаниями рыночной конъюнктуры, цен, валютных курсов; непредсказуемости политической ситуации, условий инвестирования и использования прибыли.

Тем не менее в настоящее время одним из основных инструментов финансовых аналитиков является метод дисконтированных денежных потоков, который наряду с несомненными преимуществами имеет ряд недостатков. Так в работе [26] отмечается, что на теоретическом уровне метод дисконтированных денежных потоков не учитывает вероятностный характер результатов инвестиционного проекта, и предлагается в условиях высокой неопределенности и риска использовать альтернативные методы, одним из которых является метод Монте-Карло.

Отмечается, что метод дисконтированных денежных потоков и разработанные за последние годы многочисленные альтернативные мето-

ды, частично устраняющие его недостатки, дают удовлетворительные результаты, только если дискретные и непрерывные риски находятся на низком уровне. В случае наличия значительных дискретных рисков используется метод дерева решений. При значительной непрерывной неопределенности применяется компьютерное моделирование по методу Монте-Карло.

Наконец, при наличии высокого уровня непрерывной неопределенности и значительных дискретных рисков применяется метод реальных опционов.

В работе [17], как альтернатива методу Монте-Карло, рассматривается вычислительный вероятностный анализ, реализующий подход, основанный на использовании методов построения функции плотности вероятности по эмпирической информации в условиях неопределенности для параметров инвестиционной модели и численные процедуры построения законов распределения значений соответствующих показателей.

В данном параграфе предлагаются новые подходы, основанные на процедурах построения вероятностных расширений, а также процедуры, существенно повышающие надежность численных результатов с целью снижения уровня неопределенности в данных.

Расчет NPV и IRR

Приведем пример модели и необходимых пояснений, которые используются для оценки привлекательности инвестиционных проектов предприятия, производящего товары.

Это прежде всего оценка денежных потоков CF_z, далее чистый дисконтированный доход NPV и внутренняя норма доходности IRR.

$$\begin{aligned} CF_z = &\sum_i G_i P_i (1 - AV_{Gi}) - \sum_j E_j (1 - AV_{Ei}) - \\ &- \sum_k S_k W_k T_{Wk} - \sum_p H_p (A_p + T_{Hp}) - \\ &- \sum_q T_q - \sum_n C_n R_n + Fin_z - Fout_z, \end{aligned} \tag{10.1}$$

где G_i — количество продаж i-го товара; P_i — цена i-го товара; AV_{Gi} — налог на добавленную стоимость на i-ый товар; E_j — расходы j-го вида; AV_{Ej} — налог на добавленную стоимость на приобретаемые товары по j-му виду расходов; S_k — численность персонала k-й категории; W_k —

средняя заработная плата работников k-й категории; T_{Wk} — коэффициент отчислений во внебюджетные фонды по k-й категории работников; H_p — основные фонды p-го вида; A_p — норма амортизации по p-у виду основных фондов; T_{Hp} — ставка налога, базой для расчета которого выступает p-й вид основных фондов (налог на имущество); T_q — сумма q-го налога; C_l — сумма l-го кредита; R_l — процент по l-му кредиту; TPr_z — ставка налога на прибыль в z-м году; Fin_z — прочие операционные, финансовые и инвестиционные поступления средств; $Fout_z$ — прочие операционные, финансовые и инвестиционные выплаты средств.

NPV — это сумма приведенных к текущему моменту времени чистых денежных потоков по инвестиционному проекту. Данный показатель определяется по следующей формуле:

$$NPV = \sum_{z=1}^{T} \frac{CF_z}{(1+d)^z},$$

где T — расчетный срок инвестиционного проекта в годах; d — ставка дисконтирования. Инвестиционный проект признается эффективным в случае, если $NPV > 0$.

IRR — расчетная ставка дисконтирования, при которой чистый дисконтированный доход (NPV) равен нулю. IRR определяется из уравнения

$$\sum_{z=1}^{T} \frac{CF_z}{(1+IRR)^z} = 0.$$

Инвестиционный проект признается эффективным в случае, если IRR> d.

Как правило, для оценки инвестиционных проектов необходимо рассчитывать показатели на несколько лет вперед. В условиях высокой рыночной неопределенности такие показатели, например, как G_i, P_i, имеют существенно стохастический характер. Однако при расчетах, как правило, используют только детерминированные оценки.

Такая методика расчета как на теоретическом, так и на практическом уровне не учитывает вероятностный характер входных и результирующих показателей инвестиционного проекта, делает невозможным управление проектом в ходе его реализации и соответственно затрудняет принятие эффективных управленческих решений.

Как было показано в [68], учет стохастических неопределенностей и использование вычислительного вероятностного анализа позволяют существенно поднять качество оценки проектов.

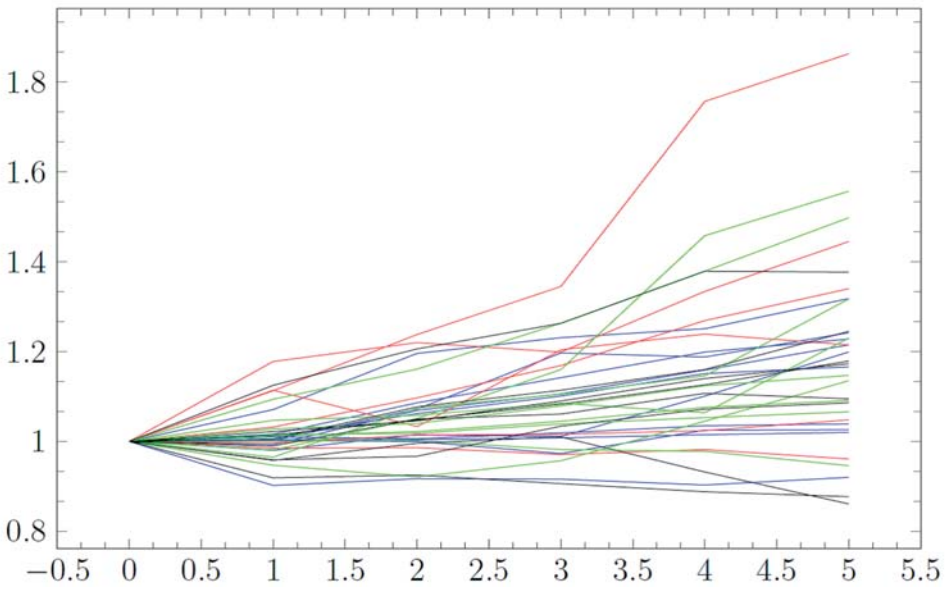


Рис. 10.1. Приведенные объемы продаж фирм-аналогов

Используя технологии вычислительной экономики и экономической аналитики [62, 132, 136], мы можем построить приведенные объемы продаж фирм-аналогов.

На рис. 10.1 показаны приведенные объемы продаж фирм-аналогов. Далее с помощью процедур агрегации могут быть построены кусочно-полиномиальные функции, аппроксимирующие плотность вероятности входных параметров.

Следовательно, используя BBA, мы можем получить оценки плотностей вероятности NPV и IRR в виде кусочно-полиномиальных функций. Следовательно можно оценить соответствующие риски. Так, если P_{NPV} — функция плотности вероятности NPV, то

$$P_u = \int_{-\infty}^{0} P_{NPV}(\xi) d\xi$$

— вероятность того, что инвестиционный проект окажется убыточным.

В качестве примера использования вероятностных расширений рассмотрим задачу оценку риска принятия решения об инвестировании проекта выпуска лекарственного препарата [26].

Компания рассматривает вопрос о приобретении для последующего производства патента нового лекарственного препарата. Стоимость патента составляет 3,4 млн долл. Решение принимается на основе анализа дисконтированных денежных потоков NPV и IRR. Горизонт расчетов составляет три года. Стандартная финансовая модель приводится в табл.

7. Согласно прогнозам компания в первый, второй и третий год проекта продаст соответственно 802 тыс., 967 тыс. и 1 132 тыс. упаковок лекарства по цене 6, 6,05 и 6,10 долл. за упаковку.

Ставка налога на прибыль равна 32 %, ставка дисконтирования — 10 %, себестоимость составляет 55 %, а операционные издержки — 15 % от цены препарата. По результатам расчетов IRR проекта составляет 15 %, а NPV — 344,8 тыс. долл.

В данном случае мы имеем дело с высоким уровнем рыночной неопределенности, поэтому стандартная финансовая модель не может дать достаточных для принятия решения результатов. Для одновременного учета неопределенности в цене, продажах, себестоимости и издержках применяется численный вероятностный анализ. Основные параметры финансовой модели — цена, объем продаж — моделируются как случайные переменные, имеющие вероятностное распределение. Численный вероятностный анализ позволит понять, какие факторы в наибольшей степени повлияют на финансовые результаты проекта. Для моделирования цены продажи (в первый, второй и третий год проекта отдельно) используется треугольное распределение. Треугольное распределение имеет три параметра — минимальное значение, максимальное значение и наиболее вероятное значение. Цена продажи в первый год имеет минимальное значение 5,90 долл., максимальное значение — 6,10 долл. и наиболее вероятное значение — 6,00 долл. Аналогично, цена продажи во второй год имеет треугольное распределение с параметрами 5,95; 6,05; 6,15 долл. Цена продажи на третий год имеет треугольное распределение с параметрами 6,00; 6,10; 6,20 долл. Объем продаж моделируется как случайная переменная с нормальным распределением. Объем продаж в первый год имеет нормальное распределение со средним значением (математическим ожиданием) 802 тыс. долл. и стандартным отклонением 25 тыс. долл. Аналогично, объем продаж во второй год имеет нормальное распределение с ожиданием 967 тыс. долл. и стандартным отклонением 30 тыс. долл. Наконец, объем продаж в третий год имеет нормальное распределение с ожиданием 1132 тыс. долл. и стандартным отклонением 25 тыс. долл. Себестоимость (процент от продаж), как предполагается, имеет треугольное распределение с минимальным значением 50 %, максимальным значением 65 % и наиболее вероятным значением 55 %. Следует отметить, что в данном случае треугольное распределение имеет не симметричную форму, а немного скошено вправо, т. е. имеется большая вероятность того, что себестоимость будет завышена, а не занижена

по сравнению с наиболее вероятным значением. Операционные издержки (процент от продаж) моделируются как нормальное распределение с ожиданием 15 % и стандартным отклонением 2 %.

Таблица 7

Стандартная финансовая модель

	Год 0	Год 1	Год 2	Год 3
Цена упаковки	—	$ 6,00	$ 6,05	$ 6,10
Количество проданных, шт.	—	802 000	967 000	1 132 000
Выручка	—	$ 4 812 000	$ 5 850 350	$ 6 905 200
Себестоимость	—	$ 2 646 600	$ 3 217 693	$ 3 797 860
Валовая прибыль	—	$ 2 165 400	$ 2 632 658	$ 3 107 340
Операционные издержки	—	$ 324 810	$ 394 899	$ 466 101
Чистый доход до налогов	—	$ 1 840 590	$ 2 237 759	$ 2 641 239
Налоги	—	$ 588 989	$ 716 083	$ 845 196
Стартовые инвестиции	–$ 3 400 000			
Чистый доход	–$ 3 400 000	$ 1 251 601	$ 1 521 676	$1 796 043

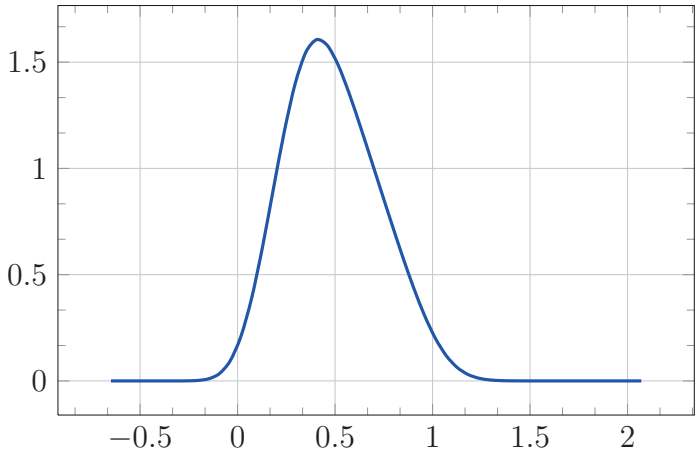


Рис. 10.2. Функция плотности вероятности NPV

На рис. 10.2, 10.3 приведены сплайновые аппроксимации функций плотности вероятности NPV и IRR для проекта. Из анализа функций плотности вероятности NPV и IRR видно, что вероятны как крайне негативные исходы, так и значительная прибыль в сравнении со стандартным анализом. Приведенный пример показывает, что применение численной вероятностной арифметики в рамках технологии визуально-интерактивного моделирования (ВИМ) [13] позволяет ЛПР увидеть воз-

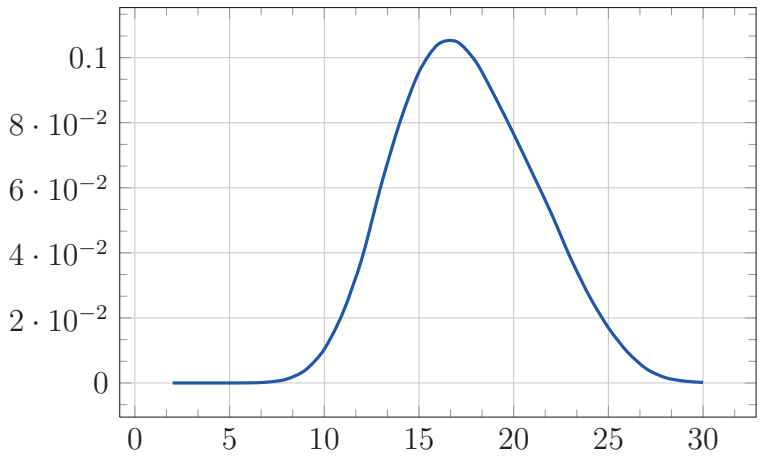


Рис. 10.3. Функция плотности вероятности IRR

можные варианты негативных исходов реализации проекта по сравнению со стандартным анализом, который дает только положительный ответ.

Результаты проведенных исследований и численные примеры показывают возможность использования гистограмм второго порядка для представления эпистемической неопределенности, что позволяет применить процедуры распространения неопределенности с целью получения достоверных знаний для выдвижения дополнительных оснований принятия решений в задачах оценки выбора инвестиционных проектов.

Проведенные практические исследования позволяют сделать два основных вывода:

• вероятностная арифметика может рассматриваться как численный метод вероятностного анализа, позволяющий работать с неопределенными данными в рамках различных практических приложений;

• способы представления информации в виде математических моделей и возможность вычислять и строить различные распределения параметров на основе эмпирических данных может использоваться как инструмент визуально-интерактивной технологии, что значительно повышает качество анализа возможных вариантов решений и дает в руки лица, принимающего решения (ЛПР), удобное средство для исследования и оценки существующих возможностей выбора.

10.2. Методика построения гарантированых оценок показателей надёжности

Задачи анализа и численного моделирования состояний и процессов функционирования сложных технических систем входят в круг практического применения ВВА. Отказы сложных технических систем ответственного назначения (ТСОН) могут стать источником техногенных чрезвычайных ситуаций. Экономические, экологические и социальные последствия подобных происшествий обусловливают необходимость совершенствования научных основ анализа техногенных рисков и обеспечения техногенной безопасности. Как правило, объекты ответственного назначения относятся к числу уникальных систем.

Обеспечение надежности является одним из главных элементов общей системы менеджмента качества ТСОН. Анализ требований к показателям надежности современных отечественных ТСОН показывает, что требования к таким показателям, как долговечность (срок активного существования (САС) или ресурс), показатель безотказности (вероятностъ безотказной работы (ВБР)), показатель интенсивности отказов и другие постоянно растут.

Оценка надежности осуществляется на основе расчета функций надежности оборудования ТСОН по информации, имеющейся на момент оценивания. Расчет функций надежности может проводиться: расчетно-экспериментальным методом, при котором показатели надежности определяются по результатам заводских, отработочных и приемочных испытаний, а итоговый показатель надежности вычисляется по математической модели для ВБР); экспериментальным методом на основе статистической обработки отказов, выявленных при заводских, отработочных и приемочных испытаниях, по различным признакам деления и по результатам оценки надежности (эффективности) различных видов испытаний оборудования, включая дополнительные отбраковочные испытания; с помощью имитационной модели функционирования системы.

Существуют трудности расчета показателей надежности. К ним можно отнести: для большинства входящих элементов показатели надёжности не заданы; значения надёжности отдельных элементов редуцированы к идеализированным (не учитывается фактическое распределение отказов); недостаточно изучены либо идеализированы механизмы деградации элементов под влиянием различных внешних и внутренних факторов; не осуществляется оценка функционального резервирования; суще-

ствуют трудности расчета показателей из-за недостаточной базы накопленной статистики об отказах и сбоях оборудования в течение срока всего существования; накопление статистики усложняется тем, что ТСОН в своем большинстве являются уникальными по технической и функциональной сложности объектами, срок активного существования слишком большой и полученная статистика может быть неактуальна для новых видов аппаратуры с более жесткими требованиями к функционированию и надежности.

Существует множество отраслевых методик анализа надежности. В ГОСТ 51901.5-2005 проведен краткий обзор часто используемых методов анализа надежности, существующих методик и разработки новых подходов для решения поставленных задач.

Для расчета надежности существует строгая математическая теория [12], и на основе этой теории алгоритмы вычисления надежности $P(t)$ в некоторый момент времени t. $P(t)$ можно представить в виде функциональной зависимости

$$P(t) = f(t, k),$$

где k — вектор параметров, который в общем случае может меняться во времени. Считается, что значения k известны. На практике это зачастую не так.

Сторонники данного подхода считают, что заводы-изготовители должны определять величины k на заводских испытательных стендах или из статистики по множеству производств. Зачастую условия испытаний на заводских стендах не совпадают с условиями эксплуатации действующей системы. При этом элементный подход к расчету надежности технической системы предполагает взаимную независимость отказов элементов самой системы, они являются полем элементарных случайных событий и ищется вероятность сложного случайного события — отказа технической системы.Отметим, что состояние любого элемента системы функционально взаимосвязано через законы сохранения с состоянием всех других элементов системы. Практика использования одного и того же вида оборудования в разных системах показывает, что они работают совершенно по-разному, с разными вероятностями отказов. Более того, один и тот же вид оборудования в разных местах одной и той же технологической системы (орбитах, группировках) также работает по-разному.

Кроме того, динамика интенсивности отказов (λ-распределение) может иметь сложную форму и явно не описываться только одним значением $P(t)$. Каждая подсистема обычно имеет собственное нетривиальное

распределение отказов. Редуцирование каждого распределения к одному значению безусловно ведёт к накоплению погрешностей при дальнейших расчетах.

Экспериментальные методы являются, по сути, единственным способом получения окончательного ответа на вопросы о правильности выполненной разработки системы, достигнутых результатов, реально достигнутом уровне надежности созданной системы. Экспериментальная оценка надежности технической системы может реализовываться в двух вариантах: во-первых, организация специальных испытаний и, во-вторых, сбор статистических данных о работе системы в условиях нормальной работы или подконтрольной эксплуатации. Аналитические методы дают возможность оценивать надежность и проводить сравнение различных вариантов исполнения технических систем и находить оптимальные решения на самых ранних этапах проектирования. К аналитическим методам — по постановке задачи — очень близок метод статистического моделирования. Сходство в том, что и тот, и другой требуют наличия данных о надежности компонентов системы. Однако способы получения результатов существенно различаются. Метод статистического (имитационного) моделирования состоит в генерировании (с помощью специальных генераторов случайных чисел) случайных отрезков времени безотказной работы и времени восстановления отдельных компонентов технической системы и «искусственном» воспроизведении таким образом процесса функционирования данной системы. В результате статистического моделирования системы получается серия частных значений искомых показателей надежности. Эти значения обрабатываются и классифицируются методами математической статистики, что позволяет получить сведения о надежности реальной системы в произвольные моменты времени. Если количество реализаций N достаточно велико, то результаты моделирования системы приобретают статистическую устойчивость и могут быть приняты в качестве оценок искомых показателей надежности. Теоретической основой метода статистического моделирования на ЭВМ являются предельные теоремы теории вероятностей.

Принципиальное значение предельных теорем состоит в том, что они гарантируют высокое качество статистических оценок показателей надежности при весьма большом числе испытаний (реализаций). В условиях уникальных ТСОН число испытаний ограничено, а имеющаяся информация не достаточна для получения гарантированных оценок показателей надежности. Оценка надежности многих видов оборудования

осуществляется индивидуально для каждого экземпляра по результатам периодических обследований.

Для таких систем необходимо применять методы, которые позволяют получать гарантированные оценки в условиях недостаточности информации и малых выборок. В работах О. В. Абрамова, А. Н. Розенбаума, В. В. Климченко отмечается, что уровень теоретических и прикладных исследований, относящихся к области управления надежностью ТСОН, еще далек от существующих потребностей [1]. Подавляющее большинство работ, посвященных проблеме учета постепенных отказов на стадии эксплуатации технических объектов, не учитывает реалии имеющегося информационного обеспечения. Практически не изученными остались вопросы гарантии безотказности и высокой эффективности функционирования технических объектов в условиях ограниченности и неопределенности исходной совокупности сведений, в том числе задачи построения критериев оптимальности для управления в таких условиях, задачи рациональной обработки имеющейся совокупности сведений для предсказания поведения объекта в будущем без всякого домысливания, а также задачи определения стратегии управления, гарантирующей безотказность и высокую эффективность функционирования технических объектов.

В работе [1] отмечается что основные трудности при решении задачи прогнозирования для синтеза стратегии эксплуатации по состоянию связаны с тем, что прогноз приходится осуществлять для каждого объекта индивидуально, при малых объемах исходной информации (по небольшому набору результатов контроля) и в присутствии помех (ошибок контроля), статистические свойства которых достоверно не известны. В этих условиях классические методы математической статистики и теории случайных процессов теряют свои привлекательные свойства, а их использование для прогнозирования приводит к существенным ошибкам и невысокой достоверности прогноза.

Известны некоторые подходы к решению задачи индивидуального прогнозирования и планирования эксплуатации при дефиците и неполной достоверности исходной информации, позволяющие получать в этих условиях достаточно надежные результаты. К их числу относится метод индивидуального гарантированного прогноза (МГП). Основная его идея состоит в том, что из множества возможных реализаций случайного процесса деградации свойств (состояния) исследуемого технического объекта, согласующихся с результатами контроля (не противоречащих им),

выбираются «наихудшие». Под наихудшими понимаются такие, которые за пределами области наблюдения (контроля) идут выше (ниже) всех остальных. Такие реализации можно называть экстремальными.

Как показано в работе [1], если в качестве модели случайного процесса изменения параметров состояния исследуемого технического объекта может быть принята структура в виде полинома Чебышева со случайными коэффициентами, то наихудшими реализациями будут экстремальные полиномы Карлина. МГП дает возможность определить некоторую область, в пределах которой гарантированно будут находиться параметры состояния системы в любой заданный момент времени. Анализ публикаций показал, что в настоящее время активно развивается аппарат непараметрического статистического оценивания. Так в работе [3] описана методика непараметрического оценивания показателей надежности уникального высоконадежного оборудования. Авторы отмечают, что при изучении опыта эксплуатации современного оборудования приходится сталкиваться с определенными особенностями. В частности, необходимо иметь в виду, что современное оборудование относится к классу высоконадежного. Его отказы — события редкие. Ввиду этого приходится иметь дело со статистическими данными малого объема. Кроме того, наблюдаемые данные помимо полных наработок (наблюдения, завершившиеся отказом) содержат всевозможную цензурированную информацию (информацию с различного рода неопределенностями). В связи с этим практика современных исследований предъявляет требования наряду с точечным оцениванием характеристик надежности, которое имеет место во многих современных подходах, проводить интервальное оценивание. В работе [3] показана возможность применения непараметрических методов оценки показателей надежности уникального высоконадежного оборудования на основе использования ядерных оценок.

Новым подходом в рамках рассмотренных задач для построения гарантированных оценок показателей надежности ТСОН в условиях ограниченного объема информации является ВВА. На основе ВВА авторами разработана методика, которая может применяться к задачам оценки параметров надежности и оптимизации структуры и режимов функционирования ТСОН. Суть методики состоит в следующем: на первом этапе проводится предварительный анализ и статистическая обработка полученных экспериментальных данных. Предварительная обработка результатов измерений и выбор методов представления необходимы для того, чтобы в дальнейшем поднять уровень достоверности полученных оце-

нок показателей надежности и корректно применять методы численного вероятного анализа для построения законов распределения и функций плотности вероятности для исследуемых показателей. На первом этапе осуществляется анализ исходных данных, в частности, изучается характер имеющейся неопределенности и далее определяется способ представления данных. При создании методологии рассматривались базовые методики оценки показателей надежности. Как правило, для организации расчетов требуется иметь:

– статистику (поток) отказов $\xi_1, \xi_2, \ldots, \xi_n$, полученную эмпирическим путём в процессе эксплуатации аналогов или отдельных составных блоков;

– нормативные показатели надёжности каждого i-го элемента системы (P_i);

– топологические характеристики проектируемой аппаратуры, включая способы подключения резервных комплектов.

На втором этапе производится оценка вероятности отказа оборудования для всего технического устройства. Подход к расчёту ВБР по известным аналитическим формулам считается классическим, и его справедливость во многом подтверждается реальными апостериорными данными по эксплуатации оборудования [3]. Одним из способов повышения достоверности оценок показателей надежности является использование интервального анализа. В этом случае оценки показателей представляются в виде интервалов. Тогда для оценки показателей надежности может быть использована интервальная арифметика [10]. Представленная методика реализует другой подход, основанный на применении ВВА. Для этих целей необходимо преобразовать исходные данные к функциям плотностей вероятности. Численный вероятностный анализ имеет развитую арифметику для работы с плотностями вероятности, представленными гистограммами, кусочно-полиномиальными функциями, гистограммами второго порядка и аналитическими функциями. ВВА позволяет вычислять плотности вероятностей для широкого класса функций от случайных аргументов [76]. В этом случае исходным типом представления вероятностных оценок надежности будет плотность вероятности, представленная различными способами, например в виде кусочно-полиномиальной функции, аналитической функции или в виде гистограмм второго порядка. Рассмотрим подробнее применение ВВА и численной арифметики для оценки значения ВБР. Данный этап позволяет выбрать соответствующий метод вычисления показателей надежности. В случае достаточ-

ного объема выборок для построения функции плотности вероятности используют гистограммы. Однако в случае уникального оборудования объем выборок весьма невелик. В этих случаях использование гистограмм очень ограничено, поэтому в качестве альтернативы применяют ядерные оценки, которые впервые были введены в работах E. Parzen и M. Rozenblatt. С целью повышения достоверности оценок показателей надежности методология предлагает подход, основанный на сглаживании эмпирической функции распределения. Третий этап представляет собой анализ результатов и использование визуально-интерактивного моделирования (ВИМ). В рамках ВИМ реализуется графическое представление полученных результатов, возможность интерактивной настройки модели [14]. Далее сформулированы некоторые требования к ВИМ-системе. Такая система должна позволять: создавать компьютерные модели специалисту в предметной области, не являющемуся специалистом в моделировании; создавать компьютерные модели сложных систем из моделей объектов, входящих в состав этих систем; скрывать аспекты ВИМ от конечного пользователя, оставляя лишь возможность настройки параметров, относящихся к компьютерным моделям объектов, понятных специалисту в предметной области; создавать компьютерные модели быстро и предоставлять возможность пользователю сосредоточиться на самой проблеме, а не на способах создания модели; осуществлять визуальный контроль исполнения модели; осуществлять анализ статистических данных и представлять их в виде, ожидаемом пользователем модели. Данная методика оценки интенсивности отказов оборудования ответственного назначения позволяет строить достоверные оценки показателей надежности сложных технических систем в условиях малых выборок. Достоверность получаемых оценок показателей надежности обеспечивается процедурами вычисления их вероятных расширений и алгоритмами построения эмпирических законов функций распределения. Предложенный подход может использоваться для оценки рисков в различных сложных технических системах в условиях ограниченного объема информации и ответственного функционирования.

10.3. Оценка показателей надежности

В параграфе рассмотрен подход построения достоверных оценок показателей надежности оборудования в условиях малых выборок статистических данных об отказах. Подход основан на использовании порядковых

статистик и случайных интерполяционных многочленов для построения достоверных оценок функций распределения.

Надежность — важный показатель любых технических систем. Её роль повышается, когда проектируется высокотехнологичное оборудование специального назначения. Например, в техническом задании на этапе проектирования оборудования космических аппаратов должен быть определен гарантированный срок его активного существования. Но такое оборудование не имеет серийного производства и данные об отказах достаточно малы. Несмотря на это, повышение точности оценки вероятности безотказной работы остаётся одной из важнейших проблем проектировщиков систем ответственного назначения [3].

Повышение надёжности безотказной работы оборудования может быть достигнуто за счёт проектных решений (структурное и функциональное резервирование), совершенствования технологических процессов и применения современных численных методов для оценки значений вероятности отказа оборудования.

Важным аспектом при анализе состояний объектов ответственного назначения по эмпирическим данным является требование получения надежных оценок изучаемых показателей как результатов математических вычислений. Эта надежность должна обеспечиваться и рядом других средств, как собственно математических, так и относящихся к компьютерному инструментарию для решения математических задач. К надежности имеют прямое отношение и вопросы корректной постановки вычислительной задачи, и обеспечение связи численных расчетов и аналитических (символьных) преобразований, и выбор надлежащей машинной архитектуры, позволяющей получить нужный результат с приемлемыми затратами вычислительных ресурсов, и выразительность языковых средств для описания алгоритмов, и многое другое.

В деле практической реализации идеи надежности применительно к численной стороне вычислений важную роль сыграли достижения интервальной математики. Корректное интервальное вычисление гарантирует выполнение важнейших свойств численного решения и прежде всего — его локализацию.

Следует отметить, что значительную часть производимых сегодня в мире вычислений нельзя назвать надежными, поскольку методы обеспечения надежности еще не получили должного распространения, а после выполнения обычных вычислений пользователи зачастую не могут сказать ничего определенного относительно важнейших свойств полученно-

го решения, в том числе и его точности.

В настоящее время для анализа состояний сложных технических систем применяются параметрический и непараметрический подход. Каждый из этих методов имеет свои плюсы и минусы. Так использование параметрических методов требует предположений о виде закона распределения наблюдаемых величин. Заметим, что во многих случаях достаточно сложно найти убедительные доказательства, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. В настоящее время для решения задач анализа статистической информации развиваются непараметрические методы, в частности, методы ядерного оценивания.

Интенсивность отказов

Вероятность безотказной работы $P(t)$ — это вероятность того, что в течение указанного времени работы не произойдет отказа. Время работы — это продолжительность, или объем работы. Частота отказов — это мера отказов за единицу времени. Частота отказов зависит от распределения отказов, которое представляет собой совокупную функцию распределения, которая описывает вероятность отказов до момента времени t

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \le \xi < t + \Delta t | t \le \xi)}{\Delta t} = \frac{f(t)}{P(t)} = \frac{f(t)}{1 - F(t)}.$$

Заметим, что $f(t) = P'(t)$ и

$$\lambda(t) = -\frac{P'(t)}{P(t)}, \tag{10.2}$$

где $P(t)$ — вероятность безотказной работы.

Пусть $(\xi_1, \xi_2, \ldots, \xi_n)$ — статистика отказов, полученная опытным путем.

Тогда

$$-\ln(z_i) = \int_0^{\xi_i} \lambda(\xi) d\xi,$$

где $z_i = P(\xi_i)$.

Для нахождения $\lambda(t)$ будем использовать метод наименьших квадратов. Представим аппроксимацию $\lambda(t)$ в виде случайной кусочно-линейной функции $l(t)$

$$\lambda(t) \approx l(t) = \sum_{i=1}^{m} a_i \psi_i(t), \quad t \in [0, T],$$

где $\{\psi_i, i = 1, \ldots, m\}$ — базис в пространстве кусочно-линейных функций, $m \leq n$. Пусть $\{0 = t_1 < t_2 < \ldots < t_m = T\}$ — сетка и $\{\psi_i, i = 1, \ldots, m\}$ определим следующим образом:

$$\psi_i(t_l) = \begin{cases} 1, & l = i, \\ 0, & l \neq i. \end{cases}$$

Тогда

$$\int_0^t \lambda(\xi)d\xi \approx \int_0^t l(\xi)d\xi = \sum_{i=1}^m a_i\varphi_i(t),$$

$$\varphi_i(t) = \int_0^t \psi_i(\xi)d\xi.$$

Для нахождения $a_1, a_2, \ldots, a_m$ рассмотрим задачу минимизации функционала

$$\Phi(a_1, \ldots, a_m) = \sum_{i=1}^n (-\ln(z_i) - \sum_{j=1}^m a_j\varphi_j(\xi_i))^2 \to \min.$$

Задача сводится к решению СЛАУ

$$G\vec{a} = b,$$

где $G = (g_{ij})$ — матрица Грама, $\vec{a} = (a_1, a_2, \ldots, a_m)$, $b = (b_i)$, $Z(\xi_i) = -\ln(z_i)$, $g_{ij} = (\varphi_i, \varphi_j)$, $b_i = (Z, \varphi_i)$ и

$$(x, y) = \sum_{i=1}^n x(\xi_i)y(\xi_i).$$

Заметим, что G — детерминированная (неслучайная) матрица. Пусть $B = G^{-1}$ — обратная матрица и b_{ij} — элементы матрицы B. Тогда

$$a_i = \sum_{j=1}^m b_{ij} \left(\sum_{l=1}^n (-\ln(z_l)\varphi_j(\xi_l)) \right)$$

или

$$a_i = \sum_{l=1}^n \gamma_l \ln(z_l),$$

где

$$\gamma_l = \sum_{j=1}^m b_{ij}\varphi_j(\xi_l),$$

$$\lambda(t) \approx \sum_{l=1}^{n} \left(\sum_{j=1}^{m} \psi_j(t) \right) \gamma_l \ln(z_l).$$

Используя вместо $z_1, z_2, \ldots, z_n$ совместную функцию плотности $p(z_1, z_2, \ldots, z_n)$, можем построить вероятностное расширение $l(t)$.

$$\boldsymbol{l}(t) = \sum_{l=1}^{n} \left(\sum_{j=1}^{m} \psi_j(t) \right) \gamma_l \ln(\boldsymbol{z}_l).$$

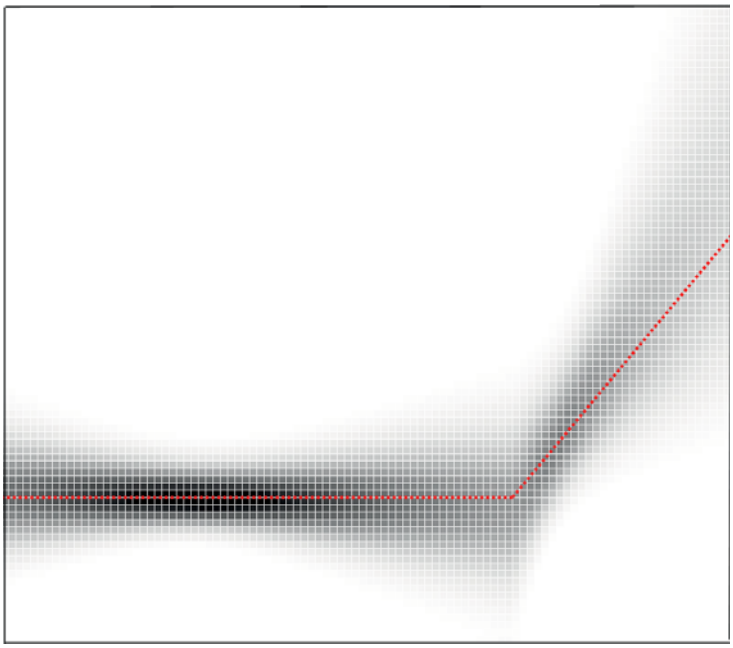


Рис. 10.4. Достоверные оценки интенсивности отказов

Пример 13. Имеем $(\xi_1, \xi_2, \ldots, \xi_n)$, $n = 29$. Предположим, $\lambda(t)$ имеет вид

$$\lambda(t) = \begin{cases} 0.2, & t \in [0, 0.7] \\ 0.2 + 12(t - 0.7) & t > 0.7 \end{cases}$$

Представим $\lambda(t)$ случайную кусочно-линейную функцию

$$\boldsymbol{l}(t) = \sum \boldsymbol{a}_i \psi_i(t),$$

$\{\psi_i\}$ — *базис в пространстве кусочно-линейных функций.*

$$\int_0^t \lambda(\xi) d\xi \approx \sum_{i=1}^{m} a_i \varphi_i(t),$$

где

$$\varphi_i(t) = \int_0^t \psi_i(t)dt.$$

В силу детерминированности матрицы МНК G и того факта, что компоненты вектора $\boldsymbol{b}$ есть линейные комбинации $\ln z_i$, вектор $\vec{\boldsymbol{a}} = G^{-1}\boldsymbol{b}$ *можно выразить как линейную комбинацию* $\ln z_i$:

$$\lambda(t) \approx \sum_{i=1}^{n} \gamma_i(t) \ln z_i,$$

где $\gamma_i(t)$ — некоторые вещественные функции.

На рис. 10.4 оттенки серого представляют плотность вероятности случайной кусочно-линейной функции $\boldsymbol{l}(t)$, точечная линия — $\lambda(t)$.

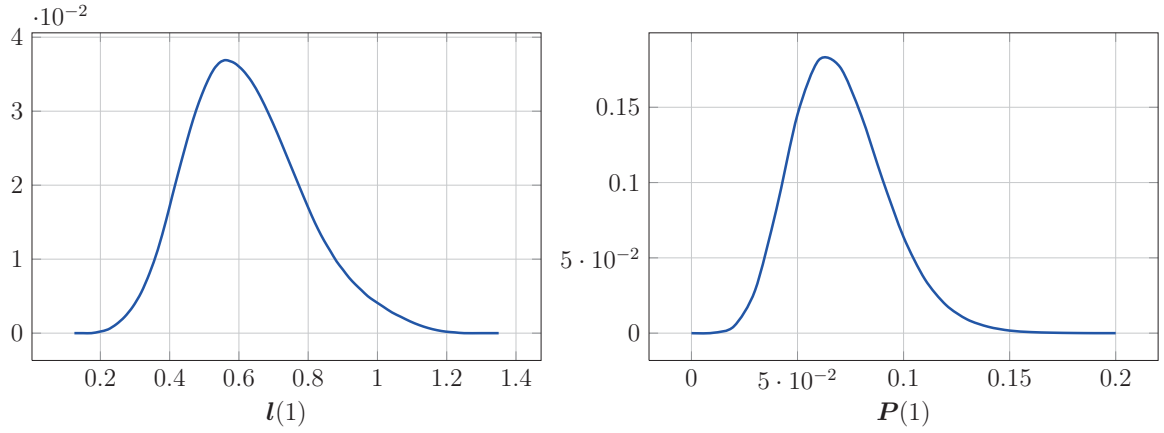


Рис. 10.5. Аппроксимация функций плотности вероятности $\boldsymbol{l}(1)$ и $\boldsymbol{P}(1)$

Используя вероятностное расширение λ, можно вычислить значения оценки функции плотности вероятности $\boldsymbol{P}(t)$ в любое время. На рис. 10.5 показана оценка функций плотности вероятности $\boldsymbol{l}(1)$ и $\boldsymbol{P}(t)$ в момент времени $t = 1$.

Используя оценки функции плотности вероятности $\boldsymbol{P}(t)$ в виде сплайнов, мы можем, например, оценить риск того, что вероятность $P(1) > 0.2$ или $P(1) < 0.05$.

Подводя итог, отметим, что проблема вычисления надежных оценок эмпирических функций распределения в условиях малой выборки и неопределенности данных может быть изучена на основе случайных интерполяционных полиномов и порядковых статистик. В рамках BBA используются арифметические операции над функциями плотности вероятности и процедуры построения вероятностных расширений. Вычислительный

вероятностный анализ обеспечивает надежное вычисление в тех случаях, когда функция плотности вероятности неизвестна, и может найти не только границы области решения, но и идентифицировать ее вероятностную структуру. Использование этого подхода для оценки параметров технических систем позволяет получить не только математические ожидания или границы этих параметров, но и оценки их внутренних распределений. Этот подход может быть использован для оценки показателей надежности специального оборудования в условиях небольшого количества статистических данных о неисправностях и для расчета рисков технических систем в условиях ограниченной эмпирической информации.

10.4. Обработка и анализ гидрологических данных спутникового мониторинга

В параграфе рассматриваются новые подходы к исследованию гидрологических данных спутникового мониторинга, которые позволяют проводить анализ изучаемых объектов и делать более точные прогнозы.

Применение методов космического и наземного мониторинга для изучения природных и других процессов сегодня сложно переоценить. В последние десятилетия в мире и в России активно обсуждается проблема экстремальных природных событий, составляющих существенную долю неблагоприятных последствий происходящих катаклизмов. В связи с этим актуализировалась проблема оценки рисков возникновения экстремальных природных событий на основе совершенствования существующих методик разработки прогнозов и создания новых информационно-аналитических технологий исследования информационных потоков, сопровождающих эти процессы. Известно, что природные события существенным образом определяются и зависят от процессов, происходящих как в атмосфере, так и на поверхности Земли, и характеризуются масштабностью и динамичностью. Для решения данной проблемы используются различные подходы, в том числе с помощью космических средств наблюдения, позволяющих получать информацию о состоянии поверхности Земли и атмосферы в режиме мониторинга на значительной территории. В этом существенное преимущество космических средств по сравнению с традиционными наземными. Другим важным преимуществом космических средств является оперативность поступления информации.

Важно отметить, что космический мониторинг все чаще и чаще рассматривается исследователями не только как способ регистрации, сбора, передачи, накопления и хранения полученной информации, но и как информационно-аналитическая технология для анализа информации о качественных и количественных характеристиках состояния атмосферы и поверхности Земли, а также оценки и прогноза тенденций изменения в них.

Одной из разновидностей мониторинга является дистанционное зондирование Земли (ДЗЗ). ДЗЗ — сбор информации с помощью приборов, установленных на вертолетах, самолетах, спутниках. Дистанционное зондирование — это получение информации об объекте по данным измерений, сделанных на расстоянии от объекта, т. е. без прямого контакта с объектом. Важным аспектом ДЗЗ является возможность контроля больших территорий; наблюдения из космоса позволяют получать информацию, обобщающую процессы на уровне региона, континента, планеты Земля. По некоторым оценкам данные из космоса стоят дешевле наземных. Возможности космического ДЗЗ позволяют получать информацию, например, о температуре поверхности Земли с разрешением в несколько метров. Информация, полученная на основе космического мониторинга, характеризуется большими объемами, различными способами регистрации и формами передачи, которую надо соответствующим образом представить и обработать. Процесс обработки и подготовки спутниковых данных для исследования природных процессов включает в себя ряд вычислительных процедур, которые должны отвечать системе требований, среди которых в первую очередь необходимо отметить снижение уровня неопределенности в данных, достоверность и наглядность полученных результатов.

Следует отметить, что использование данных только космического мониторинга не всегда достаточно и оправданно, поскольку наземные измерения также привносят свою долю достоверности о природных процессах. Поэтому к формированию базы данных для прогнозов природных явлений необходимо подходить системно, не упуская ни одной возможности повысить оправдываемость прогнозных выводов. С этой точки зрения в качестве примера можно рассмотреть автоматизированную технологию мониторинга весеннего половодья на сибирских реках [4]. Созданная информационная технология предусматривает оптимизацию параметров математической модели прогноза в ходе ее применения. Для этого используется электронная база многолетних гидрометеорологиче-

ских данных наблюдений, обеспечивающая автоматизированный перебор и моделирование возможных гидрометеорологических ситуаций. База включает ежедневные данные наблюдений десятков гидрологических постов и метеорологических станций за последние 19 лет, спутниковые данные, отличающиеся большими объёмами и многомерностью, и атрибутивные данные, содержащие информацию о ландшафтно-гидрологических особенностях районов бассейнов рек.

Концептуально-гистограммный подход к моделированию природных явлений. Важным аспектом при изучении природных явлений является выбор прогнозной модели и методов представления и расчета соответствующих характеристик. Остановимся подробнее на моделях. В настоящее время для исследования природных явлений развивается подход к моделированию, включающий решение функциональных уравнений и систем уравнений, в основу которых положены различные концепции описания физических процессов. Обычно такие модели называют концептуальными. Одним из наиболее трудных аспектов применения концептуальных моделей является калибровка выбранной модели применительно к конкретному объекту исследования. Большинство параметров модели определяются итерационным способом, вручную или автоматически, на основе исторических рядов входных и выходных данных. Из-за ограниченности и неопределенности данных, несовершенства модели и наличия внутренних связей между параметрами даже небольшое увеличение их количества способно значительно повысить трудности, связанные с калибровкой модели. Поэтому необходимо, чтобы число параметров соответствовало степени достоверности исходных данных и требуемой точности.

В качестве примера рассмотрим упрощенную модель HBV, разработанную S. Bergstrom в Шведском метеорологическом и гидрологическом институте. Она представляет собой концептуальную модель водосбора, которая преобразует осадки, температуру воздуха и потенциальное суммарное испарение в сток или приток в водохранилище. Модель описывает общий баланс воды следующим образом:

$$p - e - q = \frac{d}{dt}(sp + sm + uz + lz + vl),$$

где p — осадки, e — суммарное испарение, q — сток, sp — снежный покров, sm — влажность почвы, vl — объем озер. Модель HBV можно рассматривать как детерминированную модель с полураспределенными параметрами; водосбор разбивается на частные водосборы, также при-

меняется метод высотного районирования. Для водосборов определенного высотного положения осуществляется дополнительное деление на высотные зоны. Каждую высотную зону можно подразделить на подзоны по типу растительности, например лесные и не лесные территории. Последовательность формирования стока представляет собой функцию реагирования, преобразующую избыточную почвенную влагу в сток. Она также учитывает осадки, выпадающие непосредственно на поверхность озер, рек и других увлажненных территорий, и испарение с них. Необходимой входной информацией для модели являются количество осадков (суточные суммы), температура воздуха (среднесуточные значения) и оценки возможного суммарного испарения. В качестве альтернативы суточные значения можно рассчитать как пропорциональные температуре воздуха, но с коэффициентами пропорциональности ежемесячных значений. Более поздние версии модели HBV могут работать с данными более высокого временного разрешения, т. е. ежечасными данными.

Последние 20 лет ознаменовались интенсивным развитием методов исследования гидрологических процессов с помощью космических средств наблюдения. С помощью таких средств могут быть получены гидрологические данные на весьма значительных площадях с хорошим разрешением и точностью.

Космический мониторинг — это регистрация, сбор, передача, накопление, хранение и анализ информации о качественных и количественных характеристиках состояния атмосферы и поверхности Земли, а также оценка и прогноз тенденций изменения в них.

Система мониторинга состоит из следующих подсистем: подсистема сбора информации, подсистема хранения, подсистема анализа данных, подсистема принятия решений, подсистема выводов результатов. Важное значение для гидрологического мониторинга имеет дистанционное зондирование Земли (ДЗЗ). ДЗЗ — сбор информации с помощью приборов, установленных на вертолетах, самолетах, спутниках.

Процессы, происходящие в атмосфере и на поверхности Земли в целом, характеризуются масштабностью и динамичностью; в этой связи наземные методы сбора информации зачастую не позволяют получать данные с требуемой оперативностью и точностью. Зондирование из космоса дает информацию, в зависимости от используемой системы, с периодичностью до суток и даже часов.

Существенно, что данные из космоса стоят дешевле наземных. Таким образом, в настоящее время ДЗЗ представляет основной метод монито-

ринга.

Для гидрологических прогнозов необходимо знать информацию о состоянии поверхности Земли и атмосферы на значительной территории.

Возможности космического ДЗЗ позволяют получать информацию, например, о температуре поверхности Земли с разрешением в несколько метров. Существующие прогнозные гидрологические модели, как правило, используют значения входных параметров в ограниченном числе точек на поверхности Земли. Рассмотрим один из новых способов сбора и обработки гидрологических данных, основанный на ВВА.

Представим гидрологическую модель в виде функциональной зависимости

$$y = f(x, k),$$

где y — результат прогноза, например боковой приток, x — вектор входных значений, например температура, осадки и т. п., k — вектор параметров модели. Предположим, что мониторинг ведется по области Ω, которую представим как объединение подобластей Ω_i. ДЗЗ позволяет для каждой подобласти Ω_i представить N_i значений $x_{ij}, j = 1, ..., N_i$ и т. п. Обычно каждый набор x_{ij}, как правило, представляется как среднее значение.

Используя агрегацию, для каждой Ω_i может быть известно не только среднее значение, но и кусочно-полиномиальная функция. На рис. 10.6 приведен пример построения P_i по значениям в некоторой области Ω_i.

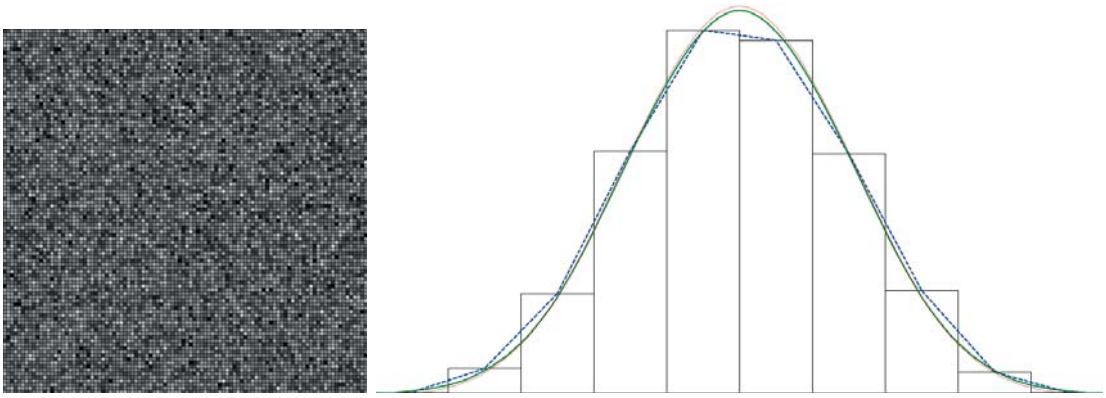


Рис. 10.6. Кусочно-полиномиальная агрегация данных ДЗЗ

Область Ω_i представляет прямоугольник, состоящий из 100×100 пикселей, каждому пикселю сопоставлено значение t_i. Для наглядности на рис. 10.6 значения t_i представлены оттенками серого цвета. Более светлые цвета соответствуют более высокой температуре. Таким образом, в

правой части рис. 10.6 представлены кусочно-полиномиальные аппроксимации функции плотности вероятности — гистограмма, частотный полигон, сплайн.

Рассмотренный подход демонстрирует эффективный способ агрегирования данных. Так, вместо 10^4 значений, представленных в подобласти Ω_i, для определения кусочно-полиномиальной функции используется лишь порядка 10^2 значений. Более того, построенное кусочно-полиномиальное приближение несет в себе массу дополнительной информации: интервал изменения значений, среднее значение, частоты и т. п.

Такой подход можно интерпретировать как построение функции плотности вероятности некоторой случайной величины. Поскольку изучаемая выходная величина модели y, характеризующая состояние гидрологической системы, часто представляется в виде некоторой функциональной зависимости входных параметров $x_1, x_2, \ldots, x_n$, рассмотрим метод представления y в виде кусочно-полиномиального приближения, т. е. ставится задача: зная кусочно-полиномиальные аппроксимации плотностей вероятности $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, построить плотность вероятности $\boldsymbol{P}y$, аппроксимирующую распределение выходной величины y.

Использование данных, полученных на основе космического и наземного мониторинга, с одной стороны, способствует повышению достоверности в изучении природных явлений и процессов во времени и в пространстве. С другой стороны, возможности космического и наземного мониторинга существенно увеличивают объемы информации, подлежащей обработке и исследованию, что затрудняет, делает не эффективным, а порой невозможным применение существующих численных процедур для построения прогнозных моделей и анализа полученных результатов. Разработанные на основе ВВА процедуры и методы кусочно-полиномиального представления, обработки и численного моделирования выходных параметров концептуальных моделей на основе использования данных космического и наземного мониторинга позволяют снизить уровень неопределенности в информационных потоках, существенно сократить время обработки и выполнения численных процедур. Данный подход позволяет также в режиме визуально-интерактивного моделирования представить необходимые данные для оперативного принятия решений.

10.5. Оптимизация выработки электроэнергии гидроэлектростанцией в условиях неопределенности

Задачи гидроэнергетики характеризуются высоким уровнем неопределенности, которая проявляется на всех стадиях информационного процесса принятия управленческого решения. Поэтому поиск методов и подходов к построению эффективных решений в условиях неопределенности является важной и практически значимой задачей. Решением разнообразных задач со стохастическими неопределенностями в данных занимается стохастическая гидрология. Для решения оптимизационных задач со стохастическими входными данными используется аппарат стохастического программирования. Особое место занимают оптимизационные задачи с неопределенными входными данными. В случае, когда входные параметры содержат различные типы неопределенности, используется математический аппарат неопределенного программирования [95]. Следует отметить работы в области интервальных неопределенностей [29, 47]. Работа [47] посвящена задачам линейной оптимизации в условиях интервальной и нечеткой неопределенности.

Вычислительный вероятностный анализ направлен прежде всего на разработку численных процедур и методов, способствующих снижению уровня неопределенности в зависимости от типа, характера, специфических особенностей, объема и ее источников на всех стадиях информационного процесса, сопровождающего принятие управленческого решения. В параграфе для решения оптимизационных задач гидроэнергетики со случайными входными данными предлагается использовать аппарат случайного программирования [113, 34, 69] (глава 8). Предполагается, что техника регрессионного анализа над эмпирическими распределениями (глава 9) позволяет построить соответствующие регрессионные модели притока воды в водохранилище.

Постановка задачи. Мощность p выработки электроэнергии ГЭС можно представить в виде [49]

$$p = Chu,$$

где C — некоторая константа; h — высота уровня воды в водохранилище, $h \in [h_{\min}, h_{\max}]$, u — количество воды, проходящей через турбины в единицу времени, $u \in [u_{\min}, u_{\max}]$.

Высота уровня воды h зависит от объема воды в водохранилище V:

$$h = h(V).$$

Объем воды в водохранилище $V(t)$, в свою очередь, зависит от $u(t)$, притока воды в водохранилище $q(t)$ и $u_x(t)$ — холостого сброса:

$$V(t) = V_0 + \int_0^t q(\xi) - u(\xi) - u_x(\xi) d\xi.$$

Пусть необходимо максимизировать выработку электроэнергии на временном отрезке $[0, T]$. Ставится задача оптимального управления

$$P(u) = \int_0^T C\ h\left(V_0 + \int_0^T q(t) - u(t) - u_x(t) dt\right)\ u(t) dt \to \max,$$

где u — управление, количество воды, проходящей через турбины в единицу времени, $u \in [u_{\min}, u_{\max}]$.

Упростим задачу, представим объем воды в водохранилище V в виде

$$V(t) = V_0 + S(h(t) - h_0),$$

где V_0 и h_0 — объем и уровень воды в водохранилище в момент времени $t = 0$ соответственно. Уровень воды h в водохранилище зависит от $u(t)$, $q(t)$ и $u_x(t)$:

$$h(t) = h_0 + (V(t) - V_0)/S = h_0 + (\int_0^t q(\xi) - u(\xi) - u_x(\xi) d\xi)/S.$$

Таким образом,

$$P(u) = C \int_0^T \left(h_0 + (\int_0^t q(\xi) - u(\xi) - u_x(\xi) d\xi)/S\right)\ u(t) dt \to \max, \tag{10.3}$$

где $q(t)$ — приток воды в водохранилище; $u_x(t)$ — холостой сброс; u — количество воды, проходящей через турбины, $u \in [u_{\min}, u_{\max}]$.

Дискретная модель. Рассмотрим дискретное приближение для модели (10.3). Это позволит свести решение исходной задачи к решению системы линейных алгебраических уравнений. Для этих целей построим на отрезке $[0, T]$ сетку: $\omega = \{t_0 < t_1 < \ldots < t_n\}$, приток воды в водохранилище за время $[t_{i-1}, t_i]$ приблизим гистограммой q_i, соответственно $\{u_{xi} | t \in [t_{i-1}, t_i]\}$ — гистограммы холостого сброса за время $[t_{i-1}, t_i]$,

$U = \{u_i | t \in [t_{i-1}, t_i]\}$ — гистограммы количества воды через турбины за время $[t_{i-1}, t_i]$:

$$P(U) = C \sum_{i=1}^{n} \left(h_0 + (\sum_{j=1}^{i} q_j - u_j - u_{xj})/S \right) \; u_i \to \max .$$

Задачу в случае известного q_x можно свести к решению системы линейных алгебраических уравнений

$$AU = b,$$

где U — вектор решения, $A = (a_{ij})$, $b = (b_i)$ — матрица и вектор правой части.

В нашем случае

$$2u_1 + u_2 + \ldots + u_n = h_0 + q_1,$$

$$u_1 + 2u_2 + \ldots + u_i + \ldots + u_n = h_0 + q_1 + q_2,$$

$$u_1 + u_2 + \ldots + 2u_i + \ldots + u_n = h_0 + \sum_{j=1}^{i} q_j,$$

$$u_1 + u_2 + \ldots + 2u_n = h_0 + \sum_{j=1}^{n} q_j.$$

В этом случае оптимальное количество воды, проходящей через турбины u_i, $i = 1, \ldots, n$, можно представить в виде линейных комбинаций от q_i, $i = 1, \ldots, n$.

В данной модели основная входная информация $q(t)$ — приток воды в водохранилище. В силу многих причин функция $q(t)$ носит случайный характер. Тогда мы получаем систему линейных алгебраических уравнений со случайной правой частью. Методы решения подобных систем подробно рассмотрены в главе 6.

Важное значение имеет способ представления $q(t)$ и соответственно кусочно-полиномиальных аппроксимаций q_i, $i + 1, ..., n$. Для этих целей на практике предлагается использовать регрессионные модели над эмпирическими распределениями (глава 9).

Для вычисления линейных комбинаций от q_i, представленных своими функциями плотности вероятности, будем использовать численные вероятностные арифметики.

Численный пример. Рассмотрим численное решение дискретной модели. Пусть $q_i \in [\underline{q}_i, \overline{q}_i]$ — равномерные случайные величины. Для определенности $n = 3$, $S = 1$, носители $q_1 = [0.1, 0.2]$, $q_2 = [0.2, 0.3]$, $q_3 = [0.3, 0.4]$, $h_0 = 0.9$.

При $n = 3$, решив систему линейных алгебраических уравнений, в силу детерминированности матрицы получаем оптимальное количество воды, проходящей через турбины в виде линейной комбинации притоков воды:

$$u_1 = \frac{-q_3 - 2q_2 + q_1 + h_0}{4},$$
$$u_2 = \frac{-q_3 + 2q_2 + q_1 + h_0}{4},$$
$$u_3 = \frac{3q_3 + 2q_2 + q_1 + h_0}{4}.$$

Таким образом, для нахождения u_i мы можем использовать численную вероятностную арифметику.

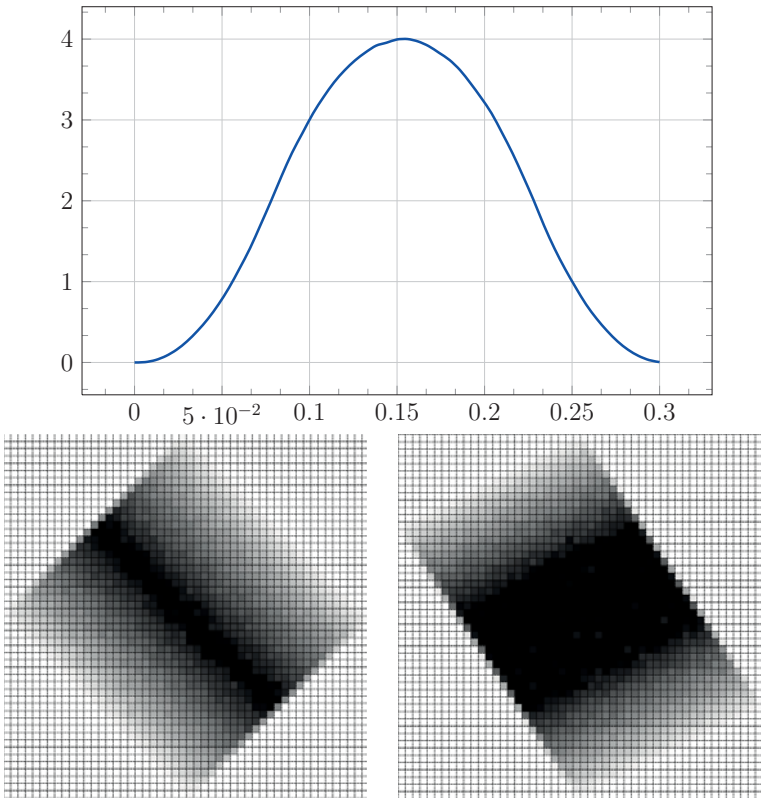


Рис. 10.7. Функция плотности вероятности u_1 и совместные плотности вероятности для (u_1, u_2), (u_2, u_3)

На рис. 10.7 представлена аппроксимация сплайнами плотности вероятности решения u_1, значения совместных плотностей вероятности для

(u_1, u_2), (u_2, u_3) представлены оттенками серого. Носители $u_1 = [0.0, 0.1]$ и $u_2 = [0.25, 0.35]$, $u_3 = [0.575, 0.725]$. Лицу, принимающему решения, представляется визуальная информация о плотности вероятности компонент решения и их совместных плотностей вероятности, что значительно облегчает принятие решений.

В параграфе рассмотрена оптимизация выработки электроэнергии гидроэлектростанцией в условиях случайных входных данных, в частности, случайного характера притока воды в водохранилище. В качестве решения использовалось количество воды, проходящей через турбины. В результате построены плотности вероятности решений, которые представлены в виде кусочно-полиномиальных функций. Далее эта информация о решении может быть использована для оценки рисков.

10.6. Технология извлечения и визуализации знаний

Эффективность принимаемых решений в большей степени зависит от изменчивости, взаимозависимости и неопределенности входных и выходных параметров, характеризующих объект исследования. В настоящее время разработано достаточно технологий извлечения знаний из данных для поддержки принятия решений. Достаточно назвать технологии KDD, Data Mining [144], VIS-технологии и другие. Однако их применение большей частью ограничено и является привилегией нескольких, чем большинства. Причин этому несколько. Одной из основных причин «плохого» распространения и внедрения в практику решения реальных задач является отсутствие доверия и адекватного восприятия к результатам моделирования лицами, принимающими решения. В статье рассматривается технология поддержки принятия решений, которая сочетает два аспекта. Первый аспект связан с процессом численного представления, расчета, анализа и извлечения знаний из неопределенной информации, которая носит неточный или случайный характер. Второй аспект связан с процессом визуализации полученных знаний в интерактивном режиме графического представления и анализа результатов моделирования для взаимодействия с лицом, принимающим решение. Такой подход позволяет потенциально улучшить процесс моделирования при решении задач принятия решений, повысить эффективность «общения» и «взаимопонимания» между пользователем и моделью процесса, а также реализовать необходимость для создания альтернативных возможностей для выбора наиболее подходящего решения.

Наличие неопределенностей во входных данных при решении многих практических задач приводит к необходимости создания методов, учитывающих эти неопределенности. Так, интервальная неопределенность привела к развитию интервальных методов. Сами интервальные числа при этом можно трактовать как случайные величины, про которые известны лишь границы их изменения. Поэтому интервальная математика сосредоточивает свое внимание на вычислении гарантированных границ множеств решений и не учитывает возможного распределения плотности вероятности полученных решений. Дальнейшие обобщения интервальной математики направлены на определение более детальных характеристик получаемых множеств решений [11].

Первый аспект рассматриваемой технологии реализуется на основе вычислительного вероятностного анализа (BBA). Второй аспект реализует идею визуально-интерактивного моделирования (VIS) и представляет собой метод, который интегрирует математическую и символическую модель при взаимодействии реального времени для графического отображения результатов моделирования.

Рассмотрим основные аспекты предлагаемой технологии на примере задачи принятия согласованного решения в условиях неопределенности входных данных. Для этого представим основные математические методы и способы формализации процесса согласования. Для изучения процесса построения согласованного решения используются различные методы исследования. Одним из таких методов является метод моделирования, который предполагает, как одно из направлений, разработку формальных моделей объекта. Понятие «согласование» можно выразить различными математическими конструкциями. Например, общность условий или ограничений, в рамках которых принимается решение, можно представить как систему линейных или нелинейных уравнений и неравенств, операторных уравнений, балансовых соотношений. Для достижения баланса целей можно использовать методы многокритериальной оптимизации, теорию кооперативных игр и другие.

Оптимизационный подход представляет собой один из наиболее часто применяемых инструментов, используемых для построения моделей принятия решений. Это связано, прежде всего, с тем, что лицо, осуществляющее выбор, стремится максимизировать свои интересы или выгоду. Такое поведение ЛПР подчиняется принципу рациональности. Выбор группы участников в отличие от индивидуального должен подчиняться определенным ограничениям, которые представляют собой балансо-

вые соотношения, в рамках которых группа лиц может рассматриваться как единое целое. Поведение каждого участника группы может описать как некоторую задачу оптимизации, в которой интерес участника представлен в виде соответствующей целевой функции. Так как участников несколько, и все они подчиняются единым балансовым ограничениям, то в случае группового выбора мы имеем систему задач оптимизации. Следует отметить, что модель в виде системы задач оптимизации с балансовыми соотношениями отражает идею компромисса или согласования интересов, ограничений и предпочтений всех участников процесса принятия решений. Данный подход может применяться для согласования решений в различных областях.

Таким образом, при формализации процесса согласования интересов используются различные математические конструкции, позволяющие рассматривать всех лиц, заинтересованных и принимающих участие в решении проблемы с позиции единого целого — системы. Информационной основой для принятия согласованных решений являются данные о значениях соответствующих показателей, многие из которых можно отнести к категории неопределенные данные.

Технология для расчета и визуализации согласованного решения

Рассмотрим задачу согласования интересов в условиях стохастической неопределенности. Важное направление при решении данной задачи — вероятностное представление имеющейся информации.

Рассмотрим метод построения согласованного решения, использующий BBA. Суть метода состоит в построении совместной функции распределения для данной системы показателей, которые могут рассматриваться как случайные величины. Для примера предположим, что модель, определяющая взаимосвязь данных показателей, представляет собой стохастическую систему линейных или нелинейных уравнений

$$f_i(x, k) = 0,\ i = 1, ..., n,$$

$x \in R^n$ — вектор решения, $k \in R^m$ — вектор решения.

Предположим, что мы согласовываем интересы двух сторон. Пусть интересы каждой стороны выражаются в некоторой паре значений (x, y). Обычно, решая проблему согласования различными подходами, мы получаем точку на плоскости $(x, y) \in R^2$. Считается, что это компромисс-

ное решение, которое удовлетворяет обе стороны. При этом в подавляющем количестве случаев предполагается, что входные значения имеют точное значение. В жизни это, безусловно, не так. Пусть неопределенность данных носит вероятностный характер (но может, к примеру, носить и нечеткий характер). Если мы будем перебирать все входные значения, каждое из которых имеет свою вероятность, то будем получать различные точки согласования (x_s, y_s) с той же вероятностью. В результате вместо одной точки получаем целое множество, которое можно представить своей совместной плотностью вероятности.

На рис. 6.11 в главе 6 приведено приближение совместной функции плотности вероятности вектора решения системы нелинейных уравнений. Видим, что есть области наиболее или наименее вероятного принятия решений. Эти области выделены оттенками серого. Каждый оттенок соответствует определенной вероятности попадания в заданную точку области. Например, черный цвет соответствует наибольшему значению вероятности «попадания» решения в данную область, белый соответствует нулевой вероятности. Заметим, что в процессе осуществления расчетов в режиме реального времени формируется таблица значений, по которой и строится графическое представление совместной функции плотности вероятности.

Имея данную информацию, визуализированную в виде соответствующей области вероятностного пространства принятия решений, у ЛПР появляются дополнительные основания для выбора наилучшего согласованного решения. На основе ВВА можно получить также дополнительные вероятностно-статистические характеристики такого множества и использовать их для дальнейшего поиска и обоснования эффективности принимаемого решения.

Таким образом, на основе применения ВВА реализуется второй аспект технологии, который связан с визуальным (графическим) представлением полученных знаний и организацией интерактивного диалога между моделью и ЛПР. Данный метод позволяет в интерактивном режиме вводить новые дополнительные условия и ограничения на значения показателей, интересующих ЛПР, а также на условия поиска согласованного решения (которые часто сложно формализовать и учесть при построении решений известными численными методами), что, безусловно, позволит найти более качественные варианты согласования интересов. Заметим, такое решение несколько отличается от решения, найденного при предположении точных входных данных, но будет более адекватно со-

ответствовать субъективным ожиданиям и объективным предпосылкам каждой стороны процесса принятия согласованного решения.

В заключение отметим, что ведение диалога между ЛПР и моделью предполагает использование динамического дисплея, с помощью которого ЛПР может изменять параметры модели и анализировать их влияние и последствия. Для повышения эффективности и качества принимаемых решений обязательно нужно учитывать специфику неопределенности, содержащуюся в данных, процессах и объектах, составляющих информационное пространство принятия решений. Одним из адекватных методов в решении такого класса задач является ЧВА, предоставляющий исследователю эффективный инструмент для работы с данными в условиях стохастической неопределенности.

10.7. Визуально-интерактивная анимация

За последнее десятилетие анализ многомерных данных стал одним из основных направлений прикладной математики, активно развивающимся и применяющимся практически во всех областях исследований. Неправильные предположения о свойствах данных или их источника, вместе со сложностями зрительной системы человека, могут привести к ложным выводам.

Для визуального исследования данных в компьютерной графике применяются различные приемы, в том числе использование перспективы, удаление скрытых линий и поверхностей, применение стереографических изображений для представления образа объекта, которое обманывает нашу зрительную систему в восприятии пространства, или объемное изображение, нарисованное на явно плоском экране компьютера. Если мы смотрим на изображение в гарнитуре виртуальной реальности, мы можем добавить параллакс движения, который поможет нам воспринимать глубину, даже если дисплей перед нами на самом деле не более чем массив цветных точек [4, 99, 143]. Наше понимание высших измерений ограничено, поэтому наши визуализации будут в основном ограничены тремя измерениями, включая дополнительно анимацию.

Данное исследование посвящено разработке техники визуально-интерактивной анимации для обработки, представления и анализа многомерных данных. Предлагается использовать метод визуально-интерактивного моделирования многомерного объема данных для получение максимально возможной информации об изучаемом облаке данных. Для этого ис-

пользуется подход, основанный на методах научной визуализации и визуальной аналитики.

Визуальная аналитика призвана организовать человеко-машинный интерфейс, усиливающий человеческие аналитические способности с помощью таких методов, как расширение оперативной памяти человека за счет использования визуализации, размещение информации в пространстве в соответствии с временными соотношениями, организация управляемой среды для работы пользователя в пространстве параметрических значений, организация визуального представления и интерфейсов, обеспечивающих человеку возможность сразу видеть, исследовать и понимать огромные информационные объемы.

Современное развитие вычислительной техники позволяет решать сложные задачи численного моделирования. Результатом численного моделирования служат многомерные массивы дискретных величин, выражающие поведение решения от входных параметров рассматриваемой задачи. Полученные таким образом многомерные численные результаты нуждаются в обработке и анализе. Следовательно, нужны инструменты, позволяющие анализировать данные, реализованные в многомерном пространстве.

Использование анимации

Термин «анимация», как правило, применяется для обозначения движения, вызванного некоторой процедурой или программой, а не пользователем. Движение, вызванное пользователем, будем называть «взаимодействием».

Для числа измерений меньше трех реализация визуального анализа данных не представляет особой сложности, потому что человек обладает двумерным зрением и укладывающимися в сознание геометрическими образами и представлениями для пространств с числом измерений $n \leq 3$. Для многомерных объектов с большим числом измерений подобных геометрических образов у человека нет. Следовательно, необходимо осуществлять проецирование во вложенные пространства со стандартным числом измерений. Помимо трех пространственных измерений, единственное другое измерение, которое мы встречаем в реальном мире — время. При достижении определенной частоты синтеза кадров существует естественная тенденция связывать различные статические изображения вместе (бета-движение). Изображения, показанные быстрее, чем

около 20 Гц, будут выглядеть непрерывно, без мерцания. Во время анимации обработка каждого кадра также должна включать восстановление визуализации для каждого нового временного шага или, возможно, считывания файла данных для получения нового временного шага. Одной из альтернатив анимации является «мозаика» дисплея с различными кадрами. Заметим, что анимация дает общее представление об объекте, в то время как взгляд на кадры показывает мелкие детали, которые в противном случае могли быть пропущены. Поэтому в идеале мы стремимся обеспечить оба механизма для отображения многомерных наборов данных [67, 131, 143, 144]. Как пример, рассмотрим объем данных, поступающих от томографии головы человека. Сечения по всему объему могут произвести впечатление общего изменения плотности ткани, но сложность заключается в том, что чем больше мы пытаемся охватить весь объем, тем меньше мы видим. Этот факт можно объяснить тем, что промежуточные сечения мешают восприятию. Данную проблему можно решить, имея одно сечение, организуя анимацию последовательно по всему телу, так чтобы кратковременная зрительная память могла реконструировать всю картину. Следует учитывать движение глаз при восприятии. Быстрые, строго согласованные движения глаз, происходящие одновременно в одном направлении, описываются понятием «саккада». Есть доказательства того, как саккада происходит между изображениями. Исследования этого феномена дают хороший способ скрыть изменения в деталях. Избежать описанных выше трудностей можно, если отображать на сечениях значения данных только в узлах некоторой сетки. На рис. 10.8 приведен точечный режим моделирования, показаны значения данных оттенками красного в узлах прямоугольной сетки. Приведены сразу несколько сечений, параллельных координатным плоскостям.

Визуально-интерактивная анимация

Основной особенностью рассматриваемого подхода визуально-интерактивного представления многомерных данных с числом измерений $n \geq 3$ является их динамическое представление в виде 2D-моделей сечений. Эффект многомерности достигается за счет выбора 2D-моделей, вариативности направлений динамических сечений и использования возможностей стереоскопического зрения. В тех случаях, когда числовые данные, представленные функцией, распределены непрерывно, например, как многомерная функция плотности вероятности или функция распре-

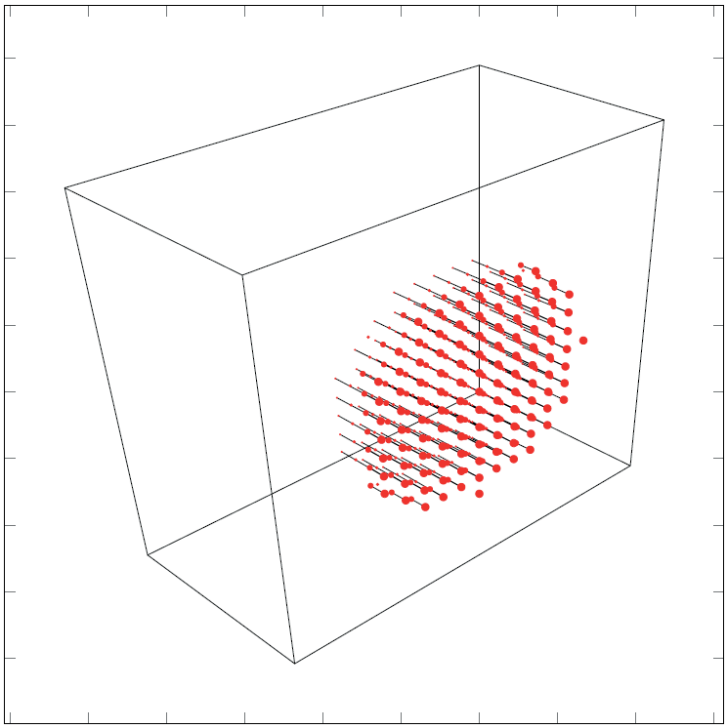


Рис. 10.8. Точечный режим представления сечений

деления температуры в объеме некоторого тела, важно знать поверхности уровня

$$S(t) = \{(x, y, z)|u(x, y, z) = t\}.$$

Для решения данной задачи предлагается отображать поверхности уровня, меняя параметр t на основе визуально-интерактивной анимации. Это позволит получить аналитическую информацию о представленных данных $u(x, y, z)$.

Остановимся подробней на представлении 3D-объекта в виде динамических сечений. Целью данного подхода является отображение движения сечений как образа 3D-объекта. Для повышения качества восприятия необходимо уменьшить эффект от перекрытия сечений. Одним из таких способов является возможность показывать на сечениях не всю информацию, а только часть. Например, применять изолинии, используя цвет, показывая информацию только в узлах сетки. Для улучшения восприятия используются стереопроекции. Рассмотрим пример воспроизведения на сечениях изолиний. В этом случае можно воспроизводить одновременно несколько близких сечений, варьируя их яркость и прозрачность. На рис. 10.9 представлен фрагмент анимации функции u пятью сечениями. Эти сечения образованы пятью параллельными плоскостями. Параметрами метода визуально-интерактивной анимации является расстояние

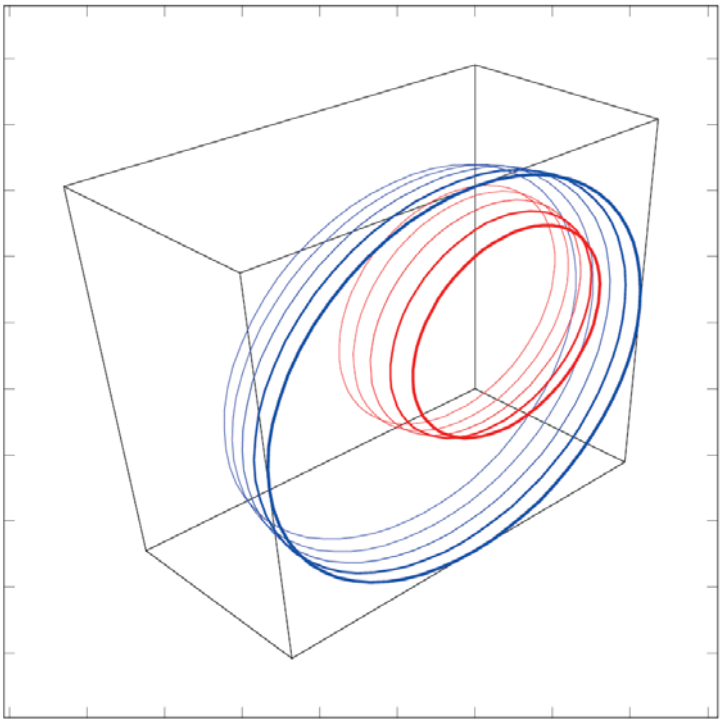


Рис. 10.9. Изолинии на сечениях $P_i, i = 1, \ldots, 5$

между плоскостями и скорость движения сечений. Толщина и прозрачность изолиний плавно меняются в зависимости от номера плоскости. Наиболее ярко изолинии представлены на переднем плане (активные сечения). Изолинии представлены в условных цветах, что соответствует значению функции u. Красный цвет соответствует значению параметра $t = t_1$, синий цвет $t = t_2$.

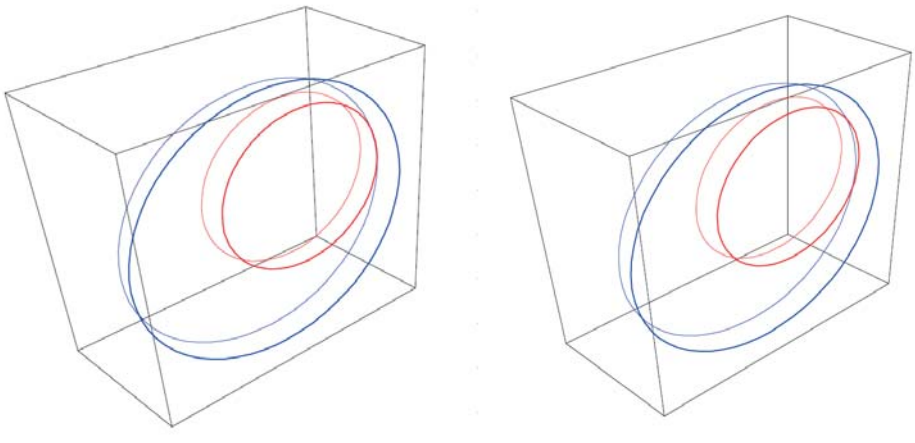


Рис. 10.10. Стереопара

Для лучшего восприятия генерируются стереопары. Построенное стереоизображение можно просматривать, используя различное оборудова-

ние. Рассмотрим наиболее доступное.

Стереоскоп — оптический прибор с двумя окулярами; обычно используется для просмотра стереослайдов, но не составляет сложности вложить туда КПК или коммуникатор с продолговатым экраном высокого разрешения. Стереодисплей — оптический инструмент, с помощью которого два плоскостных изображения комбинируются таким образом, что наблюдатель получает впечатление рельефного предмета. Можно использовать современный телевизор, позволяющий просматривать стереоизображения через специальные очки. Виртуальный шлем (VR HMD) — шлем, который показывает для каждого глаза отдельные изображения, в результате чего получается стереоэффект. Построенные стереоизображения тестировались через стереоскоп, после загрузки видеофайла на смартфон. На рис. 10.10 показан пример стереопары с двумя сечениями. Изолинии имеют различную толщину и представлены в условных цветах.

Результаты тестирования метода визуально-интерактивной анимации показали, что наиболее реалистичная картина представления трехмерных данных была получена при динамическом воспроизведении стереоизображений сечений с изолиниями в условных цветах. В настоящее время проводятся исследования по реализации данного подхода для объектов размерности больше трех.

Заключение

В монографии представлено новое направление вычислительной статистики — вычислительный вероятностный анализ (ВВА).

В работе дается всесторонний обзор современных теоретических исследований в области представления и моделирования эмпирических данных в условиях различных типов неопределенности и объема имеющейся информации. В монографии подробно рассматриваются новые направления в анализе данных, такие как функциональный анализ данных и символьный анализ данных, суть которых формируется на новых подходах к представлению эмпирической информации в виде символьных и функциональных данных.

Важное место в монографии занимают вопросы, связанные с обсуждением основных вычислительных проблем численного моделирования, включая вопросы повышения надежности результатов моделирования, эффективности организации вычислительных процессов и повышения точности численных процедур. Показано, что применение ВВА позволяет успешно решать подобные задачи. Суть данного подхода реализует идею представления эмпирических данных в виде функций распределений на основе применения кусочно-полиномиальных моделей. Применение методов и процедур ВВА позволяет представить выходные распределения вероятностей как функции входных распределений и использовать методы анализа неопределенностей, чтобы оценить влияние входных неопределённостей на неопределенность выходных параметров модели.

В работе представлены численные вероятностные арифметики и новые подходы к вычислению функций случайных аргументов — вероятностные расширения. В ВВА функции распределения рассмотрены как особый вид переменных, над которыми выполняются соответствующие операции и процедуры. Для анализа и повышения точности вычислений используется подход, основанный на правиле Рунге и экстраполяции Ричардсона. Все это позволило авторам разработать методы численного

моделирования задач со случайными входными данными. В монографии рассмотрены задачи линейной алгебры, системы нелинейных уравнений, краевые задачи для дифференциальных уравнений со случайными коэффициентами. Приведено сравнение по числу операций ВВА и метода Монте-Карло. Показано, что в ряде случаев ВВА значительно эффективней метода Монте-Карло.

В монографии рассмотрены вопросы применения ВВА к решению различных практических задач. Например, введено понятие временного ряда распределений и рассмотрены вопросы их применения в задачах метеорологии, задачах цифровой экономики, обработки данных дистанционного зондирования Земли. Обоснована актуальность применения регрессионного моделирования в пространстве распределений входных эмпирических данных для различных задач цифровой экономики.

Список литературы

1. Абрамов О. В. Мониторинг и прогнозирование технического состояния систем ответственного назначения // Информатика и системы управления. 2011. № 2. С. 4–15.

2. Альберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. М.: Мир, 1972.

3. Антонов А. В., Маловик К. Н., Чумаков И. А. Интервальная оценка характеристик надежности уникального оборудования // Фундаментальные исследования. 2011. № 12. С. 71–76.

4. Бондарев А. Е., Галактионов В. А., Чечеткин В. М. Анализ развития концепций и методов визуального представления данных в задачах вычислительной физики // Журнал вычислительной математики и математической физики. 2011. Т. 51. 4. С. 669–683.

5. Васильев Ф. П. Численные методы решения экстремальных задач.— 2-е изд., перераб. и доп. М.: Наука, 1988.

6. Вентцель Е. С. Теория вероятностей. М.: Наука, 1969.

7. Гаскаров Д. В., Шаповалов В. И. Малая выборка. М.: Статистика, 1978. 248 с.

8. Герасимов В. А., Добронец Б. С., Шустров М. Ю. Численные операции гистограммной арифметики и их применения // АиТ. 1991. № 2. С. 83–88.

9. Гнеденко Б. В. Курс теории вероятностей. М.: Наука, 1988.

10. Добронец Б. С. Интервальная математика. Красноярск: КГУ, 2004.

11. Добронец Б. С. Приближения множеств решений параметрическими множествами // Журн. Сиб. федер. ун-та. Математика и физика. Т. 2, № 3. С. 305–311.

12. Добронец Б. С. Надежность информационных систем. Сиб. федерал. ун-т. Красноярск: СФУ, 2012. 159 с.

13. Добронец Б. С., Попова О. А. Применение гистограммной математики в задачах принятия экономических решении // Тр. IX междунар. конф. по финансово-актуарной математике и эвентоконвергенции технологий. Красноярск: КГТЭН; СФУ. 2010. С. 127–130.

14. Добронец Б.С., Попова О.А. Гистограммная арифметика для визуально-интерактивного моделирования в задачах принятия экономических решений // Актуальные проблемы анализа и построения информационных систем и процессов: сб. ст. Международ. науч.-техн. конф. Таганрог: изд-во Технологического института ЮФУ. 2010. С. 44–53.

15. Добронец Б. С., Попова О. А. Численные операции над случайными величинами и их приложения // Журн. Сиб. федер. ун-та. Математика и физика. 2011. Т. 4, № 2. С. 229–239.

16. Добронец Б. С., Попова О. А. Элементы численного вероятностного анализа // Вестн. Сиб. гос. аэрокосм. ун-та им. акад. М. Ф. Решетнева. 2012. № 2. С. 19–23.

17. Добронец Б. С., Попова О. А. Численный вероятностный анализ неопределенных данных: монография. Красноярск: Сиб. федерал. ун-т, 2014.

18. Добронец Б. С., Попова О. А. Гистограммный подход к представлению и обработке данных космического и наземного мониторинга // Изв. Юж. федерал. ун-та. Технические науки. 2014. Т. 6, № 155. С. 14–22.

19. Добронец Б. С., Попова О. А. Представление и обработка неопределенности на основе гистограммных функций распределения и p-boxes // Информатизация и связь. 2014. № 2. С. 23–26.

20. Добронец Б. С., Попова О. А. Вычислительные аспекты цифровой экономики // УБС. 2020. Т. 84. Р. 114–129.

21. Добронец Б. С., Шайдуров В. В. Двусторонние численные методы. Новосибирск: Наука, 1990.

22. Дюбуа Д., Прад А. Теория возможностей. Приложения к представлению знаний в информатике. М.: Радио и связь, 1990. 288 с.

23. Дэйвид Г. Порядковые статистики. М.: Наука, 1979. 336 с.

24. Жолен Л., Кифер М., Дидри О., Вальтер Э. Прикладной интервальный анализ. М.—Ижевск: Институт компьютерных исследований, 2007. 468 с.

25. Крянев А. В., Лукин Г. В. Математические методы обработки неопределенных данных. М.: ФИЗМАТЛИТ, 2006.

26. Лукашов А. В. Метод Монте-Карло для финансовых аналитиков: краткий путеводитель // Управление корпоративными финансами. 2007. Т. 1, № 19. С. 22–39.

27. Михайлов Г. А., Войтишек А. В. Статистическое моделирование. Методы Монте-Карло: учеб. пособие для бакалавриата и магистратуры. М.: Юрайт, 2018.

28. Панов Н. В., Шарый С. П. Интервальный эволюционный алгоритм поиска глобального оптимума // Изв. Алт. гос. ун-та. 2011, № 1–2. С. 108–113.

29. Перепелица В. А., Тебуева Ф. Б. Дискретная оптимизация и моделирование в условиях неопределенности данных. М.: Академия Естествознания, 2007.

30. Попова О. А. Численное решение систем линейных алгебраических уравнений со случайными коэффициентами // Вестник ВСГУТУ. 2013. № 2 (41). С. 5–11.

31. Попова О. А. Технология извлечения и визуализации знаний на основе численного вероятностного анализа неопределенных данных // Информатизация и связь. 2013. № 2. С. 63–66.

32. Попова О. А. Гистограммный информационно-аналитический подход к представлению и прогнозированию временных рядов // Информатизация и связь. 2014. № 2. С. 43–47.

33. Попова О. А. Гистограммы второго порядка для численного моделирования в задачах с информационной неопределенностью // Изв. Юж. федер. ун-та. Технические науки. 2014. Т. 6, № 155. С. 6–14.

34. Попова О. А. Численный вероятностный анализ оптимизационных задач гидроэнергетики // Изв. Иркут. гос. ун-та. Серия: Математика. 2015. Т. 12. С. 79–92.

35. Попова О. А. Информационный подход к апостериорным оценкам погрешности численного моделирования // Информатизация и связь. 2016. № 2. С. 40–43.

36. Попова О. А. Применение численного вероятностного анализа в задачах интерполяции // Вычислительные технологии. 2017. Т. 22, № 2. С. 99–114.

37. Попова О. А. Использование экстраполяции Ричардсона для повышения точности обработки и анализа эмпирических данных // Измерительная техника. 2019. № 2. С. 18–22.

38. Попова О. А. Достоверные оценки характеристик надежности оборудования // Моделирование и анализ безопасности и риска в сложных системах. СПб.: Санкт-Петербург. гос. ун-т аэрокосм. приборостроения, 2019. С. 111–117.

39. Соболь И. М. Численные методы Монте-Карло. М.: Наука, 1973.

40. Тарасенко Ф. П. Непараметрическая статистика. Томск: ТГУ, 1976.

41. Тюрин Ю., Макаров А. Анализ данных на компьютере / под ред. В. Э. Фигурнова. М: ИНФРА–М, 2002.

42. Ужга-Ребров О. И. Управление неопределенностями. Ч. 1. Современные концепции и приложения теории вероятностей. Rēzekne: RA Izdevniecība, 2004.

43. Ужга-Ребров О. И. Управление неопределённостями. Ч. 2. Современные концепции и приложения теории вероятностей. Rēzekne: RA Izdevniecība, 2007. 388 с.

44. Ужга-Ребров О. И. Управление неопределённостями. Ч. 3. Современные невероятностные методы. Rēzekne: RA Izdevniecība, 2010. 560 с.

45. Ужга-Ребров О. И. Управление неопределённостями. Ч. 4. Комбинирование неопределённостей. Rēzekne: RA Izdevniecība, 2014. 408 с.

46. Ужга-Ребров О. И. Оценивание, анализ и распространение неопределённостей. Rēzekne: RA Izdevniecība, 2019. 582 с.

47. Задачи линейной оптимизации с неточными данными / М. Фидлер, Й. Недома, Я. Рамик, И. Рон. М.; Ижевск: НИЦ «РХД», 2008.

48. Хардле В. Прикладная непараметрическая регрессия. Пер. с англ. М.: Мир, 1993.

49. Цветков Е. В., Алябышева Т. М., Парфенов Л. Г. Оптимальные режимы гидроэлектростанций в энергетических системах. М.: Энергоатомиздат, 1984.

50. Шарый С. П. Конечномерный интервальный анализ. Новосибирск: XYZ, 2019. 635 с.

51. Шарый С. П. Интервальный анализ или методы Монте-Карло? // Вычислительные технологии. 2007. Т. 12, № 1. С. 103–116.

52. Шарый С.П. Рандомизированные алгоритмы в интервальной глобальной оптимизации // Сибирский журнал вычислительной математики. 2008. Т. 11, № 4. С. 457–474.

53. Arroyo J., Gonzalez-Rivera G. Time series modeling of histogram-valued data: the daily histogram time series of SP500 intradaily returns // Int. J. Forecast. 2012. Vol. 28. P. 20–33.

54. Arroyo J., Gonzalez-Rivera G., Maté C. Chapter forecasting with interval and histogram data. Some financial applications // Handbook of Empirical Economics & Finance / Ed. by A. Ullah, D. E. A. Giles. New York: Chapman & Hall/CRC, 2011.

55. Arroyo J., Gonzalez-Rivera G., Maté C. Smoothing Methods for Histogram-valued Time Series: An Application to Value-at-Risk // Statistical Analysis and Data Mining. 2011. Vol. 4. P. 216–228.

56. Arroyo J., Maté C. Forecasting histogram time series with k-nearest neighbours methods // Int. J. Forecast. 2009. Vol. 25. P. 192–207.

57. Berleant D. Automatically verified reasoning with both intervals and probability density functions // Interval Computations. 1993. № 2. P. 48–70.

58. Billard L., Diday E. From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis // Journal of the American Statistical Association. 2003. Vol. 98, № 462. P. 470–487.

59. Billard L., Diday E. Symbolic Data Analysis: Conceptual Statistics and Data Mining. West Sussex PO19 8SQ, England: John Wiley & Sons, 2006.

60. Brito P. Symbolic clustering of probabilistic data // Advances in Data Science and Classification / Ed. by Bock H-H Rizzi A, Vichi M. Berlin; Heidelberg: Springer, 1998. P. 385–389.

61. Clements M. Evaluating Econometric Forecasts of Economic and Financial Variables. Basingstoke, Hampshire: Palgrave Macmillan, 2005.

62. Computational Economics: a perspective from computational intelligence / Shu-Heng chen and Lakhmi Jain, editors. London: Idea Group Inc., 2006. P. 339.

63. Dias S., Brito P. Distribution and symmetric distribution regression model for histogram-valued variables // arXiv. Vol. 1303.6199v1.

64. Dias B. Linear regression with empirical distributions. Ph.D thesis. 2014. Universidade do Porto, Portugal.

65. Diday E. Probabilist, possibilist and belief objects for knowledge analysis // AnnOper Res. 1995. Vol. 55. P. 227–276.

66. Digital Economy: Impacts, Influences and Challenges. Harbhajan Kehal, editor, Varinder P. Singh, editor. Idea Group Publishing. 2005. P. 425.

67. Dill J. Expanding the Frontiers of Visual Analytics and Visualization. London: Springer-Verlag, 2012.

68. Dobronets B. S., Popova O. A. Numerical probabilistic analysis under aleatory and epistemic uncertainty // Reliable Computing. 2014. Vol. 19. P. 274–289.

69. Dobronets B., Popova O. Numerical Probabilistic Approach for Optimization Problems // Scientific Computing, Computer Arithmetic, and Validated Numerics. Lecture Notes in Computer

Science: 2016. T. 9553 / Ed. by Tucker W. Nehmeier M., Wolff von Gudenberg J. Cham: Springer, 2016. P. 43–53.

70. Dobronets B. S., Popova O. A. The numerical probabilistic approach to the processing and presentation of remote monitoring data // Journal of Siberian Federal University — Engineering and Technologies. 2016. Vol. 9, № 7. P. 960–971.

71. Dobronets B. S., Popova O. A. Improving the accuracy of the probability density function estimation // Journal of Siberian Federal University, Mathematics and Physics. 2017. Vol. 10, № 1. P. 16–21.

72. Dobronets B. S., Popova O. A. The numerical probabilistic approach to the processing and presentation of remote monitoring data Journal of Siberian Federal University. Engineering & Technologies, 2016. **9**(7), P. 960–971.

73. Dobronets B. S., Popova O. A. Piecewise Polynomial Aggregation as Preprocessing for Data Numerical Modeling // IOP Conf. Series: Journal of Physics: Conf. Series 2018. Vol. 1015. P. 032028.

74. Dobronets B. S., Popova O. A. Improving reliability of aggregation, numerical simulation and analysis of complex systems by empirical data // IOP Conf. Series: Materials Science and Engineering. 2018. Vol. 354. P. 012006.

75. Dobronets B. S., Popova O. A. Numerical Probabilistic Analysis for the Digital Economy // Sino-Russian Global Engagement Models in the Context of Digitalization of Social and Economic Processes. Conference proceedings. Krasnoyarsk: SFU, 2018. P. 12–18.

76. Dobronets B. S., Popova O. A. Computational aspects of probabilistic extensions // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. 2019. Vol. 47. P. 41–48.

77. Efron B. Bootstrap Methods: Another Look at the Jackknife // Annals of Statistics. 1979. Vol. 7(1). P. 1–26.

78. Eldred M.S. and all. DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 4.0 Users Manual // Sandia Technical Report SAND2006-6337, October 2006. P. 292.

79. Elliott G., Timmermann A. Economic Forecasting. Princeton: Princeton University Press, 2016.

80. Eubank R.L. Nonparametric Regression and Spline Smoothing. New York: CRC. 2nd ed., 1999.

81. Ewbank J. B., Foote B. L., Kumin H. J. A Method For The Solution Of The Distribution Problem Of Stochastic Linear Programming. // SIAM J. APPL. MATH. 1974. Vol. 26(2), P. 225–238.

82. Exponential smoothing methods for histogram time series based on histogram arithmetic: Rep. / Universidad Complutense de Madrid; Executor: J. Arroyo, C. Maté, S. Munoz, A. Sarabia: 2008.

83. Fan J., Gijbels I. Local Polynomial Modelling and Its Applications. London: Chapman and Hall, 1996.

84. Ferson S. What Monte Carlo Methods Cannot Do // Human and Ecological Risk Assessment. 1996. Vol. 2, № 4. P. 990–1007.

85. Ferson S., Ginzburg L. R. Different methods are needed to propagate ignorance and variability // Reliability Engineering & System Safety. 1996. Vol. 54. P. 133–144.

86. Ferson S., Hajagos J. G. Arithmetic with uncertain numbers: rigorous and (often) best possible answers // Reliability Engineering & System Safety. 2004. Vol. 85, № 1–3. P. 135–152.

87. Ferson S., Kreinovich V., Hajagos J., Oberkampf W., Ginzburg L. Experimental uncertainty estimation and statistics for data having interval uncertainty, Tech. Rep. 2007-0939, Sandia National Laboratories (2007).

88. Gasser T., Kneip A. Searching for structure in curve samples // J. Am. Stat. Assoc. 1995. Vol. 90. P. 1179–1188.

89. Gibbs A. L., Su F. E. On choosing and bounding probability metrics // Int Stat Rev. 2002. Vol. 70. P. 419–435.

90. Giordano G., Brito P. Social networks as symbolic data // Analysis and Modeling of Complex Data in Behavioral and Social Sciences / Ed. by G. Ragozini, C. Weihs, D. Vicari, A. Okada. Berlin;Heidelberg: Springer, 2014.

91. Grenander U. Stochastic processes and statistical inference // Arkiv Matematik. 1950. Vol. 1, № 3. P. 195–277.

92. Grigoriu M. Stochastic Calculus: Applications in Science and Engineering. Birkhäuser, 2002.

93. Ichino M. The quantile method for symbolic principal component analysis // Stat Anal Data Min. 2011. Vol. 4. P. 184–198.

94. Kall P., Mayer J. Stochastic Linear Programming. Models, Theory, and Computation. New York: Springer, 2005.

95. Liu B. Theory and Practice of Uncertain Programming (2nd Edition). Berlin: Springer-Verlag, 2009.

96. Loéve M. Probability Theory I, fourth edition, Vol. 45 of Graduate Texts in Mathematics. Springer, 1977.

97. Mallows C. L. A note on asymptotic joint normality // The Annals of Mathematical Statistics. 1972. Vol. 43, № 2. P. 508–515.

98. Marchuk G. I., Shaidurov V. V. Difference methods and their extrapolations. New York: Springer–Verlag, 1983.

99. Maslennikov O.P., Milman I.E., Safiulin A.E., Bondarev A.E., Nizametdinov Sh.U., Pilyugin V.V. Development of a system for analyzing of multidimensional data // Scientific Visualization, 2014. Vol. 6(4), P. 30–49.

100. Moore R. E. Interval analysis. N. J.: Prentice-Hall: Englewood Cliffs, 1966.

101. Moore R. E. Methods and Applications of Interval Analysis. Philadelrhia: SIAM, 1979.

102. Morris Jeffrey S. Functional regression // Annual Review of Statistics and Its Application. 2015. Vol. 2. P. 321–359.

103. Mosleh A., Bier V. M. Uncertainty About Probability: A Reconciliation with the Subjectivist Viewpoint // IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans. 1996. Vol. 26, № 3. P. 303–310.

104. Mayer-Schonberger V., Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York: Houghton Miffin Harcourt, 2013. P. 242.

105. Neto EAL., De Carvalho FAT. Centre and range method for fitting a linear regression model to symbolic interval data // Comput Stat Data Anal. 2008. Vol. 52. P. 1500–1515.

106. Neto EAL., De Carvalho FAT. Constrained linear regression models for symbolic interval-valued variables // Comput Stat Data Anal. 2010. Vol. 54. P. 333–347.

107. Neumaier A. Interval methods for systems of equations. Cambidge: Cambidge University Press, 1990.

108. Neumaier A. Clouds, fuzzy sets and probability intervals // Reliable Computing. 2004. Vol. 10. P. 249–272.

109. Niederreiter H. Random Number Generation and Quasi-Monte Carlo Methods. Philadelphia: SIAM, 1992.

110. Papoulis A., Pillai S. Probability, Random Variables and Stochastic Processes. Boston: McGraw-Hill, 2002.

111. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, California: Morgan Kaufmann Publishers, 1988.

112. Pedrycz W. Granular Computing: Analysis and Design of Intelligent Systems. Boca Raton, FL: RC Press/Taylor & Francis, 2013.

113. Popova O. A. Optimization problems with random data // ЖКурн. СФУ. Сер. Матем. и физ. 2013. Vol. 6, № 4. P. 506–515.

114. Popova O. A. Using Richardson extrapolation to improve the accuracy of processing and analyzing empirical data // Measurement Techniques. 2019. Vol. 62, № 2. P. 111–118.

115. Prékopa A. On the probability distribution of the optimum of a random linear program // J. SIAM Control, 1966. Vol. 4(1). P. 211–222.

116. Ramsay J. O. When the data are functions // Psychometrika. 1982. Vol. 47. P. 379–396.

117. Ramsay J. O., Dalzell C. Some tools for functional data analysis // J. R. Stat. Soc. Ser. B. 1991. Vol. 53. P. 539–572.

118. Ramsay J. O., Silverman B. W. Applied Functional Data Analysis: Methods and Case Studies. New York: Springer-Verlag, 2002.

119. Ramsay J. O., Silverman B. W. Functional Data Analysis. New York: Springer 2nd ed., 2005.

120. Rao M. M., Swift R. J. Probability Theory with Applications, Vol. of Mathematics and its Applications, second edition. Springer, 2006.

121. Richardson Extrapolation. Practical Aspects and Applications / Z. Zlatev, I. Dimov, I. Farago, A. Havasi. Berlin;Boston: Walter de Gruyter GmbH, 2018.

122. Rodriguez O., Diday E., Winsberg S. Generalization of the principal components analysis to histogram data // Workshop on Symbolic Data Analysis at the European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2004. Pisa, 2004.

123. Rudin W. Real and Complex Analysis, third edition. McGraw-Hill, 1987.

124. Rubner Y., Tomasi C., Guibas L. The Earth Mover's Distance as a Metric for Image Retrieval // International Journal of Computer Vision. 2000. Vol. 40(2). P. 99–121.

125. Schjaer-Jacobsen H. Representation and calculation of economic uncertainties: Intervals, fuzzy numbers, and probabilities // Int. J. Production Economics. 2002. Vol. 78. P. 91–98.

126. Schweizer B. Distributions are the Numbers of the Future // Proceedings of The Mathematics of Fuzzy Systems Meeting / Ed. by A. di Nola, A. Ventes ; University of Naples, Naples, Italy. 1984. P. 137–149.

127. Scott D. W. Multivariate density estimation: theory, practice, and visualization. Hoboken, New Jersey: John Wiley & Sons, 2015.

128. Shafer G. A mathematical theory of evidence. Princeton, NJ: Princeton University Press, 1976.

129. Shapiro A., Dentcheva D., Ruszczynski A. Lectures on stochastic programming: modeling and theory. Philadelphia: SIAM, 2009.

130. Soong T. T. Random Differential Equations in Science and Engineering. New York and London: Academic Press, 1973.

131. Steele J. (Eds.) Beautiful Visualization. O'Reilly Media, Inc., 2010.

132. Shu-Heng Chen Computational intelligence in economics and finance: Carrying on the legacy of Herbert Simon // Information Sciences. 2005. Vol. 170. P. 121–131.

133. Tay A., Wallis K. Density Forecasting: A Survey // Journal of Forecasting. 2000. Vol. 19, № 4. P. 235–254.

134. Taylor J. C. An Introduction to Measure and Probability. Springer, 1997.

135. Teles P., Brito P. Modelling interval time series data // Proceedings of the 3rd IASC World Conference on Computational Statistics and Data Analysis. Limassol, 2005.

136. Valeriu I. Economic Intelligence // Journal of Knowledge Management, Economics and Information Technology. Special Issue. 2013. December. P. 182–198.

137. Verde R., Irpino A. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data // Data Science and Classification, Proceedings of the IFCS. Berlin: Springer, 2006. P. 185–192.

138. Verde R., Irpino A. Dynamic clustering of histogram data: using the right metric // Selected Contributions in Data Analysis and Classification / Ed. by G. Cucumel P. Brito, P. Bertrand, F. de Carvalho. Berlin: Springer, 2007. P. 123–134.

139. Verde R., Irpino A. Ordinary least squares for histogram data based on Wasserstein distance // Proceedings of the COMPSTAT'2010 / Ed. by Y. Lechevallier, G. Saporta. Heidelberg: Physica Verlag, 2010. P. 581–589.

140. Wand M. P., Jones C. M. Kernel Smoothing. New York: Chapman & Hall, 1995.

141. Wang J., Chiou J., Muller H. Functional data analysis // Annual Review of Statistics and Its Application. Vol. 3. P. 257–295.

142. Williamson R., Downs T. Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds // International Journal of Approximate Reasoning. 1990.

143. Wright H. Introduction to Scientific Visualization. London: Springer-Verlag, 2007.

144. Zhang Q., Segall R., Cao M. Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications. Hershey, New York: Information Science reference, 2011.

145. Zhang W., Fan J. Statistical estimation in varying coefficient models // The Annals of Statistics. 1999. Vol. 27, № 5. P. 1491–1518.

146. Zhang X., Wang J. L. Varying-coefficient additive models for functional data // Biometrika. 2015. Vol. 102. P. 15–32.

147. Zhang X., Wang J. L. From sparse to dense functional data and beyond // The Annals of Statistics. 2016. Vol. 44, № 5. P. 2281–2321.

Научное издание

Добронец Борис Станиславович
Попова Ольга Аркадьевна

**Вычислительный вероятностный анализ:
модели и методы**

Монография

Редактор М. В. Саблина
Компьютерная верстка Б. С. Добронца

Подписано в печать 23.07.2020. Печать плоская. Формат 60x84/16
Бумага офсетная. Усл.-печ. л. 14,75. Тираж 500 экз. Заказ № 9830

Библиотечно-издательский комплекс
Сибирского федерального университета
660041, Красноярск, пр. Свободный, 82а
Тел. (391) 206-26-16; http://bik.sfu-kras.ru
E-mail: publishing_house@sfu-kras.ru