А.В. КРЯНЕВ, Г.В. ЛУКИН

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ НЕОПРЕДЕЛЕННЫХ ДАННЫХ

Допущено Министерством образования Российской Федерации в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению и специальности "Прикладная математика и информатика"

МОСКВА
ФИЗМАТЛИТ
2003

УДК 519.2, 519.6
ББК 22.17, 22.19
К85

Крянев А. В., Лукин Г. В. **Математические методы обработки неопределенных данных.** — М.: ФИЗМАТЛИТ, 2003. — 216 с. — ISBN 5-9221-0412-8.

В первых главах монографии изложены основные понятия параметрической и непараметрической статистики, включая понятия оценки и свойств, предъявляемых к оценкам с точки зрения их вычисления при обработке данных на компьютере. В 7–13 главах монографии изложены методы и алгоритмы восстановления регрессионных зависимостей, включая методы прогнозирования и решения задач планирования оптимальных экспериментов.

Предполагается, что читатель предварительно освоил курс теории вероятностей и математической статистики на базе, например, книги В.С. Пугачева «Теория вероятностей и математическая статистика».

В монографии представлены некоторые новые методы робастного оценивания и учета априорной информации, включая алгоритмы их численной реализации.

Основная цель монографии — ознакомить читателя с наиболее эффективными и апробированными классическими и новыми статистическими методами оценки и восстановления, научить использовать эти методы при решении конкретных задач обработки неопределенных данных.

Монография предназначена научным работникам, аспирантам, студентам старших курсов различных специальностей.

ISBN 5-9221-0412-8

Научное издание

КРЯНЕВ А.В.
ЛУКИН Г.В.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ НЕОПРЕДЕЛЕННЫХ ДАННЫХ

Редактор *Н.Б. Бартошевич-Жагель*
Оригинал-макет: *В.В. Худяков*
Оформление переплета: *А.А. Логунов*

ЛР № 071930 от 06.07.99. Подписано в печать 16.04.03.
Формат 60×90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 13,5. Уч.-изд. л. 14,5. Заказ №

Издательская фирма «Физико-математическая литература»
МАИК «Наука/Интерпериодика»
117997 Москва, Профсоюзная, 90
E-mail: fizmat@maik.ru

Неизвестная типография

ISBN 5-9221-0412-8
9 785922 104128

ОГЛАВЛЕНИЕ

Предисловие

Представляемая нами книга написана на основе курса лекций, читаемого одним из авторов на протяжении более чем 20 лет студентам старших курсов МИФИ и слушателям факультетов повышения квалификации и переподготовки.

При подготовке настоящего издания были внесены дополнения, включающие некоторые последние достижения, в том числе авторов книги, в основном по двум направлениям: робастное оценивание и оценивание с учетом дополнительной априорной информации. В книге также представлены некоторые современные методы прогнозирования временных рядов и элементы теории вейвлет-преобразования.

Книга в первую очередь предназначена лицам, использующим математические методы обработки неопределенных данных при решении прикладных задач различного содержания. Кроме того, большинство разделов книги используются студентами МИФИ, выполняющими учебно-исследовательские работы, студентами-дипломниками и аспирантами, тематика исследований которых требует применения математических методов обработки неопределенных данных.

В книге рассматриваются неопределенные данные, неопределенность которых может быть обусловлена двумя причинами: 1) стохастическим характером данных; 2) неполной априорной информацией об искомом решении, задаваемой, например, в виде принадлежности решения детерминированному множеству. Вышеуказанный неопределенный характер данных и определил название книги.

В гл. 1 представлены базовые элементы математической статистики. Анализируются основные свойства оценок и соотношения между ними, в частности эффективность и робастность оценок. Изложен традиционный материал, в том числе интервальное оценивание математического ожидания и дисперсии нормально распределенной случайной величины. Рассматриваются два универсальных метода в рамках параметрической статистики: метод моментов и метод максимального правдоподобия (ММП). Метод моментов изложен в обобщенном виде, позволяющем существенно расширить рамки его применения. С помощью неравенства Фишера–Крамера–Рао анализируется эффективность оценок. Представлены основные свойства ММП-оценок. Изложение теоретических результатов сопровождается примерами, при рассмотрении которых конструируются функции правдоподобия и находятся информация Фишера и информационная матрица Фишера, ММП-оценки и их характеристики. Особое внимание уделяется многомерному нормальному распределению.

Во 2-й главе рассматриваются методы учета дополнительной априорной информации в рамках параметрической статистики. Представлены четыре метода учета априорной информации в зависимости от ее вида. Если априорная информация об искомых параметрах имеет стохастический характер, то предлагается использовать два метода: метод Байеса и обобщенный метод максимального правдоподобия (ОММП) с заданием априорной выборки. Если же априорная информация об искомых параметрах u задается в виде принадлежности u априорному множеству R_a, то предлагается использовать два метода — минимаксный и ОММП с учетом соотношения $u \in R_a$. Анализируются схемы применения методов учета априорной информации и алгоритмы их численной реализации. Рассматриваются примеры применения методов, в частности случай, когда члены исходной выборки подчиняются нормальному закону.

В 3-й главе представлены робастные методы оценивания параметра положения в условиях наличия больших выбросов. Подчеркивается, что схема робастного оценивания параметра положения может быть без существенных качественных изменений обобщена на многомерный вариант оценивания параметров регрессионной модели. За основу робастного оценивания взят минимаксный подход Хьюбера. Предлагается итерационный метод численного нахождения робастной оценки параметра положения, основанный на геометрическом положении робастной оценки искомого параметра. Рассматривается общая схема получения робастных М-оценок, основанная на использовании функции влияния.

В гл. 4 представлены некоторые часто используемые для решения прикладных задач методы непараметрической статистики восстановления функции и плотности распределения. Подчеркивается, что задача восстановления плотности распределения по выборке, в отличие от аналогичной задачи для функции распределения, принадлежит к классу некорректно поставленных задач и поэтому может быть эффективно решена только при использовании дополнительной априорной информации об искомой плотности. Форма и объем априорной информации могут быть различными, и в зависимости от этого используются различные методы восстановления плотности распределения. В главе описаны 5 методов восстановления плотности распределения, использующие различные виды и уровни априорной информации. В методах гистограмм, Розенблатта–Парзена и корневой оценки плотности априорная информация используется для определения подходящих значений коэффициентов, аналогичных параметру регуляризации. В проекционных методах априорная информация может быть использована в виде задания априорной реперной плотности, в регуляризованном методе гистограмм априорная информация об искомой плотности задается в виде априорного класса плотностей.

В гл. 5 дается краткое изложение схем проверки гипотез о восстанавливаемом законе распределения в рамках параметрической и непараметрической статистик. Представлены критерии согласия Колмогорова, ω^2, χ^2.

Глава 6 посвящена численным методам статистического моделирования. В этой главе представлены способы моделирования случайных величин, в частности датчики равномерно распределенной нормированной случайной величины γ, основанные на методах Лемера и Неймана. Дано применение метода статистического моделирования для вычисления определенных интегралов и решения интегральных уравнений Фредгольма второго рода.

В гл. 7 изложен метод наименьших квадратов (МНК) для линейных моделей с неопределенными данными. Представлена классическая схема МНК и ее обобщения. Приводятся наиболее значимые свойства МНК-оценок и их обобщений. В конце главы дается линейная прогнозная модель, использующая МНК на этапе ее «обучения».

В гл. 8 представлены робастные схемы для линейных моделей с неопределенными данными. Все робастные схемы оценивания для линейных моделей строятся на основе функций влияния и М-оценивания. Особое внимание уделяется М-оценке Хьюбера. Предлагаются итерационные численные схемы нахождения нелинейных робастных оценок параметров линейных моделей, в частности итеративный МНК и метод вариационно-взвешенных квадратических приближений. Для нахождения робастной оценки Хьюбера предлагается эффективная итерационная процедура, сходящаяся к робастной оценке за конечное число итераций.

В гл. 9 изложены схемы учета дополнительной априорной информации в линейных моделях с неопределенными данными. Отмечается, что в условиях, когда информационная матрица Фишера исходной линейной модели вырождена или близка к вырожденной, задача оценивания параметров линейной модели принадлежит к классу некорректно поставленных задач и без учета дополнительной априорной информации невозможно получить приемлемые по точности оценки искомых параметров. В главе представлены различные схемы учета априорной информации, основанные на методах Байеса, минимаксном и ОММП. Для ОММП даны две схемы оценивания в зависимости от вида априорной информации. При применении первой схемы ОММП можно учитывать априорную информацию стохастического характера, задаваемую в виде априорной выборки. В этом случае предполагается, что совместная плотность вероятностей членов выборки зависит от искомого вектора u линейной модели. При применении второй схемы ОММП учитывается априорная информация детерминированного вида, задаваемая в виде априорного детерминированного множества, которому заведомо принадлежит искомый вектор u. В § 9.5 рассматривается также регуляризованный метод наименьших квадратов, позволяющий учитывать погрешность в элементах матрицы A исходной линейной модели $Au = f - \varepsilon$.

В гл. 10 в сжатой форме изложен метод наименьших квадратов для нелинейных моделей. Приведен итерационный метод Ньютона–Гаусса численного нахождения МНК-оценок решений нелинейных моделей. Даны регуляризованные модификации Левенберга–Марквардта итерационных процессов Ньютона–Гаусса.

Главы 11, 12 посвящены анализу и прогнозированию временных рядов. В гл. 11 представлены методы выделения детерминированных компонент временных рядов. Все представленные в главе методы основаны на представлении детерминированной компоненты в виде разложения по базисной системе функций и оценке коэффициентов разложения с помощью МНК или робастной схемы. В качестве базисной системы функций берутся: 1) система полиномов, ортогональных на множестве фиксированных значений аргумента; 2) линейные сплайны; 3) кубические сплайны; 4) вейвлеты. Предлагаются эффективные численные схемы расчета искомых коэффициентов детерминированных компонент. В гл. 12 представлено несколько сравнительно новых методов прогнозирования временных процессов. Первые два из них основаны на учете априорных экспертных оценок прогнозируемых величин и применении схем выделения детерминированных компонент, изложенных в гл. 11. Третий из представленных в гл. 12 методов прогнозирования базируется на сингулярно-спектральном анализе и выделении главных компонент исследуемого временного ряда. Указывается на соответствие между прогнозированием с помощью метода главных компонент и прогнозированием на основе линейной регрессионной модели.

В завершающей книгу гл. 13 дано краткое введение в планирование оптимальных измерений при восстановлении функциональных зависимостей. Рассматриваемая задача планирования особенно важна при проведении физических экспериментов, когда экспериментатор имеет возможность выбора значений аргумента, при которых производятся измерения восстанавливаемой функции. Вводятся различные типы оптимальности планов экспериментов и изложены методы решения задач оптимизации планов измерений для рассматриваемых типов оптимальности, включая алгоритмы их численного решения.

Ссылки на используемую литературу даны в квадратных скобках, а сам список литературы представлен в алфавитном порядке и включает в себя лишь цитируемые источники. Представленный список не претендует на полноту, а отражает лишь некоторые стороны изложенного в книге материала. Для читателя, желающего углубить свои знания по теории вероятностей и традиционным главам математической статистики, рекомендуем монографии В. С. Пугачева [62], Ж. Закса [33], М. Кендалла и А. Стюарта [38], С. Р. Рао [65], энциклопедию «Вероятность и математическая статистика» [19]. Теория матриц достаточно полно изложена в монографии Ф. Р. Гантмахера [23]. С численными методами, используемыми при обработке неопределенных данных, можно ознакомиться в книгах [20, 36, 51, 56, 70, 77].

Авторы благодарят заведующего кафедрой «Прикладная математика» МИФИ, д.ф.-м.н., профессора Н.А. Кудряшова, д.ф.-м.н., профессора этой же кафедры Т. И. Савёлову и рецензентов за просмотр рукописи и высказанные замечания, учтенные в окончательном варианте книги.

Авторы будут благодарны за все замечания и предложения, которые можно направлять по e-mail: `book@kryanev.ru`. Материалы, использованные в данной монографии, а также информацию о направлениях деятельности авторов можно найти на сайте `www.kryanev.ru`.

Москва, январь 2003 г.

А. В. Крянев
Г. В. Лукин

Введение

Статистические методы все более широко применяют при решении различных прикладных задач. В первую очередь традиционно они используются при обработке данных экспериментов, причем круг задач обработки непрерывно расширяется прежде всего за счет расширения областей, где используются экспериментальные методы, и усложнения самих экспериментов.

В последние годы в связи с запросами практики интенсивно развиваются методы исследования, основанные на измерениях косвенной информации об изучаемых системах и объектах. Часто задачи математической обработки и интерпретации неопределенных данных по косвенным проявлениям изучаемых объектов принадлежат к классу некорректных задач. Разработка статистических методов решения некорректных задач обработки и интерпретации неопределенных данных еще далека от завершения, но и в этой области уже создано много подходов и методов, а на их основе — апробированных эффективных численных алгоритмов решения поставленных задач. Знать эти методы и уметь их использовать — насущная необходимость для любого экспериментатора и специалиста, обрабатывающего неопределенные данные.

Вторая значительная особенность последнего периода развития статистических методов исследования — широкое применение современной вычислительной техники при обработке неопределенных данных. Применение компьютеров при обработке данных дает возможность решать такие задачи, решение которых совсем недавно было невозможно. Например, большинство задач обработки результатов измерений косвенных проявлений изучаемых объектов невозможно решить без применения компьютеров. Кроме того, сокращение времени обработки при применении компьютеров дает возможность создания оперативных систем управления сложными объектами, в которых слежение за поведением объекта осуществляется с помощью экспериментальных методов (например, с помощью систем различных датчиков и приборов измерения). Бурное внедрение компьютеров и вычислительных методов в статистику, и, в частности, в статистическую обработку данных, заставило заново переосмыслить место и значение многих классических методов математической статистики под углом возможности их использования с применением компьютеров и потребовало разработки самих численных реализаций, новых методов и создания эффективных численных алгоритмов обработки данных. Знать основные из этих

методов и алгоритмов, владеть ими не только полезно, но и необходимо обработчику неопределенных данных.

В последние годы в самой математической статистике, в силу ее внутреннего развития и потребности практики, созданы новые подходы и методы, более полно и адекватно учитывающие реальные ситуации, возникающие при получении неопределенных данных. Прежде всего это относится к критическому переосмыслению основных предположений, упрощений и гипотез, которые берутся в качестве основы получения тех или иных статистических подходов и методов. Всякий разумный метод обработки должен обладать свойством устойчивости в том смысле, что любое возможное (физически реализуемое) малое уклонение от базовых предположений приведет лишь к малым изменениям в конечных результатах обработки данных. Такие устойчивые методы обработки в математической статистике получили название *робастных.*

При решении прикладных задач обработку неопределенных данных, как правило, проводят в условиях, когда у исследователя помимо исходных неопределенных данных имеется дополнительная априорная информация об изучаемом объекте. Ясно, что учет этой дополнительной информации позволяет в ряде случаев существенно повысить точность конечных результатов обработки. Более того, для целого класса задач обработки данных без учета дополнительной априорной информации вообще невозможно получить приемлемый конечный результат.

В математической статистике разработаны различные методы учета дополнительной информации об изучаемом объекте и ее «соединения» в процессе обработки с неопределенными данными, ознакомить с такими методами — также одна из целей настоящей книги.

В главах 7–13 книги представлены методы и алгоритмы восстановления регрессионных зависимостей и их применения для прогнозирования.

Задачи восстановления регрессионных зависимостей встречаются во всех областях естествознания при обработке неопределенных данных.

Обширный класс регрессионных зависимостей составляют линейные задачи. К линейным регрессионным задачам относятся, например, многие задачи восстановления функциональной зависимости, задачи исследования изучаемых объектов и систем по измерениям косвенной информации об этих объектах и системах (линейные задачи идентификации). Математические модели линейных регрессионных задач — системы линейных уравнений, правая часть которых имеет случайный характер. В некоторых задачах идентификации не только правая часть исходной системы уравнений, но и оператор левой части имеет неопределенный характер и часто задается со случайными погрешностями. Из-за наличия погрешностей исходная система уравнений, как правило, является противоречивой, т. е. не существует ни одного точного решения исходной системы при заданной правой части с погрешностями. Поэтому попытки решения таких систем «в лоб» с помощью традиционных методов, ориентированных на получение точного решения, часто приводят к неудовлетворительному результату. В то же время, учитывая случайный характер погрешностей, для получения приемлемого приближенного решения этих систем естественно применять статистические методы.

Применение статистических подходов и методов для решения задач восстановления регрессионных зависимостей имеет ряд особенностей и специфических моментов, требующих качественного анализа и разработки эффективных численных алгоритмов получения приближенных решений. Ознакомить читателя с основными, наиболее эффективными и апробированными методами и численными алгоритмами решения задач восстановления регрессионных зависимостей — одна из целей настоящей книги.

В книге рассматриваются также нелинейные задачи идентификации, численное решение которых требует разработки специальных методов и алгоритмов.

Часто исследователь является не только пассивным обработчиком неопределенных данных, но и может повлиять на процесс их получения. В таких условиях естественно добиваться такого набора неопределенных данных (в рамках тех ограничений и возможностей, в которых проводится набор), при котором возможно получение наилучшего конечного результата по идентификации изучаемого объекта или системы. В книге изложены некоторые основные методы и алгоритмы решения задач планирования оптимального набора при восстановлении функциональных зависимостей.

Основные обозначения

$X_1, \ldots, X_n$ — элементы выборки
$p(x)$ — плотность вероятностей
$F(x)$ — функция распределения вероятностей
$p_\xi(x)$ — плотность вероятностей случайной величины ξ
$p(x;\ u)$ — плотность вероятностей, зависящая от параметра u
$p(x \mid u)$ — условная плотность вероятностей при фиксированном значении u
$\widehat{p}(x)$ — оценка плотности вероятностей $p(x)$
$\widehat{F}(x)$ — оценка функции распределения $F(x)$
$\widehat{u}$ — оценка параметра u
$F_n(x)$ — эмпирическая функция распределения
$\mathbf{M}\widehat{u}$ — математическое ожидание случайной величины $\widehat{u}$
$P_{\text{дов}}$ — доверительная вероятность
α_k — начальный момент k-го порядка
μ_k — центральный момент k-го порядка
$\sigma^2(\xi), \mathbf{D}(\xi)$ — дисперсия случайной величины ξ
s^2 — эмпирическая дисперсия
$\Phi(t) = \dfrac{1}{\sqrt{2\pi}} \int\limits_0^t e^{-\frac{s^2}{2}}\, ds$ — интеграл вероятностей
$\Gamma(t)$ — гамма-функция
χ_n^2 — случайная величина, распределенная по закону хи-квадрат с n степенями свободы
$I_n, I_n(u)$ — информация Фишера или информационная матрица Фишера
$L_n, L_n(u)$ — функция правдоподобия
$l_n, l_n(u)$ — логарифмическая функция правдоподобия
K_ξ — ковариационная матрица случайного вектора ξ
$\overline{Y} = \dfrac{1}{n} \sum\limits_{i=1}^n Y_i$ — средняя арифметическая
$\mathcal{N}(m, \sigma^2)$ — нормальный закон распределения с математическим ожиданием m и дисперсией σ^2
$\widehat{u}_{\text{Б}}$ — оценка Байеса параметра u
K_a — ковариационная матрица априорного распределения
u_a — центр априорного распределения параметра u
$p_{\text{апр}}(u)$ — плотность априорного распределения