А.В. КРЯНЕВ, Г.В. ЛУКИН

МЕТРИЧЕСКИЙ АНАЛИЗ И ОБРАБОТКА ДАННЫХ

Рекомендовано УМО в области ядерные физика и технологии в качестве учебного пособия

МОСКВА
ФИЗМАТЛИТ
2010

УДК 519.2+6
ББК 22.17, 22.19
К85

К р я н е в А.В., Л у к и н Г.В. **Метрический анализ и обработка данных. -** М.: ФИЗМАЛИТ, 2010. – 279 с. – ISBN 5-9221-0412-8.

Основная цель монографии – ознакомить читателя с наиболее эффективными и апробированными классическими и новыми стохастическими и детерминированными методами оценки и прогнозирования, научить использовать эти методы при решении конкретных задач обработки данных.

В монографии изложены основные понятия параметрической и непараметрической статистики, включая понятия оценок, а также требования, предъявляемые к свойствам оценок с точки зрения их вычисления при обработке данных на компьютере. В монографии представлены некоторые новые методы робастного оценивания, учета априорной информации и прогнозирования, включая алгоритмы их численной реализации. Представлены основы нового направления обработки данных - метрического анализа, позволяющего решать задачи интерполяции и прогнозирования функций одной и многих переменных на основе эффективного использования информации стохастического и детерминированного характеров об исследуемой функциональной зависимости.

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

Монография предназначена научным работникам, аспирантам, студентам старших курсов различных специальностей, использующих математические методы обработки данных.

ISBN 5-9221-0412-8

Предисловие

Представляемая нами книга является расширенным вариантом ранее изданной книги «Математические методы обработки неопределенных данных».

При подготовке настоящей книги были внесены существенные дополнения, включающие некоторые последние достижения, в том числе авторов книги, и прежде всего по новому направлению, названному метрическим анализом, что и определило название настоящей книги.

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

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 $u \in R_a$, оба из которых используют понятия метрического характера, связанные, прежде всего, с использованием расстояния в пространстве искомого вектора u. 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, $\omega 2$, $\chi 2$.

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. Значительная часть главы 12 посвящена нестационарному сингулярно-спектральному анализу.

В гл. 13 рассматриваются основы нового направления обработки данных - метрического анализа, позволяющего решать задачи интерполяции, экстраполяции и прогнозирования функций одной и многих переменных. Вводится понятие матрицы метрической неопределенности, с помощью которой определяются интерполяционные и прогнозируемые значения исследуемой функциональной зависимости. В рамках вычислительных схем метрического анализа можно оптимальным образом учитывать неопределенности как метрического, так и стохастического характеров, получая эффективные прогнозные значения. Содержание главы 13 соответствует содержанию работы авторов [51].

Глава 14 посвящена интенсивно развиваемому в последние годы детерминированному хаосу. В главе приведены широко цитируемые в научной литературе, посвященной детерминированному хаосу, примеры процессов с детерминированным хаосом. Приведены наиболее значимые известные к настоящему времени свойства детерминированных хаотических процессов, использующие понятия и инструменты метрического характера. Читателям, желающим более обстоятельно познакомиться с содержанием теории детерминированного хаоса, можно рекомендовать книгу С.П.Кузнецова [53]

В гл. 15 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 [71], Ж. Закса [38], М. Кендалла и А. Стюарта [43], С. Р. Рао [74], энциклопедию «Вероятность и математическая статистика» [22]. Теория матриц достаточно полно изложена в монографии Ф. Р. Гантмахера [26]. С численными методами, используемыми при обработке неопределенных данных, можно ознакомиться в книгах [23, 41, 59, 64, 79, 86].

Авторы благодарят заведующего кафедрой «Прикладная математика» НИЯУ 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 адресу: avkryanev@mephi.ru. Материалы, использованные в данной монографии, а также информацию о направлениях деятельности авторов можно найти на сайте www.kryanev.ru.

Москва, март 2010 г. А.В. Крянев, Г.В. Лукин