Федеральное агенство по образованию

Санкт-Петербургский государственный электротехнический университет "ЛЭТИ"

Ю. И. ИНГСТЕР А. В. МИХЕЕВ С. Н. СОЛНЫШКИН А. В. ЧИРИНА

ОСНОВНЫЕ АЛГОРИТМЫ ЧИСЛЕННОГО АНАЛИЗА

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ПАКЕТЕ MATLAB

Методические указания

Санкт-Петербург
Издательство СПбГЭТУ "ЛЭТИ"
2009

УДК 519.2
ББК 0000
А 00

А 00 Основные алгоритмы численного анализа: Метод. указания / Сост: Ю.И. Ингстер, А.В. Михеев, С.Н. Солнышкин, А.В. Чирина С. В. : Изд-во СПбГЭТУ "ЛЭТИ", 2009. ?? с.

ISBN 5-7629-0669-8

А 00 Содержит описание алгоритмов численного анализа, основанных на статистической обработке данных в пакете MATLAB. Предназначено студентам ФКТИ.

УДК 519.2
ББК 0000

ISBN 5-7629-0669-8

Введение

В настоящих методических указаниях рассматриваются алгоритмы численного анализа, основанные на статистической обработке данных с целью выявления тех или иных статистических закономерностей. Соответствующая теория излагается в курсах "Математическая статистика" и "Вычислительная математика". Теоретические закономерности могут быть проиллюстрированы путем вычислительных экспериментов. Для проведения таких экспериментов удобно использовать пакет математических и инженерных вычислений MATLAB, хорошо приспособленный для моделирования и обработки больших массивов данных.

Ниже описывается серия лабораторных работ, предназначенных, с одной стороны, для иллюстрации теоретических результатов, и, с другой стороны, для знакомства студентов с пакетом MATLAB и получения навыков работы с этим пакетом.

Для проведения лабораторных работ необходим компьютер с установленным пакетом MATLAB (версии 6 и выше) в среде WINDOWS. От студентов требуются первичные навыки работы в среде WINDOWS и знания основ программирования.

1. Лабораторная работа 1: знакомство с MATLAB

При запуске пакета MATLAB на экране компьютера возникает окно, содержащее линейки меню, инструментов и командное окно системы со знаком приглашения к работе >> (в версиях начиная с 6.5, возникают еще два дополнительных окна, содержащие историю работы: переменные и команды).

Аббревиатура "MATLAB" расшифровывается как "Матричная Лаборатория". Это означает, что основными переменными, с которыми работает пакет MATLAB, являются прямоугольные матрицы. Так n-мерный вектор рассматривается как матрица размера $1 \times n$ или $n \times 1$. Для начала работы перейдем в латинский регистр и введем матрицу, написав в командной строке

```
A=[1 2 3; 4,5,6; 7 8,2]
```

Здесь `A` — имя матрицы для дальнейших обращений, оно должно начинаться с буквы и может состоять из набора длиной до 31 символа (букв, цифр, подчеркиваний). Элементы матрицы заключены в квадратные скобки, ими могут быть целые числа, вещественные десятичные числа (целая часть отделяется точкой), комплексные числа, вводимые как $a+bi$, где a, b — вещественные числа. Элементы внутри строк отделяются пробелами или запятой, строки отделяются точкой с запятой.

Нажав клавишу "Enter", получим на экране:

```
A =
     1   2   3
     4   5   6
     7   8   2
```

С матрицей `A` можно проводить различные операции. Например, команда `B=A'` дает транспонированную (в комплексном случае — эрмитово сопряженную) матрицу, команды `C=A+B`, `D=A*B` дают, соответственно, сумму и произведение матриц, `A2=A^2` — квадрат матрицы (если эти операции определены для данных матриц, в противном случае появляется сообщение об ошибке). Для квадратной матрицы команды `d=det(A)`, `tr=trace(A)` дают определитель и след матрицы, команды `A1=inv(A)` или `A1=A^(-1)` дают обратную матрицу (если $d \neq 0$).

Поэлементные операции с матрицами обозначаются точкой, например `D1=A.*B`, `D2=A.^2`, `D3=A./B`.

Задание: выполните эти операции и посмотрите на результаты.

Логическое значение "ложь" в системе MATLAB кодируется числом 0; соответственно, "истина" — это 1. Поэтому для матриц `A`, `B` одинакового размера команда `E=A>B` даст матрицу того же размера, элементы которой e_{ij} равны 1 в тех позициях, где $a_{ij} > b_{ij}$, и нулю в противном случае (для комплексных чисел сравниваются вещественные части).

Заметим, что в системе MATLAB разрешены также записи вида `E=A+2` и `E=A>2` (подразумеваются соответствующие поэлементные операции).

Задание: введите команды: `E1=A<B`, `E2=A==B` и посмотрите на результаты.

Команды `sum(A)`, `prod(A)` вычисляют суммы и произведения элементов столбцов матрицы `A` (для векторов, т.е. матриц размера $n \times 1$ или $1 \times n$ — суммы и произведения элементов).

Задание: введите эти команды и посмотрите на результаты. Что вычисляют команды `sum(A')`, `prod(A')`?

С помощью команд типа `A(i,k)`, `A(i,:)`, `A(i:j,k)` можно выделять элементы, строки, столбцы и подматрицы матрицы `A`.

Задание: выполните операции

`a=A(2,3)`, `X=A(:, 2)`, `Y=A(1:2,3)`, `Z=A(1:2,2:3)`

и посмотрите на результаты.

Для функции одной переменной $f(x)$, команда $f(A)$ дает матрицу с элементами $f_{ij} = f(a_{ij})$. В пакете MATLAB присутствуют все основные элементарные и большое число специальных функций. Имеются различные возможности формировать функции пользователя.

Задание: выполните операции `sin(A)`, `tan(B)`, `log(A1)`, `exp(A2)`.

Специальные матрицы и массивы. Команды `zeros(n,m)`, `ones(n,m)` формируют $n \times m$-матрицы, составленные из 0 и 1 соответственно. Команда `eye(n)` формирует единичную $n \times n$-матрицу.

Задание: сформируйте матрицы `ones(2,3)`, `zeros(3,2)`, `eye(3)` и посмотрите результаты.

Команда `x=n:m` при $n \leqslant m$ формирует вектор-строку чисел от n до m с шагом 1. Команда `x=a:h:b` дает вектор-строку чисел от a до b с шагом h. Точка с запятой в конце команды блокирует вывод результатов на экран, но сохраняет их для дальнейшего использования.

Задание: введите команды

```
x=0:0.01:2*pi;  y=sin(x);  z=cos(x);  plot(x,y, x,z),  grid
```

и нажмите "Enter". Обратите внимание на точку с запятой, которой закрыты три первых команды (наблюдать на экране мелькание трёх массивов из более чем шестисот чисел было бы утомительно). В результате в новом окне выведутся графики функций $\sin(x)$ и $\cos(x)$ на промежутке $[0, 2\pi]$. Команда `plot` обеспечивает вывод графиков, по умолчанию проводя линейную интерполяцию между точками вывода (для комплекснозначных функций выводятся вещественные части и выдается соответствующее

предупреждение). Команда `grid` формирует координатную сетку. Имеется также возможность управлять цветами графиков и символами, которыми выводятся точки графиков.

Задание: введите команду `plot(x,y,'g+', x,z,'b:'); grid`. В результате график $\sin(x)$ выведется зелеными (green) плюсами, а график $\cos(x)$ — голубой (blue) пунктирной линией.

Для получения дополнительной информации о возможностях регулировки параметров вывода графиков подайте команду `help plot`.

Случайные числа. В базовой версии MATLAB имеется возможность моделировать $n \times m$-матрицы из независимых псевдослучайных чисел с равномерным распределением на интервале $[0, 1]$ (команда `rand(n,m)`) и со стандартным нормальным распределением $\mathcal{N}(0, 1)$ (команда `randn(n,m)`).

Задание: подайте команду `rand(3,4)`. Повторите эту команду несколько раз (вызывая ее из буфера с помощью клавиш $\uparrow$, $\downarrow$ и нажимая "Enter") и убедитесь, что получаются разные матрицы, составленные из чисел между 0 и 1.

Пример статистического анализа. Смоделируем выборку длины $n = 100$ из стандартного нормального распределения $\mathcal{N}(0, 1)$. Выведем ее на график, построим гистограмму, оценим среднее и среднеквадратическое отклонение (СКО). Для этого введем команды

```
n=100,  x=randn(1,n);   figure(1),   plot(x,'r*'),  grid
figure(2),   hist(x),   grid
RE=[mean(x), std(x)];   RT=[0,1];   [RE;RT]
```

На первом рисунке видим "облако" случайных точек (значения ординат в подавляющем большинстве заключено в интервале $[-3, 3]$ — вспомним "закон трёх сигм"). Гистограмма на втором рисунке напоминает колоколообразную "кривую Гаусса". Вектор `RE` содержит "экспериментальные"оценки, а вектор `RT` содержит теоретические значения среднего и СКО. Они выводятся в виде матрицы, что позволяет сравнить теоретические и экспериментальные результаты.

Чтобы ознакомиться с командой `hist` построения гистограммы и связанными с ней командами, подайте команду `help hist`. Аналогично познакомьтесь с командами `mean` (вычисление среднего значения) и `std` (вычисление среднеквадратического отклонения).

В описанном выше режиме "прямых вычислений"неудобно выполнять сколько-нибудь длинные последовательности действий. Удобнее создать специальный файл (с расширением `.m`), в который следует записать программу вычислений.

Задание: создайте файл `prog1.m`, содержащий команды из последнего примера. Запустите его несколько раз (подайте в диалоге команду `prog1`). Измените внутри файла значения параметра n на 1000 и 10000, а команду `hist(x)` на `hist(x,30)` и повторите запуск. Убедитесь, что с увеличением n форма гистограммы становится все более похожа на кривую Гаусса, а различия между экспериментальными и теоретическими оценками уменьшаются примерно в 3 и в 10 раз соответственно (закон $\sqrt{n}$).

Замечание. Программу удобно начинать с команд `clc` и `clear`, обеспечивающих очистку экрана и массивов переменных. Введите эти команды в файл `prog1.m` и запустите его еще раз.

Файлы описанного вида принято называть *файлами-сценариями*. Наряду с ними бывают ещё *файлы-функции*, содержащие в отличие от первых некий заголовок. Например, пусть файл `myf.m` содержит строки

```
function y=myf(x)
y=x.^2+4*x+5;
```

Тогда команда `t=myf(1)` возвратит $t = 10$. Обратите внимание на то, что возведение в степень производится поэлементно (предполагается, что эта функция должна будет применяться в том числе и к массивам). Все вычис-

ляемые команды внутри функции должны закрываться точкой с запятой. Принципиальны ещё два обстоятельства. Во-первых, имя функции обязано совпадать с именем файла. Во-вторых, все переменные внутри функции локальны (в частности, их имена никак не связаны с именами "внешних" переменных). Напротив, все переменные внутри сценариев глобальны, т.е. их значения видны снаружи — в диалоге и внутри других сценариев.

Создание и редактирование m-файлов удобнее всего осуществлять внутренними средствами MATLAB (через пункт "File" меню). Предварительно следует установить в качестве рабочего каталога какую-либо папку на доступном вам диске; именно в ней будут сохраняться как редактируемые m-файлы, так и сохраняемые вами результаты вычислений.

2. Лабораторная работа 2: Моделирование случайных величин

Напомним, что в пакете MATLAB можно моделировать случайные величины (СВ) со стандартными равномерным и нормальным распределениями. На их основе моделируется широкий класс законов распределения. Например, пусть $U \sim \mathcal{U}(0,1)$ — СВ со стандартным равномерным распределением. Тогда СВ $X = a + (b-a)U$ имеет равномерное распределение $\mathcal{U}(a,b)$ на интервале (a,b). Если $U \sim \mathcal{N}(0,1)$ имеет стандартное нормальное распределение, то СВ $X = a + \sigma U$ имеет нормальное распределение $\mathcal{N}(a,\sigma^2)$.

В общем случае выделим три основных подхода к моделированию:

- моделирование дискретных СВ на основе ряда распределения;
- моделирование непрерывных СВ на основе функции распределения;
- моделирование СВ на основе их свойств.

2.1. Моделирование дискретных СВ на основе ряда распределения

Пусть задан конечный ряд распределения, т.е. таблица из двух строк

X	x_1	x_2	$\cdots$	x_m
P	p_1	p_2	$\cdots$	p_m

где $x_i,\ i = 1, ..., m$ — возможные значения СВ, $p_i = P(X = x_i) > 0$ — их вероятности, $\sum_{i=1}^{m} p_i = 1$.

Пусть $U \sim \mathcal{U}(0, 1)$ — СВ со стандартным равномерным распределением. Разобьем интервал $[0, 1)$ на интервалы $\Delta_i = [z_i, z_{i+1})$ длины p_i. Пусть $X = x_i$, если $U \in \Delta_i$. Тогда $P(X = x_i) = P(U \in \Delta_i) = p_i$, т.е. СВ X имеет заданный ряд распределения.

Пример. Пусть ряд распределения имеет вид

X	1	2	3	4
P	0.1	0.2	0.3	0.4

Приведем программу MATLAB, моделирующую выборку длины $n = 100$ для этого ряда распределения.

```
XX=[1, 2, 3, 4];   P=[0.1, 0.2, 0.3, 0.4]   % Исходные данные
m=length(XX);
Z(1)=0;            % Построение интервалов разбиения
for i=1:m,   Z(i+1)=Z(i)+P(i);    end
Z(m+1)             % Контроль вычислений: должно получиться 1
n=100;             % Объем выборки
for k=1:n
   U=rand(1);        i=0;
   while U>Z(i+1),   i=i+1;     end;
   % - нашли номер i интервала, содержащего U
   X(k)=XX(i);     % k-е значение СВ X
end;        % - получили массив X из n значений требуемой СВ
hist(X),  grid     % Контроль вычислений: вывод гистограммы.
```

Задание 1. Напишите ряд распределения СВ, принимающей 5 значений. Смоделируйте выборки длины $n = 100,\ 1000,\ 10000$ СВ с этим рядом распределения. Сопоставьте значения на графике гистограммы и вероятности p_i при разных n.

2.2. Моделирование непрерывных СВ на основе функции распределения

Пусть $F(x)$ — непрерывная и строго монотонная на интервале $[a, b]$ функция распределения, $F(a) = 0$, $F(b) = 1$, $-\infty \leqslant a < b \leqslant \infty$. В этом случае существует непрерывная обратная функция $G = F^{-1}$, заданная на интервале $(0, 1)$ и принимающая значения в интервале (a, b) (значение $x = G(u)$ есть решение уравнения $F(x) = u$). Пусть $U \sim \mathcal{U}(0, 1)$ — СВ со стандартным равномерным распределением. Тогда СВ $X = G(U)$ имеет функцию распределения $F(x)$, т.е. $P(X < x) = F(x)$ для всех $x \in (-\infty, \infty)$.

Упражнение. Докажите это равенство.

Пример. Моделирование СВ с показательным распределением. Функция распределения показательного закона $\mathcal{E}xp(\lambda), \lambda > 0$ имеет вид $F(x) = 1 - e^{-\lambda x}$, $x \geqslant 0$. Решая уравнение $F(x) = u$, $u \in (0, 1)$, находим $G(u) = -\ln(1 - u)/\lambda$. Таким образом если $U \sim \mathcal{U}(0, 1)$, то $X = -\ln(1 - U)/\lambda \sim \mathcal{E}xp(\lambda)$. Заметим еще, что если $U \sim \mathcal{U}(0, 1)$, то и $1 - U \sim \mathcal{U}(0, 1)$. Поэтому для моделирования можно использовать формулу $X = -\ln(U)/\lambda$. Для контроля моделирования вспомните вид плотности распределения показательного закона, его математическое ожидание и дисперсию.

Приведем программу MATLAB, моделирующую выборку длины $n = 100$ из показательного распределения с параметром $\lambda = 2$.

```
lambda=2,   n=100     % Ввод исходных данных
U=rand(n,1);    X=-log(U)/lambda;
hist(X),  grid   % - сравните гистограмму с плотностью распр.
[mean(X), std(X); 1/lambda, 1/lambda]
% - сопоставили выборочные среднее и СКО (верхняя строка)
%   и их теоретические значения (нижняя строка)
```

Задание 2. Смоделируйте выборки длины $n = 100,\ 1000,\ 10000$ СВ с показательным распределением. Сопоставьте графики гистограмм и выборочные средние и СКО при разных n.

Задание 3. Выведите формулу моделирования и напишите программу

для моделирования СВ с распределением Вейбулла с функцией распределения $F(x) = 1 - \exp(-\lambda x^{\alpha}),\ x \geqslant 0,\ \lambda > 0,\ \alpha > 0$.

2.3. Моделирование СВ на основе их свойств

Пример 1. Моделирование СВ с распределением Лапласа. Распределение Лапласа $\mathcal{L}(a, \lambda)$ имеет плотность распределения

$$f(x) = \lambda \cdot e^{-\lambda|x-a|/2}.$$

Можно показать, что если X_1 и X_2 — независимые СВ с показательным распределением $\mathcal{E}xp(\lambda)$, то СВ $X = a + X_1 - X_2$ имеет распределение Лапласа $\mathcal{L}(a, \lambda)$.

Упражнение. Используя это свойство, покажите, что если $X \sim \mathcal{L}(a, \lambda)$, то $E(X) = a,\ D(X) = 2/\lambda^2$.

Это свойство позволяет моделировать СВ с распределением Лапласа с помощью моделирования СВ с показательным распределением (см. раздел 2.2.). Программа моделирования выборки длины $n = 100$ для параметров $\lambda = 2,\ a = 1$ имеет следующий вид.

```
lambda=2,  a=1,   n=100     % Ввод исходных данных
U=rand(2,n);   Y=-log(U)/lambda;    X=a+Y(1,:)-Y(2,:);
hist(X),  grid   % - сравните гистограмму с плотностью распр.
%   и их теоретические значения (нижняя строка)
[mean(X), std(X); a, sqrt(2)/lambda]
% - сопоставление выборочных и теоретических средних и СКО
```

Пример 2. Моделирование СВ с распределением хи-квадрат.

По определению, распределение хи-квадрат χ^2_m с m степенями свободы есть распределение СВ вида $\sum_{i=1}^{m} Z_i^2$, где $Z_i \sim \mathcal{N}(0, 1)$ — независимые стандартные нормальные СВ.

Упражнение. Используя это представление, покажите, что если $X \sim \chi^2_m$, то $E(X) = m,\ D(X) = 2m$. Используя центральную предельную теорему (ЦПТ) теории вероятностей, покажите, что при $m \to \infty$ последовательность СВ вида $Y_m = (X - m)/\sqrt{2m}$ сходится по распределению к стандартной нормальной СВ.

Это представление дает метод моделирования СВ с распределением хи-квадрат. Приведем программу моделирования выборки длины $n = 100$ из распределения хи-квадрат χ^2_m при $m = 5$.

```
m=5;   n=1000;                              % Ввод исходных данных
U=randn(m,n);  X=sum(U.^2);
hist(X),   [mean(X), std(X); m, sqrt(2*m)]    % -  контроль
```

Задание 4. Смоделируйте выборки длины $n = 100,\ 1000,\ 10000$ СВ с распределением χ^2_{10}. Сопоставьте графики гистограмм и выборочные средние и СКО при разных n.

Задание 5. Смоделируйте выборки длины $n = 1000$ СВ с распределением χ^2_m при $m = 10, 30, 100$. Постройте гистограммы величин $Y_m = (X - m)/\sqrt{2m}$ и убедитесь, что с увеличением m их распределение приближается к распределению стандартного нормального закона.

3. Лабораторная работа 3: эмпирическая функция распределения

3.1. Эмпирическая функция распределения "в точке"

Эмпирическая функция распределения (ЭФР) строится по выборке $X_1, ..., X_n$ и ее значение есть $F_n(t) = K(t)/n$, где $K(t)$ есть число элементов выборки, меньших чем t. Если $X_1, ..., X_n$ — независимая выборка с функцией распределения (ФР) $F(t)$, то $K(t)$ при каждом t представляет собой случайную величину с биномиальным распределением $B(n, p)$ и $p = F(t)$.

Упражнение. Выведите отсюда, что $E(F_n(t)) = F(t)$ (т.е. ЭФР есть несмещенная оценка ФР) и что $D((F_n(t)) = F(t)(1 - F(t))/n$ (откуда следует, что ЭФР есть состоятельная оценка ФР). Используя центральную предельную теорему (ЦПТ) теории вероятностей, покажите, что при $n \to \infty$ и для значений t таких, что $0 < F(t) < 1$, последовательность СВ вида $Y_n(t) = (F_n(t) - F(t))\sqrt{n/(F(t)(1 - F(t)))}$ сходится по распределению к стандартной нормальной СВ.

В пакете MATLAB по заданному вектору $X = (x_1, ..., x_n)$ значение ЭФР в данной точке t можно вычислить с помощью команды `mean(X<t)`

(логическое выражение `X<t` порождает вектор, составленный из нулей и единиц, и среднее значение элементов такого вектора как раз и равно доле элементов вектора X, меньших t).

Задание 1. Смоделировать выборку длины $n = 100$ из нормального распределения $\mathcal{N}(1; 4)$, вычислить значение ЭФР в точке $t = 3$. Сравнить его со значением теоретической ФР.

Первая часть задания выполняется командами

```
n=100;  t=3;  a=1;  sigma=2;
X=a + sigma*randn(1,n);    Fe=mean(X<t);
```

Для выполнения второй части задания нужно уметь вычислять значение ФР нормального закона $\mathcal{N}(a; \sigma^2)$, которая имеет вид $F(t) = \Phi\left(\frac{t-a}{\sigma}\right)$, где $\Phi(t)$ — ФР стандартного нормального закона $\mathcal{N}(0; 1)$:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-t^2/2} dt.$$

Используя toolbox "Statistics", достаточно воспользоваться функцией `normcdf` (см. `help normcdf`) и добавить команду `F=normcdf((t-a)/sigma)`.

Для сравнения значений эмпирической и теоретической ФР вычислим нормированную разность $Y_n(t)$ с помощью команды

`Yn=(Fe-F)*sqrt(n/(F*(1-F)))`.

Из близости распределения СВ $Y_n(t)$ к стандартному нормальному распределению следует, что абсолютные значения величины Y_n будут, как правило, не превосходить 3 (закон 3-х сигм). Чтобы убедиться в этом, можно повторить несколько раз программу моделирования или модифицировать ее для M-кратного моделирования:

```
n=100;  m=1000;  t=3;  a=1;  sigma=2;
for i=1:m
   X(i,:)=a+sigma*randn(1,n);
   Fe(i)=mean(X(i,:)<t);   F=normcdf((t-a)/sigma);
   Y(i)=(Fe(i)-F)*sqrt(n/(F*(1-F)));
end;
hist(Y),  grid
```

Задание 2. Смоделировать выборку длины $n = 200$ из показательного распределения $\mathcal{E}xp(2/3)$, вычислить значение ЭФР в точке $t = 2$. Сравнить его со значением теоретической ФР. То же сделать для равномерного распределения $\mathcal{U}(1, 4)$.

3.2. Эмпирическая функция распределения "в целом"

Для построения графика ЭФР вместо исходной выборки $X = (x_1, ..., x_n)$ удобно использовать *вариационный ряд* — последовательность элементов выборки $X^{var} = (x_{(1)}, ..., x_{(n)})$, упорядоченная в порядке их возрастания. В частности,

$$x_{(1)} = \min_{i=1,...,n} x_i, \quad x_{(n)} = \max_{i=1,...,n} x_i;$$

величина

$$med_n = \begin{cases} x_{(k)}, & n = 2k - 1, \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), & n = 2k, \end{cases}$$

называется *выборочной медианой*. В пакете MATLAB вариационный ряд строится с помощью команды `Y=sort(X)`, а выборочная медиана вычисляется командой `med(X)`.

ЭФР $F_n(t)$ принимает значение i/n в интервале между членами вариационного ряда, то есть при $t \in (x_{(i)}, x_{(i+1)}]$, $i = 1, ..., n - 1$, равна 0 при $t \leqslant x_{(1)}$ и равна 1 при при $t > x_{(n)}$. В точках вариационного ряда $F_n(t)$ имеет скачки $1/n$ (которые суммируются, если некоторые точки вариационного ряда совпадают). Если исходная выборка $X = (x_1, ..., x_n)$ задана, то программа для построения графика ЭФР на интервале $(x_{(1)}, x_{(n)})$ имеет вид

```
n=length(X); Y=sort(X);  T=1/n:1/n:1;  stairs(Y,T);  grid
```

Команда `stairs(Y,T)` (см. `help stairs`) аналогична команде `plot(Y,T)`, только соседние точки соединяются не наклонными линиями, а вертикальными и горизонтальными отрезками так, что график приобретает ступенчатый вид. Другой вариант этой команды: `[Ys,Ts]=stairs(Y,T)` ничего не рисует, но возвращает два массива `Ys` и `Ts`, по которым можно построить аналогичный график обычной командой `plot`. Следующая программа выводит на один график оба варианта линии:

```
Y=sort(X);  T=1/n:1/n:1;  [Ys,Ts]=stairs(Y,T);
plot(Ys,Ts, Y,T,':');  grid
```

(различие между этими линиями не превышает $1/n$ для непрерывных $F(t)$, что вполне допустимо при $n \geqslant 50 \div 100$).

Для сравнения ЭФР и теоретической ФР не в отдельных точках, а "в целом" обычно используют величины

$$\rho_n = \sup_{-\infty<t<\infty} |F_n(t) - F(t)|, \qquad K_n = \sqrt{n} \cdot \rho_n,$$

которые характеризуют максимальное (по вертикали) расстояние между графиками этих функций. Согласно теореме Гливенко, $\rho_n \to 0$ с вероятностью 1 и, следовательно, по вероятности при $n \to \infty$. А вот распределение СВ K_n стремится при $n \to \infty$ (для непрерывных $F(t)$) к специальному распределению, не зависящему от $F(t)$ и называемому *распределением Колмогорова*. Аналитический вид ФР Колмогорова $K(t)$ достаточно сложен, приведем лишь таблицу ее квантилей, т.е. значений k_γ, таких что $K(k_\gamma) = \gamma$.

γ	0.800	0.850	0.900	0.950	0.980	0.990	0.995	0.998	0.999
k_γ	1.073	1.138	1.224	1.358	1.517	1.628	1.731	1.858	1.949

Это позволяет строить *асимптотическую доверительную полосу*, внутри которой график неизвестной непрерывной ФР будет лежать с вероятностью, близкой к заданной доверительной вероятности γ при достаточно больших n. При заданном γ границы этой полосы $F^{\pm}_{n,\gamma}$ определяются по формулам

$$F^{-}_{n,\gamma}(t) = \max(F_n(t) - k_\gamma/\sqrt{n},\ 0), \quad F^{+}_{n,\gamma}(t) = \min(F_n(t) + k_\gamma/\sqrt{n},\ 1).$$

Упражнение. Объясните, почему в этих формулах используются операции max и min.

Для непрерывной $F(t)$ вычисление значения ρ_n может проводиться на основе сравнения значений $F_n(x_{(k)})$, $F_n(x_{(k)} + 0)$ и $F(x_{(k)})$ только в точках

вариационного ряда (*упражнение*: объясните, почему?) по формулам

$$\rho_n = \max_{k=1,\dots,n} \max\left\{\left|F(x_{(k)}) - \tfrac{k-1}{n}\right|;\ \left|F(x_{(k)}) - \tfrac{k}{n}\right|\right\}.$$

Пример. Смоделировать выборку длины 100 из нормального распределения $\mathcal{N}(1,4)$, построить графики ЭФР и теоретической ФР, вычислить расстояние ρ_n и K_n, построить доверительную полосу с уровнем доверия $\gamma = 0.9$.

```
n=100;  a=1;  sigma=2;  kgamma=1.224;
X=a+sigma*randn(1,n);   Y=sort(X);   F=normcdf((Y-a)/sigma);
T=0:1/n:1-1/n;
rho=max(max(abs(F-T), abs(F-T-1/n)));
K=sqrt(n)*rho
F1=max(T-kgamma/sqrt(n), 0);   F2=min(T+kgamma/sqrt(n), 1);
plot(Y,T, Y,F,'--', Y,F1,':', Y,F2,':');   grid
```

Задание 3. Напишите аналогичные программы для выборок из показательного и равномерного распределений.

Величины K_n, вычисляемые в программе, можно использовать для оценки функции распределения Колмогорова.

```
n=100;   a=1;  sigma=2;  m=3000;
for i=1:m
   X=a+sigma*randn(1,n);  Y=sort(X);
   F=normcdf((Y-a)/sigma);  T=0:1/n:1-1/n;
   rho=max(max(abs(F-T), abs(F-T-1/n)));
   K(i)=sqrt(n)*rho;
end;
Y=sort(K);   T=1/m:1/m:1;   plot(Y,T);  grid
```

Задание 4.

1) Повторите моделирование с измененными параметрами `a`, `sigma` и убедитесь, что характер графиков не меняется.

2) Напишите аналогичные программы для выборок из показательного и равномерного распределений и убедитесь, что характер графиков не меняется.

3.3. Критерий Колмогорова

Критерий Колмогорова используется для проверки гипотезы $H_0 : F = F_0$ о том, что ФР $F(t)$, из которой получена выборка, совпадает с заданной ФР $F_0(t)$. Он состоит в следующем. Задается допустимая вероятность ошибки I рода $\alpha \in (0, 1)$ (обычно $\alpha = 0.1,\ 0.05$ или 0.01), полагается $\gamma = 1 - \alpha$ и вычисляется значение статистики Колмогорова K_n, причем в качестве $F(t)$ используется заданная ФР $F_0(t)$. Гипотеза H_0 отвергается при $K_n > k_\gamma$; в противном случае говорят, что гипотеза H_0 не противоречит данным, содержащимся в выборке.

Пример. Пусть $\alpha = 0.1,\ n = 100$. Проверяется гипотеза $H_0 : F = \Phi$ о том, что выборка получена из стандартного нормального закона, в то время как выборка получена из нормального распределения $\mathcal{N}(a, 1)$ с параметром $a > 0$. Оценить, при каком минимальном a можно достаточно устойчиво отвергнуть гипотезу: вероятность β ошибки II рода (принять H_0 в то время как она не верна) мала — порядка 0.1.

Модифицируем предыдущую программу и будем подбирать значение a. В качестве начального значения примем $a = 1$.

```
n=100;   a=1;  sigma=2;  m=1000;   tgamma=1.224
for i=1:m
   X=a+sigma*randn(1,n);  Y=sort(X);
   F=normcdf((Y-a)/sigma);  T=0:1/n:1-1/n;
```

```
    rho=max(max(abs(F-T), abs(F-T-1/n)));
    K(i)=sqrt(n)*rho;
end;
beta=mean(K>tgamma)
```

Величина `beta` — оценка вероятности ошибки II рода β, она уменьшается с ростом a. При $a = 1$ получается слишком малое $\beta \approx 0$; последовательно уменьшая или увеличивая a, найдем, что $\beta \approx 0.1$ при $a \approx 0.34$.

Задание 5.

1) Приведите аналогичное моделирование для $n = 400$. Как изменится a?

2) Напишите аналогичную программу для показательного распределения выборки $X \sim \mathcal{E}xp(\lambda)$, приняв в кчестве гипотезы $H_0 : X \sim \mathcal{E}xp(1)$. Пусть $n = 50$ и $\alpha = 0.95$. При каком минимальном $\lambda > 1$ можно получить вероятности ошибки II рода $\beta < 0.4$?

3) Проверьте на основе моделирования, что распределение χ^2_m при больших m близко к нормальному распределению $\mathcal{N}(m, 2m)$. При каких минимальных m нельзя устойчиво различить ($\beta \geqslant 0.8$) выборки длины $n = 100$ из этих распределений?

4. Лабораторная работа 4: сравнение статистических оценок параметра положения

Пусть случайная величина X имеет функцию распределения $F(x, a) = H(x - a)$, где $H(x)$ — симметричная функция распределения: $H(x) = 1 - H(-x)$, a — параметр положения (его также называют параметр сдвига). Если X имеет конечное математическое ожидание, то оно совпадает с медианой: $E(X) = med(X)$=a.

Мы будем изучать следующие оценки параметра a:

- выборочное среднее $a1 = \overline{x}_n$,
- выборочная медиана $a2 = med_n(X)$,
- полусумма выборочных максимума и минимума $a3 = (X_{(1)} + X_{(n)})/2$.

Рассматриваются три распределения выборки:

- нормальное распределение $\mathcal{N}(a, \sigma^2)$,
- распределение Лапласа $\mathcal{L}(a, \lambda)$,
- равномерное распределение $\mathcal{U}\left(a-\frac{\delta}{2};\ a+\frac{\delta}{2}\right)$ на отрезке $\left[a-\frac{\delta}{2};\ a+\frac{\delta}{2}\right]$.

Параметры σ, λ и δ предполагаются известными.

Качество оценки $T_n(X)$ определяется величиной ее среднеквадратического риска $R_n^2 = E(T_n(X) - a)^2$. Для несмещенных оценок величина R_n^2 совпадает с дисперсией оценки $D(T_n(X))$. Точное значение этой величины часто вычислить не удается, но можно оценить ее порядок при объеме выборки n, стремящемся к бесконечности.

Асимптотика дисперсий изучаемых оценок приведена в следующей таблице.

	$\mathcal{N}(a, \sigma^2)$	$\mathcal{L}(a, u)$	$\mathcal{U}\left(a-\frac{\delta}{2};\ a+\frac{\delta}{2}\right)$
$\overline{X}_n$	σ^2/n	$2/n\lambda^2$	$\delta^2/12n$
$med_n(X)$	$\pi\sigma^2/2n$	$1/n\lambda^2$	$\delta^2/4n$
$(X_{(1)} + X_{(n)})/2$	$\sigma^2\pi^2/24\ln n$	$\pi^2/12\lambda^2$	$\delta^2/2n^2$

Задание 1. Какая статистика предпочтительнее для оценки параметра положения для: а) нормального распределения? б) распределения Лапласа? с) равномерного распределения? Обоснуйте выбор в каждом случае. Укажите статистику, которая для одного из распределений не является состоятельной.

Сравним оценки с помощью математического моделирования. Смоделируем m выборок длины n с заданным распределением, а в качестве оценки среднеквадратического риска R_n возьмем среднеквадратическое отклонение (СКО) оценки: чем оно меньше, тем лучше оценка. Введем вектор среднеквадратических отклонений $R = [r1, r2, r3]$ для оценок $a1$, $a2$ и $a3$. *Задание 2.*

1) Смоделируйте $m = 100$ выборок X объема $n = 10; 100; 1000$ с заданным распределением.

2) Вычислите значения оценок $a1$– $a3$:

```
a1=mean(X);   a2=median(X);   a3=(min(X)+max(X))/2;
```

3) Найдите значения разброса $r1$– $r3$:

```
r1=std(a1);   r2=std(a2);   r3=std(a3); R=[r1,r2,r3]
```

Как влияет увеличение объема выборки на точность оценки?

4) Сравните экспериментальные результаты с теоретическими для разных объемов выборки n.

5. Лабораторная работа 5: линейные статистические модели и метод наименьших квадратов

Пусть имеется набор n измерений неизвестной функции $f(x)$ в точках x_l (будем называть их *узлами*) со случайными ошибками (*шумами*) ε_l:

$$y_l = f(x_l) + \varepsilon_l, \quad l = 1, \ldots, n. \tag{1}$$

Требуется по набору измерений $\{(x_l, y_l),\ l = 1, \ldots, n\}$ оценить функцию $f(x)$. Конечно, для этого нужно задать некоторую конкретную модель этой функции. Мы будем рассматривать линейные модели вида

$$f(x) = f(x, \theta) = \sum_{k=1}^{m} \theta_k \varphi_k(x), \tag{2}$$

где $(\theta_1, \ldots, \theta_m) = \theta^T$ — вектор неизвестных параметров, а $\varphi_1(x), \ldots, \varphi_m(x)$ — заданный набор функций, которые называют *базисными*. Предполагается, что количество параметров m меньше, чем количество измерений n (обычно — много меньше).

Обозначим через A матрицу размера $n \times m$ с элементами $a_{lk} = \varphi_k(x_l)$. Эту матрицу называют *регрессором*; каждый её столбец представляет собой совокупность узловых значений одной из базисных функций. Пусть Y — вектор-столбец $(y_1, \ldots, y_n)^T$ результатов измерений. Тогда равенства (1), (2) можно переписать в виде

$$Y = A\theta + \varepsilon, \tag{3}$$

где $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ — вектор ошибок. Соотношение (3) называют *общей линейной статистической моделью*. При этом по заданному вектору Y и матрице A требуется оценить неизвестный вектор параметров θ.

Метод наименьших квадратов (МНК) состоит в том, что оценку неизвестной функции ищут в виде $\widetilde{f}(x) = f(x, \widetilde{\theta})$, где вектор $\widetilde{\theta}$ находится из условия минимизации суммы квадратов отклонений измеренных значений y_l от значений неизвестной функции в точках измерения:

$$\sum_{l=1}^{n} \left| y_l - f(x_l, \widetilde{\theta}) \right|^2 = \min_{\theta} \sum_{l=1}^{n} |y_l - f(x_l, \theta)|^2 . \tag{4}$$

Если бы шумы отсутствовали, то вектор θ являлся бы точным решением системы уравнений $A\theta = Y$. При наличии шумов эта система, скорее всего, не будет иметь решений, поскольку переопределена (количество уравнений превышает количество неизвестных). Требование (4) сводится к минимизации обычной (квадратичной) нормы *вектора невязок* $Y - A\theta$. Вектор $\widetilde{\theta}$, дающий такой минимум, принято называть *псевдорешением* системы $A\theta = Y$. Из курса линейной алгебры известно, что псевдорешение является точным решением вспомогательной системы

$$A^* A\,\theta = A^*\,Y \qquad \Longleftrightarrow \qquad B\,\theta = C \tag{5}$$

(такое решение всегда существует), где для сокращения записи введены обозначения $B = A^*A$ (квадратная матрица размера $m \times m$) и $C = A^*\,Y$ (столбец высоты m). Здесь A^* — матрица, эрмитово сопряжённая к матрице A, т.е. полученная из неё транспонированием и комплексным сопряжением всех элементов. Таким образом, $b_{jk} = \sum_{l=1}^{n} \overline{a_{lj}} \cdot a_{lk}$ и $c_k = \sum_{l=1}^{n} \overline{a_{lk}} \cdot y_l$; искомый вектор параметров $\widetilde{\theta}$ находится как $\widetilde{\theta} = B^{-1}C$.

Будем называть столбец $\widetilde{Y} = A\,\widetilde{\theta}$ *вектором подгонки* и обозначать соответствующий вектор невязок как $\widetilde{R} = Y - \widetilde{Y}$. С геометрической точки зрения, вектор подгонки $\widetilde{Y}$ есть не что иное как ортогональная проекция вектора измерений Y на некоторое линейное подпространство („образ“ матрицы A). Поэтому векторы подгонки и невязок ортогональны, т.е. $(\widetilde{Y}, \widetilde{R}) = \sum_{l=1}^{n} \widetilde{y}_l \cdot \overline{\widetilde{r}_l} = 0$.

Пусть ошибки ε_i являются случайными величинами и удовлетворяют условиям

$$E(\varepsilon_l) = 0, \quad E(\varepsilon_l^2) = \sigma^2 > 0, \quad E(\varepsilon_l \varepsilon_k) = 0 \quad \text{при } l \neq k.$$

Тогда в рамках модели (3) оценки МНК являются *несмещенными*, т.е.

$$E(\widetilde{\theta}) = \theta, \quad E(\widetilde{Y}) = A\theta.$$

При $m < n$ можно построить несмещенную оценку дисперсии шумов σ^2:

$$\widetilde{\sigma}_n^2 = \frac{1}{n-m}\|r\|^2 = \frac{1}{n-m}\sum_{i=1}^{n}|y_i - \widetilde{y}_i|^2 .$$

5.1. Полиномиальная регрессия

Пусть $\varphi_k(x) = x^{k-1},\ k = 1, \ldots, m$, то есть неизвестная функция в (2) ищется в виде полинома степени $m-1$ с неизвестными коэффициентами θ_k. В этом случае говорят, что соотношения (1), (2) описывают *модель полиномиальной регрессии*. При $m = 2$ неизвестная функция является линейной. В этом случае говорят о *модели простой линейной регрессии*.

Для модели полиномиальной регрессии

$$a_{lk} = x_l^{k-1}, \quad b_{jk} = \sum_{l=1}^{n} x_l^{k+l-2}, \quad c_k = \sum_{l=1}^{n} y_l x_l^{k-1}$$

и матрица B невырождена при $m \leqslant n,\ x_l \neq x_k$ при $l \neq k$. В общем случае вектор параметров $\widetilde{\theta}$ и полином налучшего приближения $f(x, \widetilde{\theta})$ следует находить составлением и решением системы (5). Для этих целей в пакете MATLAB предусмотрена встроенная команда

`p=polyfit(X,Y,m)`.

Здесь X, Y — вектора исходных данных, m — степень полинома, p — вектор коэффициентов полинома, записанного как

$$p(1)x^m + p(2)x^{m-1} + \ldots + p(m+1).$$

Значения полинома $y = f(t, p)$ для массива t значений аргумента можно вычислить с помощью команды

`y=polyval(p,t)`.

В частности, вектор подгонки $\widetilde{Y} = Z$ получается с помощью команды `Z=polyval(p,X)`.

Приведем пример программы MATLAB, реализующей МНК непосредственно и с использованием специальных команд для случая квадратичной регрессии (комментарии даны в скобках после соответствующих команд):

```
    %% Создаем столбец из 20 узлов
    %% в заданных границах с равномерным шагом:
xmin=0;  xmax=1;  n=20;  X=(xmin:(xmax-xmin)/(n-1):xmax)';
```

```
   %% Формируем "истинную" (неизвестную) функцию:
p1=3;  p2=-2;  p3=1; p=[p1 p2 p3]
 Y0=p1+p2*X+p3*X.^2;
   %% Генерируем экспериментальные данные:
sigma=0.1;                                  % (уровень шумов)
Z=sigma*randn(n,1);  Y=Y0+Z;
plot(X,Y, X,Y,'*'), grid                              % Контроль
   %% Формируем регрессор:
A1=ones(n,1);  A2=X;  A3=X.^2;   A=[A1,A2,A3];
   %% Непосредственно оцениваем коэффициенты [p1,p2,p3]:
B=A'*A;  C=A'*Y;  pn1=(B\C)',
Yn1=A*pn1';    % (вектор подгонки)
   %% Оцениваем средствами  MATLAB:
pn3=polyfit(X,Y,2); pn2=fliplr(pn3)
  Yn2=polyval(pn3,X);
plot(X,Y,'*', X,Y0, X,Yn1,'+', X,Yn2,'o'), grid   % Контроль
   %% Проверяем ортогональность векторов подгонки и невязок:
R=Y-Yn1;   sum(R.*Yn1)
   %% Оцениваем уровень шумов
sn=sqrt(sum(R.^2)/(n-3))                    % (сравнить с sigma)
```

Для модели простой линейной регрессии есть явные формулы:

$$\widetilde{\theta}_1 = \widehat{y} - b\,\widehat{x}, \quad \widetilde{\theta}_2 = \widetilde{K}_{XY}/\widetilde{D}_X,$$
$$f(x,\widetilde{\theta}) = \hat{y} + \widetilde{\theta}_2(x-\hat{x}),$$

где $\widehat{x}$ и $\widehat{y}$ — выборочные средние

$$\widehat{x} = \frac{1}{n}\sum_{l=1}^{n} x_l, \quad \widehat{y} = \frac{1}{n}\sum_{l=1}^{n} y_l;$$

$\widetilde{K}_{XY}$ — выборочная ковариация, $\widetilde{D}_X$ — выборочная дисперсия:

$$\widetilde{K}_{XY} = \frac{1}{n}\sum_{l=1}^{n}(x_l-\widehat{x})(y_l-\widehat{y}), \quad \widetilde{D}_X = \frac{1}{n}\sum_{l=1}^{n}(x_l-\widehat{x})^2.$$

Можно также построить *асимптотический доверительный интервал* $\Delta_\gamma(x) = (f_\gamma^-(x), f_\gamma^+(x))$ для значений неизвестной линейной функции $f(x)$,

такой что вероятность события $f(x) \in \Delta_\gamma(x)$ близка к заданной доверительной вероятности γ при больших n. Границы доверительного интервала имеют вид

$$f_\gamma^{\pm}(x) = f(x, \widetilde{\theta}) \pm \frac{t_\gamma\, s_n}{\sqrt{n}} \left(1 + \frac{(x - \widehat{x})^2}{\widetilde{D}_X}\right)^{1/2}$$

(заметим, что ширина доверительного интервала минимальна при $x = \widehat{x}$). Здесь s_n — оценка уровня шумов, и t_γ — двусторонняя γ-квантиль стандартного нормального распределения, определяемая соотношением $\Phi(t_\gamma) - \Phi(-t_\gamma) \equiv \gamma$. Это число может быть определено из следующей таблицы:

γ	0.800	0.850	0.900	0.950	0.980	0.990	0.995	0.998	0.999
t_γ	1.282	1.440	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Приведем программу для оценки простой линейной регрессии и построения границ доверительных интервалов для значений функции с уровнем доверия $\gamma = 0.95$.

```
xmin=0;  xmax=1;  n=20;  X=(xmin:(xmax-xmin)/(n-1):xmax)';
c1=3;  c2=-2;   Y0=c1+c2*X;          % "Истинная" функция
   %% Экспериментальные данные
sigma=0.2;  Z=sigma*randn(n,1);   Y=Y0+Z;
plot(X,Y0, X,Y,'o'), grid            % Контроль
   %% Явное оценивание:
xm=mean(X);  ym=mean(Y);
K=sum((X-xm).*(Y-ym))/(n-1);    DX=(std(X)^2);
cn1(2)=K/DX;  cn1(1)=ym-cn1(2)*xm;
Yn1=ym+cn1(2)*(X-xm);
   %% Непосредственное оценивание средствами MatLab:
cn2=polyfit(X,Y,1)
Yn2=polyval(cn2,X);
   %% Контроль:
figure(1);  plot(X,Y,'*', X,Y0, X,Yn1,'+', X,Yn2,'o'), grid
R=Yn1-Y;  sum(R.*Yn1)                % Проверка ортогональности
sn=sqrt(sum(R.^2)/(n-2))             % Шумы - сравнить с sigma
   %% Границы доверительных интервалов:
```

```
tgamma=1.96;   delta=tgamma*sn*(((1+(X-xm).^2/DX))/n).^(1/2);
f1=Yn1-delta;   f2=Yn1+delta;        % (границы)
figure(2);  plot(X,Y0, X,Yn1,':', X,f1,'--', X,f2,'--'),  grid
```

5.2. Тригонометрическая регрессия

Пусть $X_i \in [0,T],\ \omega = 2\pi/T$;

$$\varphi_1(x) = 1, \quad \varphi_{2k}(x) = \sin(k\omega x), \quad \varphi_{2k+1}(x) = \cos(k\omega x), \quad k = 1,\dots,r,$$

то есть неизвестная функция в (2) ищется в виде тригонометрического полинома степени r

$$f(x,\alpha,\beta) = \alpha_0 + \sum_{k=1}^{r} (\alpha_k \cos(k\omega x) + \beta_k \sin(k\omega x))$$

с неизвестными коэффициентами $\alpha_k,\ \beta_k$. В этом случае говорят о *модели тригонометрической регрессии*. Эту модель удобно представить в комплексной форме

$$f(x,\theta) = \sum_{k=-r}^{r} \theta_r \psi_k(x), \qquad \psi_k(x) = e^{ik\omega x}. \tag{6}$$

Здесь $i \equiv \sqrt{-1}$; коэффициенты α_k, β_k и θ_k связаны соотношениями

$$\theta_0 = \alpha_0; \qquad \theta_{\pm k} = \alpha_k \pm i\,\beta_k, \quad k = 1,\dots,r;$$
$$\alpha_k = \frac{1}{2}(\theta_k + \theta_{-k}), \quad \beta_k = \frac{1}{2i}(\theta_k - \theta_{-k}), \quad k = 1,\dots,r.$$

Если (как оно обычно и бывает) задача вещественна, то $\theta_{-k} \equiv \overline{\theta_k}$ (в частности, $\theta_0 \in \mathbb{R}$); при этом

$$\alpha_k = \operatorname{Re}\theta_k, \quad \beta_k = \operatorname{Im}\theta_k.$$

Обычно тригонометрическую регрессию применяют к равноотстоящим узлам: $x_l = (l-1)T/n,\ l = 1, \ldots, n$ (последний узел $x_{n+1} = T$ не используется, т.к. в силу периодичности значение в нём совпадает со значением в начальном узле $x_1 = 0$). Столбцы матрицы A при этом оказываются *ортогональными*, т.е.

$$b_{jk} = \sum_{l=1}^{n} a'_{lk} a_{lj} = 0 \quad \text{при всех} \quad k \neq j.$$

Тогда матрица $B = A^* A$ диагональна с элементами на диагонали

$$b_{kk} = \sum_{l=1}^{n} |a_{lk}|^2, \quad k = 1, \ldots, m.$$

Система (5) при этом решается сразу:

$$\widetilde{\theta}_k = \frac{c_k}{b_{kk}}, \quad k = 1, \ldots, m.$$

Для комплексной формы модели имеем $b_{kk} = n$ (т.к. модули комплексных экспонент равны единице), поэтому

$$c_k = \sum_{l=1}^{n} y_l \, e^{-2\pi i \, (l-1)k/n}, \quad k = -r, \ldots, r. \tag{7}$$

При этом

$$f(x, \widetilde{\theta}) = \sum_{k=-r}^{r} \widetilde{\theta}_k e^{i \, k\omega x}, \quad \widetilde{\theta}_k = c_k/n;$$

в частности, вектор подгонки имеет вид

$$\widetilde{y}_l = f(x_l, \widetilde{\theta}) = \sum_{k=-r}^{r} \widetilde{\theta}_k \, e^{2\pi i \, k(l-1)/n}, \quad l = 1, \ldots, n. \tag{8}$$

Переход (7) от вектора Y к вектору C соответствует *дискретному преобразованию Фурье* (ДПФ) и *фильтрации*, представляющей собой в данном случае обнуление величин c_k с номерами $|k| > r$, а переход (8) от вектора C к вектору $\hat{Y}$ соответствует *обратному дискретному преобразованию Фурье* (ОДПФ). В пакете MATLAB ДПФ и ОДПФ реализуются командами `fft` и `ifft` (см. `help fft`). При этом, если `Z=iff(Y)`, то величины z_k и c_k связаны соотношениями

$$c_k = z_{k+1},\ k = 0, 1, \ldots, r; \quad c_{-k} = z_{n-k+1},\ k = 1, \ldots, r.$$

Заметим, что при $r \geqslant n/2$ вектор Y восстанавливается точно, т.е. $\widetilde{Y} = Y$ (это соответствует задаче тригонометрической интерполяции). Однако

даже при отсутствии шумов функция $f(x)$ может не восстанавливается точно. Именно, если разложение в ряд Фурье функции имеет вид

$$f(x) = \sum_{k=-\infty}^{\infty} \theta_k \, e^{i\,k\omega x},$$

то коэффициенты Фурье $\widetilde{\theta}_k$ восстановленной функции имеют вид

$$\widetilde{\theta}_k = \sum_{m=-\infty}^{\infty} \theta_{k+nm}, \quad |k| \leqslant n/2, \tag{9}$$

т.е. к коэффициенту θ_k добавляются коэффициенты $\theta_{k\pm n}, \theta_{k\pm 2n}$ и т.д. Этот эффект называется *наложением частот* и связан с дискретизацией значений x. Влияние этого эффекта на точность восстановления функции зависит от n и от скорости убывания коэффициентов Фурье θ_k с ростом $|k|$, которая в свою очередь связана со свойствами гладкости периодического продолжения функции $f(x)$.

Приведем пример программы построения вектора $\widetilde{Y}$ значений аппроксимирующего тригонометрического полинома степени $r = 3$ для функции $f(x) = 1/\left(1 + (x - T/2)^4\right)$, заданной на интервале $[0, T]$, $T = 2$ по измерениям в $n = 100$ точках со случайными ошибками со среднеквадратичным отклонением $\sigma = 0.05$.

Вначале создаем функцию пользователя

```
function y=f5(x,T);
y=1./(1+(x-T/2).^4);
```

Сохраняем этот файл с именем `f5.m` (напомним, что имя файла обязано совпадать с именем функции).

Программа построения вектора $Y1 = \widetilde{Y}$ может иметь следующий вид.

```
T=2;  n=100;  r=3;  sigma=0.05          % исходные данные
X=0:T/n:T-T/n;
Y0=f5(X,T);                  % Вектор значений функции
Y=Y0+sigma*randn(1,n);       % Вектор "измеренных" значений
Z=fft(Y);                    % ДПФ "измеренного" вектора
Z1=Z .* ((1:n<=r+1) | (1:n>=n-r+1));   % Фильтрация
Y1=ifft(Z1);      % Значения тригонометрического полинома
sn=sqrt(sum(abs(Y-Y1).^2)/(n-2*r-1))   % Оценка уровня шумов
```

```
plot(X,Y0,'--', X,Y,'*', X,Y1), grid   % Контроль результатов
delta=sqrt(sum(abs(Y0-Y1).^2)/n)       % Ошибка аппроксимации
```

Строка, реализующая фильтрацию, содержит команду логического сложения "|", связывающую два строковых логических выражения. Напомним, что результатами логических выражений являются числа 0 и 1, поэтому массив Z1 будет совпадать с массивом Z только по нужным индексам, а все остальные его элементы окажутся нулевыми. Команда abs, использованная при оценке ошибок, нужна потому, что из-за погрешностей округления оцениваемые разности будут содержать мнимые компоненты (пусть и очень маленькие).

Величина delta, вычисляемая в последней команде, оценивает среднеквадратическую ошибку аппроксимации $\delta_{n,r}$ функции $f(x)$ в точках наблюдения:

$$\delta_{n,r}^2 = \sum_{l=1}^{n} E\left(f(x_l) - f(x_l, \widetilde{\theta})\right)^2.$$

Эту ошибку нельзя вычислить по измеренным данным, так как точные значения $f(x_l)$ неизвестны. Тем не менее полезно выяснить, как зависит эта ошибка от степени r полинома, если мы не знаем заранее, что неизвестная функция есть тригонометрический полином некоторой степени. Можно показать, что

$$\delta_{n,r}^2 = \frac{(2r+1)\sigma^2}{n} + \sum_{r<|k|\leqslant n/2} |\widetilde{\theta}_k|^2, \qquad (10)$$

где коэффициенты $\widetilde{\theta}_k$ определяются (9). Первое слагаемое в (10) характеризует случайную ошибку, связанную с наличием шумов, и оно возрастает с ростом r. Второе слагаемое - систематическую ошибку, связанную с отбрасыванием высоких частот при аппроксимации (т.е. слагаемых в разложении с номерами $|k| > r$), и оно убывает с ростом r. Из этой зависимости видно, что существует оптимальное значение r^*, минимизирующее значение ошибки $\delta_{n,r}$ и зависящее от уровня шумов σ, длины выборки n и

свойств функции f. Изучение подобных задач есть предмет непараметрической статистики, в рамках которой, в частности, изучается задача выбора оптимальной степени полинома r^* по экспериментальным данным, и мы не будем здесь их рассматривать подробнее.

Оценим влияние степени r полинома на точность аппроксимации путем моделирования. Для этого модифицируйте приведенную выше программу: выполните часть программы (начиная с ДПФ) в цикле по `r=1:n/2-1` (исключив контроль результатов путем вывода на график), вычислите значения `sn(r)` и `delta(r)` и выведите их на график в зависимости от r.

Из графика видно, что теоретическая ошибка вначале резко убывает с ростом r, но затем начинает возрастать и можно определить оптимальное значение r^*, обеспечивающее наименьшую ошибку аппроксимации.

Отметим, что величина $sn(r)$, которая вычисляется по экспериментальным данным, вначале убывает с ростом r, но затем стабилизируется на значении, близком к σ. При этом начало участка стабилизации близко к оптимальному значению r^*. Это наблюдение лежит в основе ряда методов оценки оптимального значения r^* по экспериментальным данным.

Список литературы

1. Боровков А. А. Теория вероятностей. – М.: Наука, 1986.
2. Боровков А. А. Математическая статистика. – М.: Наука, 1984.
3. Боровков А. А. Математическая статистика. Дополнительные главы. – М.: Наука, 1984.
4. Ибрагимов И. А., Хасьминский Р. З. Асимптотическая теория оценивания. – М.: Наука, 1979.
5. Ивченко Г. И., Медведев Ю. И. Математическая статистика. – М.: Высшая школа, 1984.
6. Иглин С.П. Математические расчеты на базе MATLAB. БХВ-Петербург, 2005.
7. Кетков Ю.Л., Кетков А.Ю., Шульц М.М. MATLAB 7: программирование, численные методы. БХВ-Петербург, 2005.
8. Кокс Д., Хинкли Д. Теоретическая статистика. – М.: Мир, 1978.
9. Мартынов Н. Н. MATLAB 7. Элементарное введение. – КУДИЦ-Образ, 2005.
10. Мэтьюз, Джон Г., Финк, Куртис Д. Численные методы. Использование MATLAB. – Вильямс, 2001.
11. Половко А.М., Бутусов П.Н. MATLAB для студента. – БХВ-Петербург, 2005.
12. Феллер В. Введение в теорию вероятностей и ее приложения. Т 1, 2. – М.: Мир, 1984.
13. Härdle, W. Applied Nonparametric Regression. – Cambridge Univ. Press, 1900.
14. Tsybakov, Alexandre B. Introduction to Nonparametric Estimation. – Springer Series in Statistics, 2009

Ингстер Юрий Измайлович
Михеев Артём В.
Солнышкин Сергей Николаевич
Чирина Анна В.

Основные алгоритмы численного анализа. Статистическое моделирование в пакете MATLAB

Методические указание

Редактор И. Г. Скачек

Подписано в печать . Формат 60 × 84 1/16. Бумага офсетная.
Печать офсетная. Гарнитура „Times“. Печ. л. 2,0.
Тираж 200 экз. Заказ .

Издательство СПбГЭТУ „ЛЭТИ“
197376, С.-Петербург, ул. Проф. Попова, 5