Софийски Университет Св.Климент Охридски
Факултет по математика и информатика
Вероятности, Операционни изследвания и Статистика

доц. ДИМИТЪР Л. ВЪНДЕВ

Записки
по
Приложна статистика 2

СОФИЯ, юни, 2003

Съдържание

Предговор

Тези лекции са продължение на лекциите от (**?**), Това е вариант от 2003 г. Извинявам се за за закъснението и обърквaнето.

Тема 13

Корелационна матрица

Както видяхме в предните лекции посветени на регресията, един от главните изчислителни проблеми на регресията е лошата обусловеност на матрицата $(X'X)$. Проблемът не е само изчислителен, но е статистически, тъй като обратната на тази матрица влиза в всички оценки на дисперсията. Ще разгледаме:

- проблемите с матрицата $(X'X)$;
- начините за пресмятане на корелационна матрица;
- начините за корекция (хребетова регресия).

13.1 Проблеми с $X'X$

Нека изследваният модел е от вида

$$y = Xa + \varepsilon, \tag{13.1}$$

където $y, \varepsilon \in R^n, a \in R^m, X \in R^n \times R^m$, грешките $\varepsilon \in N(0, \sigma^2 I)$. Тук y и X са наблюденията, а σ^2 и a са неизвестни. Съгласно теоремата на Гаус-Марков оценката се получава във формата:

$$\hat{a} = (X'X)^{-1}X'y \tag{13.2}$$

$$cov(\hat{a}) = \hat{\sigma}^2(X'X)^{-1} \tag{13.3}$$

$$\hat{\sigma}^2 = \frac{1}{n-m}\|y - X\hat{a}\|^2. \tag{13.4}$$

Ако векторът ε има нормално разпределение, то неизместената оценка по метода на най-малките квадрати $\hat{a}$ за вектора a има минимална дисперсия за класа на всички неизместени оценки на този вектор.

При отсъствие на предположението за нормалност на вектора ε, то тази оценка има минимална дисперсия за стеснения клас на *линейните* неизместени оценки. Оттук следва, че оценката $\hat{a}_j$ на параметъра a_j ще има минимална дисперсия за съответния клас оценки, но това изобщо не означава, че самата дисперсия ще бъде малка.

По-точно, ако матрицата $X^{'}X$ е близка до изродена, така че най-малкото и собствено число, например λ_{p-1}, е близко до нула, то "пълната дисперсия"

$$\sum_{j=0}^{p-1} var[\hat{a}_j] = \sigma^2\, tr[(X^{'}X)^{-1}] = \sigma^2 \sum_{j=0}^{p-1} \lambda_j^{-1} > \sigma^2 \lambda_{p-1}^{-1}$$

може да се окаже твърде голяма за практически цели. (Тук сумирането е от 0 до $p-1$, за да се отчете наличието на константа в модела, при запазване на броя на параметрите p.)

Освен това, че при малки собствени числа на матрицата $X^{'}X$ има чисто изчислителни трудности при обръщането и, проблеми възникват и при работа с ковариационната матрица на вектора на оценките $cov(\hat{a}) = (X^{'}X)^{-1}\hat{\sigma}^2$.

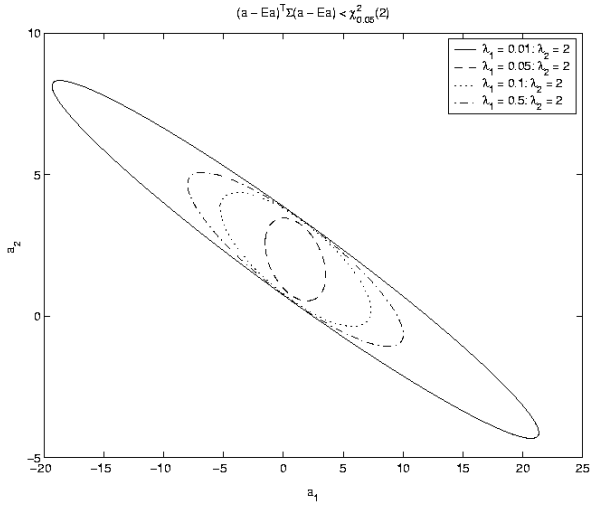
Например ако искаме да построим доверителен интервал за a, конструирането му става по следния начин:

$$(\hat{a} - a)^{'}(X^{'}X)^{-1}(\hat{a} - a)\hat{\sigma}^2 < \mathbf{Q}_{\alpha,\chi^2_{\mathbf{p}}},$$

където $\mathbf{Q}_{\alpha,\chi^2_{\mathbf{p}}}$ е α-процентният квантил на χ^2 разпределение с p степени на свобода.

Горният израз задава доверителна елипса за стойностите на параметъра a с ниво на съгласие α (т.е. с вероятност $(1-\alpha)$ векторът на параметрите a се покрива от тази елипса). Ако обаче матрицата $(X^{'}X)^{-1}$ е "лоша" в смисъл, че някои от собствените и числа са малки, то елипсоидът може да бъде сплеснат. При това положение може да се получи ситуация, за която самата доверителна област не е твърде голяма за фиксирано ниво на съгласие, но въпреки всичко границите за отделните параметри са големи.

В такава ситуация можем да направим некоректното заключение, че оценените параметри са незначими, макар че истинският проблем е от изчислителен характер. Геометрична илюстрация на това за случая в $\mathbf{R}^2$ е дадена на фиг.13.1.



Фиг. 13.1: Доверителна елипса за различни стойности на собствените числа на матрицата $\Sigma = (X^{'}X)^{-1}$

13.2 Пресмятане на R и примери

Тук ще се спрем на възприетите в програмите методи за пресмятане на крос-продукт и корелационна матрица.

13.2.1 Мащабиране

Първото необходимо действие (и така са написани всички статистически програми) е центрирането и стандартизирането на данните X и Y. Тогава матрицата $(X'X)$ става корелационна - с единици по диагонала. Нейното поведение става значително по предсказуемо. Освен това записването на регресионния модел (със свободен член) става в пространство с единица по-ниска размерност.

Една от причините за лошата обусловеност на матрицата $(X'X)$ може да бъде това, че в различните стълбове на X са записани числа с различна размерност. Когато моделът е със свободен член (intercept), обикновено първата колонка на X се състои само от единици. Това допълнително увеличава вероятността за израждане.

Изходът е в предварително правилно мащабиране на колоните на X с трансформацията $\tilde{x}_{ij} = (x_{ij} - \overline{x}_j)/\sigma_j$. Така получената матрица, обаче може да има "нулеви" колони. Такъв е случая със свободния член – съответната колона се нулира. Нека отстраним всички такива колони и означим получената матрица с $\tilde{X}$. Сега моделът може да се запише еквивалентно във формата:

$$\tilde{y} = \tilde{X}b + e, \tag{13.5}$$

Преходът от едните към другите параметри е очевиден и няма да се спираме на него тук. Да отбележим, че сега матрицата $R = \tilde{X}'\tilde{X}$ формално става почти "корелационна" - с $n-1$ по диагонала и остава положително полу-определена.

Като мярка за степента на близост до изроденост е възприето да се използува отношението h на минималното към максималното собствени числа - наречено число на обусловеност.

Пример 13.1 *Лоша матрица. Да разгледаме простия полиномен модел:*

$$y_i = a_0 + a_1 i + a_2 i^2 + a_3 i^3 + e_i, \quad i = 1, 2, \ldots, 9.$$

Да пресметнем двете матрици $X'X$ и R и степента им на изроденост.

$$X'X = \begin{pmatrix} 9 & 45 & 285 & 2025 \\ 45 & 285 & 2025 & 15333 \\ 285 & 2025 & 15333 & 120825 \\ 2025 & 15333 & 120825 & 978405 \end{pmatrix}, \quad h = 1.3015e-007$$

Резултатът е получен със следната "програма" на MATLAB:

```
e0=ones(9,1); e1=[1:9]'; e2=e1.*e1; e3=e1.*e2;
x=[e0,e1,e2,e3]; ss=x'*x;
[u,d,v]=svd(ss); h=d(4,4)/d(1,1)
```

Продължаваме със следните "изчисления" пак на MATLAB:

```
y=x-e0*mean(x); y=y(:,2:4);tx=y./(e0*std(y));
ss=tx'*tx; [u,d,v]=svd(ss); h=d(3,3)/d(1,1)
```

Получаваме следния резултат:

$$(n-1)*R = \begin{pmatrix} 8.0000 & 7.8022 & 7.4392 \\ 7.8022 & 8.0000 & 7.8989 \\ 7.4392 & 7.8989 & 8.0000 \end{pmatrix}, \qquad h = 2.1932e-004 \tag{13.6}$$

Резултатът е убедителен - с просто скалиране на променливите подобрихме с 3 порядъка обусловеността на матрицата $X'X$. Матрицата, изписана в равенство (13.6), е само пропорционална на корелационната, но това не оказва влияние на нейната обусловеност.

13.2.2 Плаващо средно

Този метод е особено подходящ за пресмятане на средна стойност и крос-продукт матрица при масиви с огромен размер. Независимо от натрупването на грешки по време на пресмятанията, той дава добри резултати. Даже прекъсвания на входния поток информация не повреждат вече получените резултати.Използува се в пакета BMDP (Dixon 1981). Всяко наблюдение се задава с вектор $x(i)$ и тегло $w(i)$. Обикновено $w(i) = 1$. В паметта се поддържат:

- сумата от теглата w;
- текущия вектор на плаващите средни x;
- текуща крос-продукт матрица s.

В началото всички тези променливи се нулират. Една стъпка на алгоритъма се задава със следната програма:

```
w = w+w(i)
d = x(i)-x
x = x + d/w
s = s + w*d*(x(i)-x)'
```

13.2.3 Загубени стойности

На практика често се срещат непълни данни. В определени позиции на матрицата X отсъствуват по различни причини измерените величини. Прието е тези позиции да се наричат загубени (missing values). Нека положим $K_j = \{k : x_{k,j}$ не е загубено наблюдение$\}$. Ако няма загубени стойности, то и $K_j = \{1, 2, \ldots, n\}$.

Средните стойности $\overline{x}_j$, например, пресмятаме по формулата:

$$\overline{x}_j = \frac{1}{\#(K_j)} \sum_{k \in K_j} x_{k,j}, \tag{13.7}$$

Извадъчната корелационна матрица се пресмята, най-общо казано, по формулата:

$$r_{i,j} = \frac{\sum_{k \in K_{i,j}} (x_{ki} - \overline{x}_i)(x_{k,j} - \overline{x}_j)}{\sqrt{\sum_{k \in K_i} (x_{k,i} - \overline{x}_i)^2 \sum_{k \in K_j} (x_{k,j} - \overline{x}_j)^2}},$$

където $K_{i,j}$ са подходящо избрани подмножества на множеството от възможни наблюдения. Ако няма загубени стойности $K_{i,j} = \{1, 2, \ldots, n\}$.

В Statistica се предприемат специални мерки при пресмятането на корелационна матрица, когато в данните се срещат загубени стойности.

- (casewise) изключват се от пресмятанията всички наблюдения, в които има загубени стойности ($K := \bigcap_m K_m, K_j := K, K_{i,j} := K$);
- (pairwise) смятане по двойки ($K_{i,j} := K_i \bigcap K_j$);
- (substitution by means) заместване със средна.

Първата възможност може (и трябва) да се използува при наличие на достатъчен брой пълни наблюдения. Тя осигурява "правилна" корелационна матрица - положително определена. Към втората и (особено третата) прибягваме в противен случай.Да отбележим, че професионалните пакети като BMDP (Dixon 1981) предлагат специални програми за "възстановяване" на загубените стойности базирани в това число и на регресионен анализ.

13.3 Методи за корекция

Когато, след всички предприети по-горе мерки, не сме в състояние да получим "добра" корелационна матрица за предикторите, т.е. тя е почти изродена, а всички предиктори ни трябват за интерпретацията на регресионния модел, преминаваме към хребетова регресия.

В следващите сметки винаги ще предполагаме, че работим с корелационна матрица $R = (X'X)$, т.е. данните са центрирани и нормирани. Първо да отбележим, че матрицата R може да бъде представена във формата

$$R = UDU', \quad D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & d_m \end{pmatrix} \tag{13.8}$$

Ковариационната матрица на оценката има вида:

$$cov(\hat{a}) = \sigma^2 (X^{'} X)^{-1} = \sigma^2 R^{-1}.$$

Средно квадратичната грешка на стандартната оценка е ($R^{-1} = UD^{-1}U'$):

$$\mathbf{E} \, \|\hat{a} - a\|^2 = tr(cov(\hat{a})) = \sigma^2 \sum_{i=1}^{m} \frac{1}{d_i} \tag{13.9}$$

За да намалим тази грешка вместо стандартната $\hat{a} = R^{-1}X'Y$, да разгледаме нова оценка: $\tilde{a} = R^{-}X'Y$. Матрицата R^{-} ще определим по-късно. За тази нова оценка имаме:

$$\begin{aligned} \mathbf{E}\,\tilde{a} &= R^{-}X'\mathbf{E}\,(Xa+E) = R^{-}Ra, \\ cov(\tilde{a}) &= \sigma^2 R^{-}RR^{-}. \end{aligned} \tag{13.10}$$

Сега за средно квадратичната грешка получаваме:

$$\begin{aligned} \mathbf{E}\,\|\tilde{a}-a\|^2 &= tr(cov(\tilde{a})) + \|\mathbf{E}\,\tilde{a}-a\|^2 \\ &= \sigma^2\, tr(R^{-}RR^{-}) + \|(R^{-}R-I)a\|^2. \end{aligned} \tag{13.11}$$

За определянето на подходяща матрица R^{-} се използуват два метода.

13.3.1 Метод на Marquardt

Нека $R^{-} = (R+\varepsilon I)^{-1}$, където I е единичната матрица, а ε е подходящо подбрана константа. Нека пресметнем нейната средно квадратична грешка съгласно формула (13.11).

$$\begin{aligned} \mathbf{E}\,\|\tilde{a}-a\|^2 &= \sigma^2\, tr(R^{-}RR^{-}) + \|(R^{-}R-I)a\|^2 = \\ &= \sigma^2 \sum_{i=1}^{m} \frac{d_i}{(d_i+\varepsilon)^2} + \varepsilon^2 \|R^{-}a\|^2. \end{aligned} \tag{13.12}$$

Тук използувахме, че

$$R^{-}R - I = (R+\varepsilon I)^{-1}(R+\varepsilon I) - (R+\varepsilon I)^{-1}\varepsilon - I = -\varepsilon R^{-}$$

Така подбирайки ε във формула (13.12) можем да постигнем значително намаление на грешката по сравнение с формула (13.9). Да отбележим, че $\|R^{-}a\|^2 \le \|R^{-1}a\|^2$, което е фиксирано число макар и неизвестно.

Методът на Marquardt така, както беше показан тук, всъщност е частен случай на по-общия метод на Levenberg-Marquardt. Последният е модификация на метода на Нютон-Гаус. Повече подробности за тези методи и приложенията им в регресионния анализ могат да бъдат намерени в (Демиденко 1989).

13.3.2 Главни компоненти

Казано накратко, анализът на главните компоненти включва математически процедури, трансформиращи известен брой (евентуално) корелирани променливи в помалко на брой некорелирани променливи, наречени главни компоненти.

Главните компоненти се подреждат според степента, с която описват поведението на първоначалните променливи – първите главни компоненти са с най-голямо значение, всеки следващ обяснява все по-малка част от "неописаните" зависимости.

Традиционно методът на главните компоненти се прилага при описването на зависимостите присъстващи в ковариационна матрица (за повече подробности вж. глава Факторен анализ).

Друг проблем, който налага използването на този метод е подобрената интерпретация на регресионния модел подобна на тази във факторния анализ - откликът се представя като функция от нови изкуствени променливи - фактори.

Нека сега зафиксираме някое $k < n$ и определим матрицата R^- по следния начин:

$$R^- = \begin{pmatrix} d_1^{-1} & 0 & 0 & \cdots & 0 & 0 & 0 \cdots & 0 \\ 0 & d_2^{-1} & 0 & \cdots & 0 & 0 & 0 \cdots & 0 \\ \cdots & & & & & & & \\ 0 & 0 & 0 & \cdots & 0 & d_k^{-1} & 0 \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \cdots & 0 \\ \cdots & & & & & & & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \cdots & 0 \end{pmatrix}.$$

Нека пресметнем средно квадратична грешка на оценката $\tilde{a} = R^- X'Y$ съгласно формула (13.11).

$$\begin{aligned} \mathbf{E}\,\|\tilde{a} - a\|^2 &= \sigma^2\, tr(R^- R R^-) + \|(R^- R - I)a\|^2 = \\ &= \sigma^2 \sum_{i=1}^{k} \frac{1}{d_i} + \| - U D_{n-k} U' a\|^2. \end{aligned} \tag{13.13}$$

Тук матрицата D_{n-k} е диагонална с $(n-k)$ единици накрая в диагонала.

И пак, като подбираме k можем така да балансираме двете грешки във формула (13.13), че сумата им да бъде значително по-малка от тази на формула (13.9).

Тема 14

Факторен анализ

Факторният анализ е възникнал като задача при обработката на данни от психосоциологични анкети. При тези проучвания обикновено се предполага, че изследваното явление се описва с неизвестен (малък) брой фактори, които не можем да измерваме директно. Вместо тях можем да наблюдаваме голям брой променливи, които са функции от тях (за простота тези функции обикновено се приемат за линейни).

14.1 Обзор

Търси се представяне на данните във формата:

$$X = FL + E. \tag{14.1}$$

Тук $X \in R^{n\times m}$ е матрицата на центрираните и нормирани наблюдения, а $E \in R^{n\times m}$ - грешките.

Във факторния анализ са приети следните названия:

- Факторни тегла се наричат коефициентите на разлагане на оригиналните променливи по факторите - $L \in R^{k\times m}$.
- Факторни стойности се наричат оценените стойности на факторите за всяко наблюдение - $F \in R^{n\times k}$.
- Общност - това е относителната дисперсия на всяка променлива в новото ѝ описание. Т.е. общностите показват доколко стойностите на конкретната променлива могат да бъдат предсказвани или възстановени от факторните стойности.

Основните задачи на факторния анализ са:

1. Да се определи броят на факторите, достатъчни за описание на явлението (втората размерност на матрицата F);
2. Да се намерят факторите. Това означава да се изчислят коефициентите на линейните функции L, представящи променливите (посредством факторите) и стойностите за така определените фактори върху отделните обекти F.

3. Да се интерпретират получените резултати в термини на предметната област на данните. Разбира се, последната задача трябва да се решава съвместно от статистик и съответен специалист.

Такава ситуация не е изключение. В статистическите изследвания се получават разумни резултати само при тясно сътрудничество между специалистите от двете области.

14.1.1 Методики на факторния анализ

За да навлезем в кухнята на факторния анализ трябва да поясним какво разбираме под термина разсейване на извадката. Ако си представим наблюденията като облак от точки в пространството, то естествена характеристика на този облак биха били неговата форма и размер. Те се отразяват до известна степен в ковариационната матрица. Естествено е, че тази характеристика не би трябвало да зависи от координатните системи, в които са правени измерванията. Единствената инвариантна в този смисъл статистика е корелационната матрица.

Точно тя е и обект на класическия факторен анализ. (Методика А). Пълната дисперсия на извадката не се използува от тази методика.

Съществуват обаче редица варианти на факторния анализ, които използуват други аналогични матрици, като интерпретацията е естествено различна. Използува се факта, че изчислителните му процедури се нуждаят съществено само от свойството симетричност на корелационната матрица.

Интересен пример за използуване на факторния анализ е така наречената "методика В". Тя се състои в замяна на редовете и стълбовете в матрицата от данни - т.е. разглеждане на наблюденията като променливи. Получените от такава матрица корелации говорят за близостта на наблюденията, а факторите определят групи от близки обекти. Естествено е използуването на тази методика, когато имаме голям брой променливи и неголям брой наблюдения.

Друга възможност се появява с наличието на групиращи променливи. Тогава използуването на съответните аналози на корелацията дава възможност за редица нови интерпретации на получените променливи. Например, използуването на вътрешно групови корелации отразява разсейването на данните, след като е отстранено влиянието на групиращите променливи. Факторите в този случай описват променливи, които са центрирани с груповите средни. Когато пък използуваме между групови корелации, всъщност описваме разсейването точно на тези групови средни стойности и по този начин влиянието на групиращите променливи. Тези особености определят интерпретацията на получените с факторния анализ резултати. Накрая ще подчертаем, че процедурата на факторния анализ не е напълно формализирана от статистическа гледна точка. Все още няма статистически критерий за проверка на правилността на получените решения. Единствен критерий за адекватност на получените модели е тяхната практическа използваемост.

14.2 Главни компоненти

Предложеният от Хотелинг (Harman 1972) метод на главните компоненти позволява да определим доколко определен брой фактори описват явлението, т.е. донякъде да решим задача 1. Той има и това достойнство, че същевременно решава и втората задача. Фактически в този метод тя се решава преди първата.

Главните компоненти съответстват на осите на елипсоида на разсейване на точките, представящи обектите (данните), в пространството на наблюдаваните променливи. Техният брой е равен на броя на променливите. Главните компоненти могат да бъдат разглеждани като фактори, т.е. като нови променливи, които изцяло описват извадката и са независими. Тази независимост позволява да представим общата дисперсия на извадката като сума от дисперсиите на новите променливи.

Отстранявайки част от факторите, тези които имат незначителна дисперсия (разсейване) можем да дадем задоволително описание на свойствата на извадката с намален брой нови променливи.

14.2.1 Корелационна матрица

Корелационната (или ковариационна) матрица на случаен нормален вектор в m - мерното пространство е симетрична и неотрицателно определена. Следователно тя може да бъде представена в т.н. спектрална форма:

$$S = \sum_{i=1}^{m} a_i e_i e_i' \tag{14.2}$$

Тук числата $\{a_1, a_2, \ldots, a_m\}$ се наричат собствени числа, а векторите $\{e_i, i = 1, 2, ...m\}$ - собствени вектори, защото удовлетворяват уравнението:

$$Se_i = a_i e_i \tag{14.3}$$

Те са ортогонални, с единична норма и образуват базис в m-мерното евклидово пространство. Представянето (14.2) е единствено, ако сред собствените числа $a_{(.)}$ няма еднакви.

14.2.2 Метод на главните компоненти

Хотелинг е показал, че решението на задача 2., формулирана в първия параграф, при фиксирано k се дава от подпространството, образувано от първите k собствени вектора. Действително, ако разгледаме цялото разсейване на извадката - числото $S^2 = \sum ||x_i||^2$ (предполагаме центрираност на данните), то естествено е да търсим това подпространство H с размерност k (за фиксирано $k, 1 \leq k \leq m$), за което експерименталните точки биха се преместили минимално при своето проектиране P върху него. Тогава най-малко би се изменила и дисперсията S^2, което се вижда от тъждеството:

$$S^2 = \sum ||x_i||^2 = \sum ||P_k x_i||^2 + \sum ||x_i - P_k x_i||^2$$

14.2.3 Брой на факторите

Изборът на броя на факторите k в представянето на информацията за извадката, обикновено се прави от изискването възможно най-пълно да бъде представено общото разсейване на извадката ((Yan and Wale 1974), стр.102):

$$\sum_{i=1}^{n} ||P_k x_i||^2 \geq .95 \sum_{i=1}^{n} ||x_i||^2 \tag{14.4}$$

Друг, също често срещан критерий на практика, е избора само на собствените вектори със собствени стойности, които са по-големи от единица. Наистина, когато използуваме корелационна матрица $\sum a_i = tr\, R = m$. Той се нарича критерий на Кайзер. Всъщност, само окончателните резултати могат да оправдаят правилния избор на k.

14.3 Въртения на факторите

Когато броят на факторите е избран, решението получено по метода на главните компоненти може да бъде записано така 14.1:

$$X = FL + E. \tag{14.5}$$

Тук $X \in R^{n\times m}$ е матрицата на центрираните и нормирани наблюдения, $F \in R^{n\times k}$ - са факторните стойности, $L \in R^{k\times m}$ - са факторните тегла, а $E \in R^{n\times m}$ - грешките.

Решението по метода на Хотелинг дава минимум на функционала $tr\, E'E$. Наистина, нека разгледаме следното представяне на матрицата $X = UDV$, обикновено се нарича svd или singular value decomposition. Тук $U \in R^{n\times n}$ и $V \in R^{m\times m}$ са ортогонални, а $D \in R^{n\times m}$ - "диагонална" с неотрицателни елементи ($d_{i,i} \geq d_{i+1,i+1}$).

14.3.1 Дуално представяне

Тогава имаме следното представяне на ковариационната матрица

$$X'X = V'D'U'UDV = V'D'DV = \sum_{i=1}^{m} d_{i,i}^2 v_i v_i', \tag{14.6}$$

Аналогично е представянето на матрицата XX', но собствените вектори са в друго пространство. Имаме и $d_{i,i}^2 = \lambda_i$.

Така като фиксираме k и означим с $()_k$ операцията "орязване"на първите k стълба на една матрица, получаваме $F = n^{-1/2}(UD)_k$, $L' = (V')_k$. Числата $d_{i,i}^2/n$ са дисперсиите на факторите, а $\sum_{i=1}^{k} l_{i,j}^2$ са така наречените *общности* на променливите. Те обясняват каква част от тяхната дисперсия е обяснена с така подбраните фактори.

Матрицата $L \in R^{k\times m}$ се нарича понякога "сурова" и много често се представя в нормиран вид - стълбовете и се нормират с квадратен корен на общностите. Това облекчава интерпретацията.

14.3.2 Вариамаксна трансформация

Така избраните фактори са линейни функции на оригиналните променливи, но е трудно да бъдат интерпретирани. За по-лесно решаване на задача 3 - намиране на удобни за интерпретация фактори, съществуват редица други методи. Един от най-прилаганите е така наречената вариамаксна трансформация. За да изложим идеята й ни е необходима малко терминология.

Сред многото методи, облекчаващи интерпретацията на факторите, може би най-популярен е методът на *варимакс* - трансформацията ((Yan and Wale 1974), стр.122). Идеята на варимакс - трансформацията е така да бъдат променени факторите, че те да запазят добрите си свойства (пълнотата на описание на извадката като цяло) и да получат по-добри интерпретативни качества. Преразпределят се факторните тегла като големите нарастват, а малките намаляват, което води до това, че всеки фактор се обяснява от по-малко на брой променливи. Разбира се, за това се плаща. Факторите вече не са така независими. Променят се общностите на променливите.

Той се състои в максимизиране на функционала:

$$\max_K var(KL) = \max_K \sum_{i=1}^{k} \sigma^2(i), \tag{14.7}$$

по всевъзможния избор на факторите във вече фиксираното подпространство H. Тук $\sigma^2(i)$ е "дисперсията" на квадратите на (нормираните с общностите или сурови) факторни тегла, а $K \in R^{k\times k}$ е ортогонална матрица. Теглата на променливите в даден фактор са записани по редове в L.

Максимизирането на дисперсията води до увеличаване на разликата между големите и малки тегла. Така се вижда по-добре кои променливи са ясно представени в даден фактор - те получават тегла близки до единица.

Когато максимизираме дисперсията по променливи (по стълбове) на тези тегла, съответния метод се нарича *квартимакс*. Той също може да се прилага към суровите или нормирани тегла. Съществуват варианти на едновременна максимизация на претеглена сума на двата функционала.

- *биквартимакс* - двата функционала са с равно тегло;
- *еквимакс* - варимакс функционалът е с тегло $k/2$.

Тъй като цел на всички тези методи е подобряване на интерпретацията, оправданието за избор на един или друг е чисто субективното впечатление от изводите.

Сериозно математическо изложение на факторния анализ може да се намери в (Harman 1972), а ред математико - статистически критерии в (Г.П.Климов 1975), стр.228 - 240. От приложна гледна точка процедурите му са изложени в (Yan and Wale 1974).

14.4 Пример

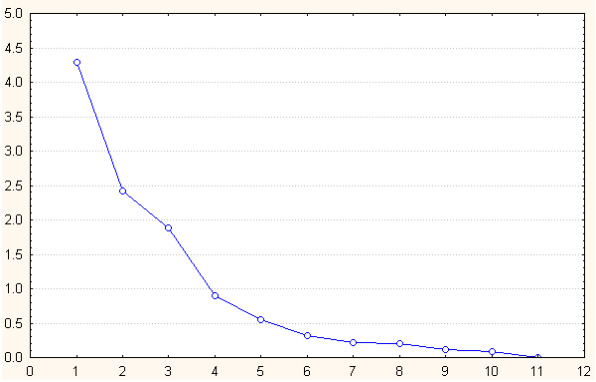
Ще използуваме данни, описани подробно в П1.1. Тези данни са събирани с цел изследване на нивото на престъпност (променлива R). ето защо тя би следвало да се

изследва като функция на останалите променливи, които характеризират донякъде наблюдаваните обекти (щати).

С помощта на факторния анализ ще потърсим по-компактно описание на щатите. Това описание може да се използува после и в регресионния анализ за изучаване на влиянието на така установените фактори върху престъпността. По тези причини променливата R засега ще бъде изключена от данните. Променливата S (северни щати - 0, южни - 1) е естествено да се разглежда като групираща.

Междугруповата корелация се използува, когато броят на групите е голям и се интересуваме от влиянието на групиращите променливи. Ковариационни матрици във факторния анализ се използуват главно в методика B или когато променливите са измерени в еднакви скали.

Ще изследваме последователно и двете смислени възможности - с или без групиращата променлива. Нека първо изпълним ГЛАВНИ КОМПОНЕНТИ - намираме собствените вектори и собствени числа на избраната матрица.



Фиг. 14.1: БРОЙ

Следва да определим БРОЙ НА ФАКТОРИТЕ . Виждаме, че само първите четири главни компонента са с по-големи от единица собствени стойности и изчерпват 85% от общото разсейване.

Eigenvalues (coxsnell.sta) Extraction: Principal components				
	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	5,651763	43,47510	5,65176	43,47510
2	2,520225	19,38635	8,17199	62,86145
3	1,905915	14,66089	10,07790	77,52234
4	,975680	7,50523	11,05358	85,02757

Таблица 14.1: Брой на факторите

Както се вижда, добавянето на пети фактор няма да подобри съществено описанието на данните. Избраните четири фактора подлагаме на ВАРИМАКС трансформация. Резултатът - факторната матрица изглежда така:

Factor Loadings (Varimax normalized) (coxsnell.sta) Extraction: Principal components				
	Factor 1	Factor 2	Factor 3	Factor 4
AGE	-,754775	-,335004	-,216132	,131880
ED	,714933	,301968	-,165533	,456024
EX0	,271013	,936232	,025773	,071201
EX1	,286739	,930553	,014500	,058661
LF	,301987	,057314	-,422428	,676215
M	,143431	-,083612	,208971	,923847
N	-,026020	,686080	,067863	-,42640
NW	-,911532	,071407	-,015773	-,15538
U1	,148925	-,116386	,933160	,163664
U2	-,064235	,215195	,913280	-,16023
W	,674558	,649078	,011219	,145555
X	-,805650	-,431225	,032204	-,11271
S	-,847109	-,132825	,011559	-,25299
Expl.Var	4,027817	3,116707	2,008122	1,90093
Prp.Totl	,309832	,239747	,154471	,146226

Таблица 14.2: Факторни тегла

Подобен резултат получаваме и като използуваме общата корелационна матрица (без включена S).

Factor Loadings (Varimax normalized)(coxsnell.sta) Extraction: Principal components				
	Factor 1	Factor 2	Factor 3	Factor 4
AGE	-,311928	,096624	-,207659	-,779254
ED	,267644	,491360	-,159864	,717398
EX0	,921361	,089664	,026148	,302728
EX1	,913643	,076772	,014232	,321696
LF	,065698	,717774	-,404737	,227964
M	-,105279	,920085	,230250	,114229
N	,720774	-,397862	,061881	-,035906
NW	,097894	-,195179	-,012646	-,898593
U1	-,116102	,149181	,936299	,131003
U2	,219377	-,180263	,910103	-,045086
W	,616907	,177462	,009573	,705527
X	-,398354	-,146393	,036147	-,822760
Expl.Var	2,997946	1,930522	1,996722	3,384701
Prp.Totl	,249829	,160877	,166393	,282058

Таблица 14.3: Факторни тегла без S

Нека сравним тези две факторни матрици. За целта да отбележим в таблиците големите (по-големи от .65) факторни тегла. Веднага се вижда, че факторите (колоните) са разместени, но останалите разлики между двете матрици са незначителни. С други думи, групирането по северни и южни щати не оказва влияние на направлението на факторите, а само променя дисперсиите им.

Да се опитаме да направим интерпретация на така получените фактори, като се придържаме към номерацията в матрицата в таблица 14.2 и кондензираме нейното представяне.

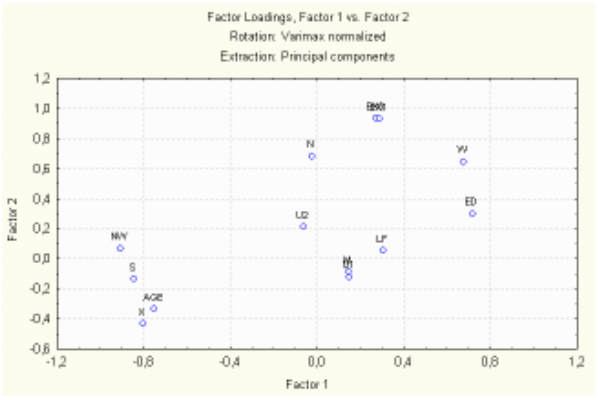
Factor Loadings				
	F1	F2	F3	F4
AGE	-,7			
ED	,7			
EX0		,9		
EX1		,9		
LF				,6
M				,9
N		,6		
NW	-,9			
U1			,9	
U2			,9	
W	,6			
X	-,8			
S	-,8			

Таблица 14.4:

В първия фактор (който така силно зависи от стойностите на S) освен тази променлива участвуват възрастта, ниския образователно ниво, съотношението бели-черни, ниския доход и високата степен на социално неравенство (AGE, ED, NW, X , W, S).

Всичко това може би отразява положението на цветното население в САЩ, което е концентрирано в южните щати. Нека го наречем "север-юг" (N/S).

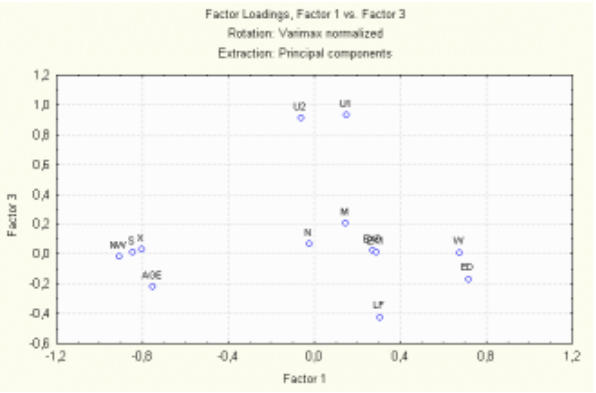
В таблица 14.3 той е FACTOR4, но това е грешка на програмата - би трябвало факторите да са наредени по големината на дисперсията си.



Фиг. 14.2: Фактор2

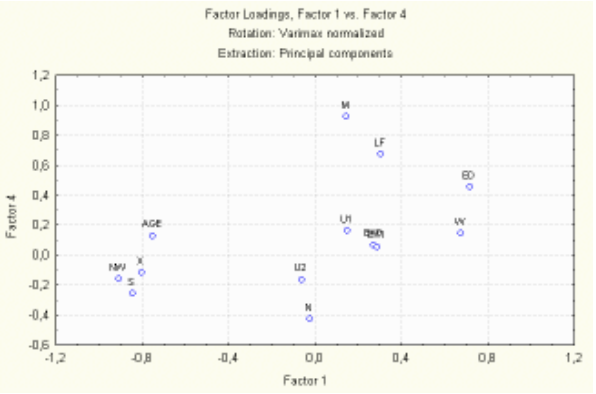
Вторият фактор свързва с големи тегла разходите за полиция, населението, отчасти и средният доход (EX0, EX1, N , W). Имаме известно основание да свържем този фактор с усилията полагани от управлението на щата в борбата с расовите проблеми, престъпността и изобщо поддържането на реда.

Естествено е тези усилия да са пропорционални на средния доход и да се отразяват главно от полицейските разходи. Нека го наречем условно по името на най-силната променлива "полиция" (POLICE). В таблица 14.3 FACTOR1.



Фиг. 14.3: Фактор3

Третият фактор обединява двата вида безработица (U1, U2). Нека го наречем "безработица" (UNEMPLOIMENT).



Фиг. 14.4: Фактор4

Последният, четвърти фактор обединява броят на младите мъже, трудовата им активност и донякъде образователното ниво (M, LF , ED).

Като че ли този фактор отразява активността на мъжете в живота, трудовия потенциал на щата, но негативно зависи от броя на населението. Да го наречем активност (ACTIVITY) и отбележим, че тя се проявява предимно при щатите с малко население.

В таблица 14.3 това е FACTOR2.

По този начин създадохме четири нови (изкуствени) променливи и можем да пресметнем техните стойности за различните щати с цел по-нататъшно изследване. Това правим с помощта на матрицата F ФАКТОРНИ СТОЙНОСТИ. Трябва да я запазим във нов файл. Ще запазим заедно с тези 4 фактора (във формата от таблица 14.3) и променливите R и S, за да можем по лесно да интерпретираме резултатите.

Като резултат от регресионната програма получаваме следната таблица:

Regression Summary for Dependent Variable: R						
R= .78582340 RI= .61751842 Adjusted RI= .57087432						
F(5,41)=13.239 p<.00000 Std.Error of estimate: 253.48						
	BETA	St. Err. of BETA	B	St. Err. of B	t(41)	p-level
Intercpt	869.254	59.9711	14.49457	.000000		
S	.130516	.171688	105.4388	138.7006	.76019	.451490
FACTOR1	.725941	.099730	280.9004	38.5902	7.27905	.000000
FACTOR2	.359947	.109889	139.2803	42.5213	3.27554	.002150
FACTOR3	.040019	.096597	15.4851	37.3778	.41429	.680824
FACTOR4	.008797	.161588	3.4040	62.5257	.05444	.956847

Тук се интересуваме само от значимите $BETA$ коефициенти. Така лесно интерпретираме престъпността като предимно зависеща от

1. (POLICE) - променливи (EX0, EX1, N , W),
2. (ACTIVITY) - променливи (M, LF , ED).

Тема 15

Дискриминантен анализ

Тук ще представим една процедура от многомерния анализ на данни базирана на вероятностен модел. Другото название на процедури от този тип е разпознаване на образи.

15.1 Основни понятия

Тази статистическа процедура се използува, когато се нуждаем от "прогнозиране" стойностите на групираща променлива. Понякога това се нарича класификация или разпознаване на образи. Нека нашата извадка е нееднородна или с други думи, се състои от няколко групи наблюдения с различни вероятностни характеристики. Целта ни е да се научим от тази извадка, по зададени параметри на дадено наблюдение, да определим принадлежността му към класа, от който произлиза.

В първата си част, фазата на обучение, процедурата на дискриминантния анализ обработва тази информацията от т.н. обучаваща извадка с цел да я кондензира в тъй наречените решаващи правила. Когато те са получени, естествено е те да бъдат изпробвани върху обектите от обучаващата извадка или върху други обекти с известен клас.

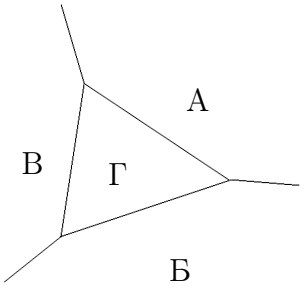
При положение, че тези обекти (или поне достатъчно голям процент от тях) бъдат класифицирани правилно, можем да очакваме, че разпознаващите правила са добри и коректно ще работят и за обекти от неизвестен клас.

Разбира се, конкретното прилагане на различните методики на дискриминантния анализ има ред тънкости. Тук ще се спрем по - подробно на най - разпространената процедура за стъпков линеен дискриминантен анализ. Тя притежава ред недостатъци, но и някои преимущества - в частност дава прости решаващи правила.

В линейния дискриминантен анализ се строят линейни дискриминантни функции от предикторите. За всеки клас има точно една такава функция. Правилото за класификация изглежда така:

Наблюдението се класифицира към класа с максимална дискриминантна функция.

15.1.1 Класификационни области



Областта от стойности на предикторите, при попадане в която наблюденията се класифицират към даден клас, е изпъкнал многоъгълник (възможно отворен). Тя се нарича класификационно множество на класа. При два предиктора и 4 класа класификационните множества биха могли да изглеждат по указания начин.

Фиг. 15.1: Класификационни области

Когато класовете са ясно разграничени, не е трудно те да бъдат отделени. Когато обаче те се пресичат, въпросът за оптимален (с най - малко грешки) избор на класификационни правила е сложен и изисква допълнителна априорна информация.

Линейният дискриминантен анализ предполага, че разпределенията на количествените променливи (предикторите) вътре в класовете са нормални и се различават само по средните си стойности. Тогава процедурата произвежда оптимални решаващи правила. Разбира се, тя може да се използува и при случайна (по групиращата променлива) извадка, но появата на празен клас е недопустима.

Когато броят на количествените променливи е по - голям, за простотата на решаващите правила е съществено да се отберат за предиктори само тези променливи, които носят важната за разделянето информация. В това помага статистиката на Махалонобис. Тя позволява да се провери хипотезата за съвпадане на груповите средни на предикторите като цяло. За простота и тук, вместо критичната област за статистиката, се използува вероятността съответно разпределената случайна величина да не надхвърли стойността на статистиката. Тази вероятност расте докато променливите допринасят за по - доброто разделяне на класовете и започва да намалява, когато предикторите станат твърде много. Естествено, добро разделяне може да се очаква, само когато хипотезата се отхвърля с висока вероятност.

Нека вече са избрани най - добрите променливи за предиктори. Това още не означава, че са построени класификационните множества. Да напомним, че основна цел на дискриминантния анализ е да се получи правило за причисляване на едно ново наблюдение към даден клас. За това наблюдение може да съществува априорна информация за неговата възможна принадлежност към класовете. Прието е такава информация да бъде формулирана в термини на априорни вероятности, които са необходими за определяне на оптимални класификационни правила.

Ако такава информация не съществува, естествено е априорните вероятности на класовете да се приемат за равни. Когато пък извадката е случайна и новото наблюдение се избира по същия начин, може те да се приемат за пропорционални на обема на класовете в обучаващата извадка.

Изборът на априорните вероятности фиксира оптимални дискриминантни функции и класификационни правила. Не е удобно обаче, за всяко ново наблюдение да се въвеждат априорни вероятности. Това е свързано и със значителни изчислителни трудности, особено когато броят на класовете е голям. Един възможен начин за

заобикаляне на това неудобство е представянето на класовете с помощта на няколко групиращи променливи. Такова представяне съответствува и на редица практически задачи. Ако са известни стойностите на поне една от групиращите променливи, това е съществена априорна информация - фиксирането на тази променлива е еквивалентно на задаването на нулева априорна вероятност за поне половината от класовете.

15.2 Вероятностна формулировка

15.2.1 Бейсов подход

Нека допуснем, че са известни вероятностите $\{p(g)\}$, груповите средни $\{m(g)\}$ и вътрешно - груповата ковариационна матрица

$$C(g) = C, \quad (g = 1, 2, \ldots, G)$$

Тогава по формулата на Бейс, апостериорната вероятност за класификация в класа g на наблюдението $(x, .)$ ще бъде

$$q(g) = c.p(g).f(x, m(g), C). \tag{15.1}$$

Тук f е плътността на нормалното разпределение със средна стойност $m(g)$ и ковариационна матрица C, а c е нормираща константа (такава, че $\sum q(g) = 1$).

15.2.2 Класификационни правила

Съгласно принципа за максимално правдоподобие, класифицира се по правилото:

$$\hat{g} = \max h : q(h). \tag{15.2}$$

Класификационните правила могат да бъдат записани във вида:

$$p(\hat{g}).f(x, m(\hat{g}), C) \geq p(h).f(x, m(h), C), \quad (h = 1, 2, \ldots, G). \tag{15.3}$$

След логаритмуване и съкращаване, получаваме:

$$b(\hat{g})'x + a(\hat{g}) \geq b(h)'x + a(h), \quad (h = 1, 2, \ldots, G), \tag{15.4}$$

Векторът $b(g)$ и числото $a(g)$ се получават по формулите:

$$b(h) = m(h)'C^{-1}, \qquad a(h) = logp(h) - m(h)'C^{-1}m(h). \tag{15.5}$$

Оттук се вижда, че в неравенствата (15.4) участвуват линейни функции относно променливите и това обстоятелство е дало името на линейния дискриминантен анализ.

15.2.3 Априорни вероятности. Модели

За оценка на априорните вероятности $\{p(g)\}$ можем да използуваме най - добрите им оценки $\{n(g)/N\}$ при случайна извадка или друга априорна информация. За оценка на $\{m(g)\}$ и C се използуват вътрешно - груповите средни и обединената извадъчна вътрешно - групова ковариация.

Когато групиращите променливи са повече от една, броят на класовете G нараства. Вероятността за поява на празни клетки ($n(g) = 0$) при случайна извадка с ограничен обем рязко се увеличава. Затруднява се и оценката за $\{m(g)\}$. В такива случаи се препоръчва използуването на оценки, получени от линеен модел, като се направят съответните проверки с методите на дисперсионния анализ. Съответно, ще се промени и оценката за C. Аналогично, за оценяване на честотите $n(g)$ могат да се прилагат тъй наречените логаритмично - линейни (log - linear) модели.

15.3 Стъпков дискриминантен анализ

Аналогично на стъпковия регресионен анализ и тук е възприета концепцията за избор на подходящ набор от количествени променливи, с които да построим модела. Единственото средство, което ни трябва са аналозите на P(F-to-enter) и P(F-to-remove). Те се строят аналогично на регресията, но ролята на сумите от квадрати играят

- вътрешно - груповата ковариационна матрица
- между - груповата ковариационна матрица.

Както и в едномерния случай, така и в многомерния е верно следното равенство:

$$\sum_i \sum_j (x_{ij} - \bar{x})(x_{ij} - \bar{x})' =$$

$$\sum_i \sum_j (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' + \sum_i n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

$$SS = SS_{in} + SS_{mod}. \tag{15.6}$$

Матрицата SS_{in} се тълкува като "сума от квадрати", отговаряща на разсейването на данните около техните локални средни и с нейна помощ се строи оценка за вътрешно - груповата ковариационна матрица C:

$$\hat{C} = \frac{1}{N - G + 1} SS_{in}.$$

Лесно се вижда, че по диагонала на матричното равенство (15.6) стои добре известното ни разлагане на сумата от квадрати в дисперсионния анализ.

Матрицата SS_{mod} се тълкува като "сума от квадрати", отговаряща на разсейването на груповите средни и с нейна помощ се строи оценка за между - груповата ковариационна матрица C_{mg}:

$$\hat{C}_{mg} = \frac{1}{G - 1} SS_{mg}.$$

Съгласно концепцията на Фишер един дискриминантен анализ ще бъде толкова по-добър, колкото е по-малко частното на двете детерминанти:

$$\Lambda = \frac{|SS_{in}|}{|SS|},$$

наричано критерии на Уилкс.

При преминаването от дадена размерност k към размерност $k+1$ имаме изменение:

$$\lambda = \frac{\Lambda_{k+1}}{\Lambda_k}$$

Това мултипликативно изменение, разбира се, зависи от това коя променлива сме избрали и колкото по-малко е, толкова по-добре. То се нарича "partial lambda" статистика. (C.R.Rao 1965) е показал, че ако новата променлива не влияе на разделянето статистиката

$$F = \frac{(n-m)}{k}\frac{(1-\lambda)}{\lambda}$$

има разпределение на Фишер. Точно тази статистика служи за избор на променливите за отстраняване и въвеждане в модела точно както в стъпковия регресионен анализ.

По-подробно човек може да се запознае с теорията в книгите на (Wilks 1967) и (T.W.Anderson 1958), а с алгоритмичните в частта на (Jennrich 1977) в сборника (Einslein, Ralston, and Wilf 1977), послужил при създаването на пакета (Dixon 1981).

Тема 16

Нелинейна регресия

Както вече стана ясно в главата Линеен регресионен анализ, термините "линеен" и "нелинеен" се отнасят само за параметрите на модела. По правило, процедурата за нелинейна регресия се използува, когато имаме много основателни съображения за механизма на явлението и вида на модела. Ако моделът на изследваното явление е неизвестен, линейните модели са достатъчно универсални и гъвкави за описание му и са за предпочитане.

16.1 Въведение

Понякога и нелинейните модели могат да се преобразуват в линейна (по параметрите) форма. Например, в антропологията често се използува модела:

$$z = c.x^a.y^b, \tag{16.1}$$

който лесно се преобразува в линеен, чрез логаритмуване на променливите:

$$logz = logc + a.logx + b.logy. \tag{16.2}$$

Такова преобразуване не е възможно обаче за модел от вида:

$$z = c + x^a + y^b. \tag{16.3}$$

В тези случаи се използуват методите на нелинейната регресия. За съжаление, задачата за оценка на нелинейни параметри притежава ред неудобства: изчислителната процедура е много по-трудоемка, интерпретацията на резултатите е много по-несигурна от статистическа гледна точка. Нормалното разпределение на грешката вече не ни осигурява нормално разпределение на оценките при краен брой наблюдения. Това е причината, поради която на всички статистически резултати, свързани с регресията (доверителни граници на коефициентите, проверка на хипотезата за адекватност и др.), трябва да се гледа като на приблизителни (асимптотични), ориентировъчни , не съвсем строги .

Сериозен проблем в нелинейната оптимизация като цяло (и в регресията в частност) е въпросът за *единственост* на полученото решение. Този въпрос е тясно

свързан с избора на начално приближение за оценяваните параметри или за наличната априорна информация за техните стойности.

По своята същност нелинейната регресия е много по-близко до задачата за апроксимация, отколкото до традиционната математическа статистика. Причината е в сравнително по-голямата тежест на изчислителните алгоритми по отношение на статистическите изводи. Както и в линейната регресия, тук се минимизира грешката от модела, като сума от квадратите на остатъците SSr (метод на най-малките квадрати). Намирането на този минимум е сложно, итеративно и е обект на методи за нелинейна оптимизация (вж. (Pollard 1982)). Тези методи се разделят на два големи класа според принципа си на действие.

16.1.1 Градиентни методи

Методите от първия клас използуват предположението за гладкост на минимизираната функция по параметрите (наличие на първи производни). Всички те представляват различни модификации на метода на Нютон - търсене на минимума в посока на градиента. Такива модификации са необходими за увеличаване на скоростта на сходимост при нерегулярни (далечни по форма от квадратичната) функции.

Аналитичното изчисляването на градиента може да се избегне (методи на изкуствения градиент). Естествено това увеличава изчислителната работа, тъй като се използуват крайни разлики за оценки на стойностите на производните. Когато производните са прости функции, целесъобразно е да се използува "естествен" градиент - по-малко изчисления и по-малко допълнителни параметри за процедурата (нарастванията за крайните разлики).

16.1.2 Безградиентни методи

Вторият клас методи се използуват при по-слаби предположения. Например не е необходимо функцията да е диференцируема. Друго положително качество на методите от този клас е сравнително по-добрата им устойчивост в зависимост от избора на началното приближение. Те са по принцип по-бавни. Освен това отсъствието на градиент не дава възможност да се изчисли ковариацията на оценените параметри, т.е. точността на получените оптимални оценки.

При линейната регресия най-информативните експериментални точки в пространството на предикторите винаги са на границата на областта. По-голяма информативност означава по-малка дисперсия на оценките за параметрите при същия брой наблюдения. При нелинейната регресия, поради сложността на апроксимираната функция (като функция от параметрите и предикторите), определянето на оптимални точки за наблюдение е сложно. Те могат да се окажат както на границата на областта, така и вътре в нея. Намирането им е предмет на теорията на планирането на експеримента (вж (Kowalik and Osborn 1968)).

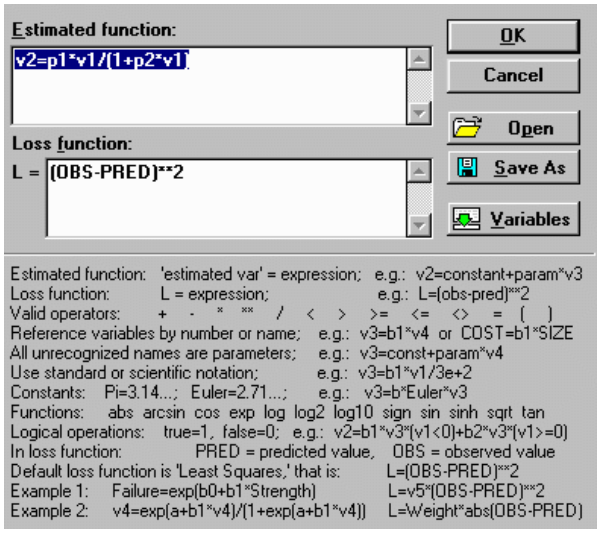
16.2 Нелинейна регресия в програмите

За илюстрация на възможностите и проблемите на метода избрахме следния пример **??** от приложението.

Предлага се следния емпиричен модел за описание на реакцията:

$$y = p(1).x/(1 + p(2)x) + e. \tag{16.4}$$

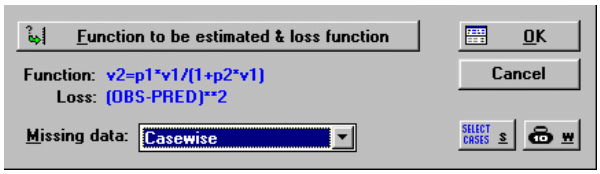
Процедурата предвижда въвеждане на функцията на отклика, задаване на начални стойности на параметрите и намиране на оптималната им оценка. В пакета Statistica това става в няколко прозореца.



Фиг. 16.1: Въвеждане на функция

В първия прозорец (виж фиг.16.1) със средствата на прост език се въвежда функцията на отклика, която ще играе ролята на регресионен модел. Може да се въведе и функция на загубите. По подразбиране това сумата от квадратите на остатъците.

Често обаче е необходимо да се ограничи изменението на параметрите в определена област. Тогава е полезно към минимизирания функционал да се добавят и "глоби" за напускане от параметрите на областта.

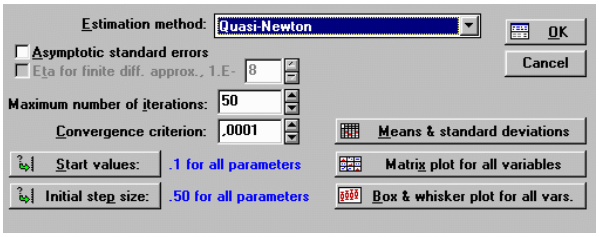


Фиг. 16.2: Въвеждане на тегла

В втория прозорец преди стартиране на методите за оптимизация е предвидена и възможността да се въведе тегло на отделното наблюдение. Теглата трябва предварително да са приготвени в отделна колонка на матрицата данни.

Тук, както и в другите регресионни програми на пакета е предвидена и обработка на "загубените " стойности.

Изборът

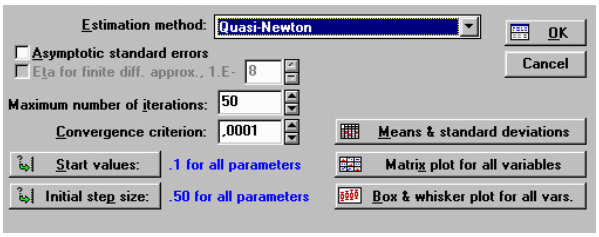


Фиг. 16.3: В третия прозорец (виж фиг.16.4) се избира метод и се задават начални стойности на параметрите. Ако производните не се пресмятат от програмата от предната стъпка - ВЪВЕЖДАНЕ НА ФУНКЦИЯ, необходимо е да се определи и нарастване за оценката им.

на началните приближения е критичен, особено за градиентните методи. Желателно е с помощта на РИСУНКИ и ОСТАТЪЦИ да се провери преди пресмятанията, доколко въведените начални стойности са близки до реалността.

16.2.1 Методи за оптимизация

В тази секция разглеждаме методите реализирани в пакета СТАТЛАБ, описан в (Вындев и Матеев 1988). Същите методи са реализирани в Statistica. Идеята за обединяване на описаните по-долу методи в една програма е заимствана от (Yamaoka, Tanigawara, Nakagawa, and Uno 1981).



След избиране на метод се въвеждат необходимите допълнителни за метода параметри и се извършват пресмятанията. Предвидена е възможност за прекъсване на итерациите и с намеса от клавиатурата.

Фиг. 16.4: Избор на метод

Оптимизираните (най-добрите получени до момента) параметри на регресионното уравнение заместват началните стойности. Произвеждат се и остатъци от модела. Реализирани са следните четири метода:

- метод на Нютон;
- метод на Нютон с подтискане на градиента;
- метод на Маркуард (модификация на Флетчер);
- симплекс метод (Нелдер и Миид).

Първите три метода могат да работят както с изкуствен, така и с естествен градиент. В Statistica градиента е винаги изкуствен. Със средствата на специализирания и език, обаче може и да се преодолее това обстоятелство.

16.2.2 Градиентни методи

При зададени начални стойности на параметрите се изчисляват производните на функцията по тези параметри за всяко наблюдение. Ако във функцията не са определени производните, то програмата строи тяхна оценка с помощта на зададеното нарастване. В този случай е много важно оценяваните параметри да бъдат от един порядък.

С получените производни се строи линейно (по параметрите) приближение на регресионната функция. По метода на най-малките квадрати се оценяват параметрите на приближения модел. Получените нови стойности на параметрите служат за начални при следващата итерация. Процедурата се повтаря докато измененията на параметрите станат незначителни. Тези итерации променят параметрите по посока обратна на градиента на SSr (сумата от квадратите на остатъците), от където и методите носят името си "градиентни".

При добра, гладка в областта на минимума, функция (SSr) и правилно избрани начални стойности, този метод е най-бърз. Понякога, когато функцията е "лоша" или началните стойности са на неподходящо място, се налагат корекции на получения градиент за да стане процедурата устойчива. Тук са реализирани два типа корекции. Те се прилагат, когато при поредната итерация метода на Нютон не намалява достатъчно SSr.

Първата корекция е разполовяване стъпката на изменение на параметрите (по градиента). Нарича се подтискане. Разполовяването се повтаря докато се получи удовлетворителна SSr или се изчерпи разрешеният брой разполовявания. Както и при стандартния метод на Нютон, итерациите продължават докато SSr престане значимо да се изменя.

Втората, предложена в (Fletcher 1971), променя и направлението на изменението. Известна е като модификация на Флетчер на алгоритъма на Маркуард. Тази промяна се задава с константата CF, която се добавя към диагонала на матрицата на нормалните уравнения. Изборът на тази константа не е особено критичен, тъй като тя се коригира автоматично от процедурата.

Освен остатъците, градиентните методи добавят към комплекта (от данни и текущи резултати) ковариационната матрица на оценените параметри.

16.2.3 Симплекс метод

Този метод, предложен от Нелдер и Миид ((Nelder and Meed 1965)), не се нуждае от градиент. Той е описан подробно в (Kowalik and Osborn 1968) и е принципно различен.

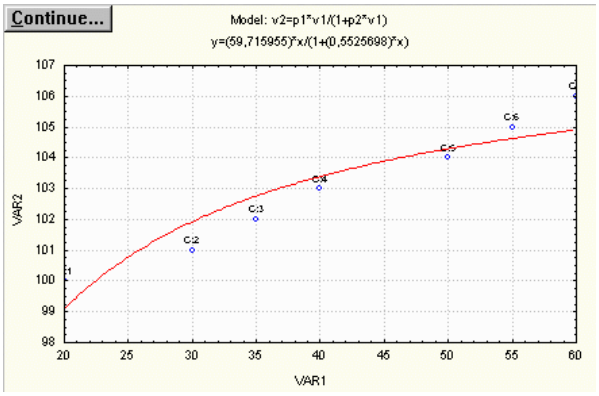
Тук се следят стойностите на SSr във върховете на един симплекс в пространството на параметрите. Симплекс се нарича най-простия възможен многостен в многомерното пространство - в равнината симплексът е просто триъгълник.

Методът се състои в преместване на този симплекс в пространството на параметрите, като може да променя и големината му. Преместването става винаги със заместване на само един връх - този с максимална стойност на целевата функция - той се отразява относно противоположната страна на симплекса или се мести вътре в него.

При два параметъра, триъгълникът би се местил в равнината. Промяната на симплекса се определя от стойностите на SSr във върховете му, параметрите на метода са АЛФА, БЕТА и ГАМА (някои програми подсказват препоръчителни стойности за тях).

Симплекс методът не пресмята ковариационна матрица на оценките. Препоръчва се след неговото използуване да се проведе и поне една итерация с метода на Нютон, която произвежда тази матрица. При симплекс метода е особено критичен изборът на еднакви скали за параметрите.

16.2.4 Графика



Предизвиква изпълнението на специализирани за целите на регресионния анализ графична програми. Рисуват се прогнозирани и наблюдавани стойности на отклика, като за хоризонтална ос служи произволна променлива от оригиналните данни. За нашия пример се получава естествената графика.

Фиг. 16.5: Резултати

От представената рисунка се вижда, че предположения модел не може да бъде адекватен - наблюдава се закономерно изместване н остатъците.

16.3 Теория

В книгите ((Химмельблау 1973), (Химмельблау 1975)) са описани сравнително подробно методите за оптимизация и нелинейната регресия. Тук са дадени само основните формули. Предполага се следния модел:

$$y = f(a) + e, \quad y, f, e \in R^N, \quad a \in R^k, \tag{16.5}$$

където y е зависимата променлива, e - грешката, а f - функция на отклика при фиксирана стойност на неизвестния параметър a в различните наблюдения. Тук

$$\mathbf{E}\, e = 0 \quad \text{и } COV(e) = \sigma^2.W^{-1}, \tag{16.6}$$

където W е диагонална матрица от тегла и σ^2 е неизвестната дисперсия на грешката e.

Разбира се, зависимостта на отклика от предикторите, която тук не е указана явно, се отразява във вектора $f(a)$. Той може да има вида

$$f_1(a) = f(x_1, a), f_2(a) = f(x_2, a), ..., \tag{16.7}$$

но това не е задължително.

В предположение, че грешката е нормална, максимално - правдоподобната оценка може да бъде получена по метода на най-малките квадрати като се минимизира функционала:

$$SSr(a) = (y - f(a))'W(y - f(a)) = e(a)'We(a). \tag{16.8}$$

Прилага се следната линеаризация на модела f:

$$f(a) = f(a_0) + grad\, f(a_0).(a - a_0). \tag{16.9}$$

Тъй като в много случаи аналитичната форма на частните производни на f по a, т.е. $grad\, f(a)$, е твърде сложна, изчислително те могат да се заместват с оценки, получени по метода на крайните разлики. Ако A е новата $N \times K$ матрица от градиенти на f във всяко наблюдение, взети при фиксирана стойност $a = a_0$, условието за минимум на (16.8) може да бъде записано леко:

$$(A'WA)(a - a_0) = A'We(a_0). \tag{16.10}$$

Тук с $e(a_0)$ е означен векторът на остатъците на модела при зададената стойност на параметрите a_0,т.е. $e(a_0) = y - f(a_0)$. Тогава решението на уравнение (16.10) относно a дава новата стойност на параметрите. Това собствено е и итеративния метод на Нютон. Всъщност на всяка стъпка се провежда една линейна регресия и се оценяват коефициентите на един нов линеен модел. Условие за спиране на итерациите е векторът $(a - a_0)$ (нарастването) да има стойност близка до нула. Това всъщност означава, че производната на функционала $SSr(a)$ става нула в точката a_0.

Съществуват ред модификации на този метод. Методът на Маркуард променя матрицата $A'WA$ като добавя към диагонала й константа, която се променя по време на итерациите.

За оценка на ковариационната матрица на параметрите се използува линеаризацията (16.9) (по аналогия с линейната регресия):

$$COV(a) = (A'WA)^{-1}\sigma^2, \tag{16.11}$$

а за оценка на σ^2 оценката $SSr/(N - K)$. Разбира се, тъй като (16.9) е само едно приближение на f, то на (16.11) може да се гледа като на една ориентация даже и в случая, когато моделът е адекватен.

За сравняване на модели с различен брой параметри се използува информационния критерий на Акаике [(Akaike 1973)].

Тема 17

Клъстерен анализ и многомерно скалиране

Това е една от най-популярните процедури на анализа на данни. Терминът е използуван за първи път от (Tryon, 1939) и всъщност зад него се разбират много различни класификационни техники. Пълен обзор на най-използуваните алгоритми има (Hartigan 1975). Думата клъстер означава група от близко лежащи обекти. Цел на клъстерния анализ е разкриване на евентуално скрита групировка на обектите. Обекти на клъстерния анализ могат да бъдат както наблюденията, така и променливите. Достатъчно е между съответните обекти да е зададено някакво разстояние или мярка за близост (сходство). Така например за променливите това може да бъде корелационната матрица.

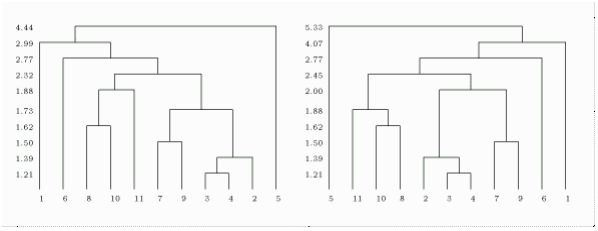
Многомерното скалиране е метод за представяне на наблюдаваните обекти като точки в пространство с понижена размерност, така че разстоянията между тях по възможност да се запазят. Той намира широко приложение в социалните науки, психологията и статистиката. И при тази процедура е достатъчно между съответните обекти да е зададено някакво разстояние. Множество алгоритми са описани в (Torgerson 1958),(Cox and Cox 1994).

17.1 Иерархичен клъстерен анализ

Иерархичния клъстерен анализ съдържа ред процедури, които се базират на последователното обединяване на най-близките клъстери. В началото всеки обект се разглежда като клъстер, състоящ се от един елемент и, следователно, разполагаме с n клъстера.

На всяка стъпка процедурата намира двата най-близки клъстера и ги обединява в един. Т.е. броят на клъстерите намалява с единица.

Всички агломеративни процедури като резултат произвеждат (хоризонтална или вертикална) дендрограма (вж.фиг.17.1):



Фиг. 17.1: Две еквивалентни дендрограми

Ден-

120

дрограмата е граф - дърво, в което всеки възел отразява една стъпка от процеса на обединяване. т.е. той носи и допълнителната информация за величината на разстоянието между двата клъстера в момента на обединение.

Нека означим с I и J два клъстера и множествата индекси на обекти от тях. Видовете стратегии за обединяване на клъстери се базират на следните видове разстояния между клъстерите:

1. Проста връзка - single link (nearest neighbor).

$$d(I,J) = \min_{i\in I, j\in J} d_{i,j}.$$

2. Пълна връзка - complete link (furthest neighbor).

$$d(I,J) = \max_{i\in I, j\in J} d_{i,j}.$$

3. Средно разстояние - Unweighted pair-group average.

$$d(I,J) = \frac{1}{|I|+|J|} \sum_{i\in I, j\in J} d_{i,j}.$$

4. Средно претеглено разстояние - Weighted pair-group average. За това разстояние вече е необходима информация за обектите. Тук се използува същото разстояние, но при решенията се използуват и размерите и центроидите на клъстерите.

Следния пример е взет от учебника на пакета ((Dixon 1981)).

Пример 17.1 *Данните на здравните индикатори за 11 страни са дадени в 17.1. Това са относителен брой на*

1. *(DDNT) - лекари и зъболекари;*
2. *(PHRM) - аптекари;*
3. *(NURS) - сестри;*
4. *(HOSPB) - болнични легла;*

5. *(ANIM) - процент на животинска мазнина в храната;*

6. *(STRCH) - процент на нишесте в храната;*

7. *(LIFEXP) - продължителност на живот.*

	DDNT	PHRM	NURS	HOSPB	ANIM	STRCH	LFEXP
1 Algeria	129	023	350	3392	21	57	35
2 Iran	329	107	290	1113	24	60	51
3 Iraq	241	081	235	1898	28	57	54
4 Jordan	284	096	241	1712	25	49	52
5 Lebanon	933	191	564	4071	35	50	60
6 Libia	338	041	612	3215	24	55	57
7 Morocco	094	026	233	1516	21	57	53
8 Syria	254	070	140	1163	13	69	52
9 Tunizia	114	339	248	2967	21	57	53
10 Tyrkey	412	057	306	1738	16	71	55
11 UAR	483	131	454	2225	15	73	54

Таблица 1.Здравни индикатори

За разстояние между отделните случаи да вземем Евклидовото разстояние между стандартизираните данни. Таблица 2 съдържа горната триъгълна част на така получените разстояния:

A dissimilarity matrix for 11 objects										
	2	3	4	5	6	7	8	9	10	11
1	3.97	3.84	3.82	6.83	4.07	3.52	4.40	2.99	4.30	4.75
2		1.39	1.57	5.33	3.49	2.07	2.45	2.59	2.32	2.83
3			1.21	5.10	3.11	1.73	3.01	1.85	2.78	3.54
4				5.08	3.31	2.00	3.30	2.19	3.29	3.88
5					4.44	6.48	7.10	5.82	6.09	5.33
6						3.36	4.65	2.77	3.51	3.59
7							2.37	1.50	2.50	3.82
8								2.91	2.50	2.90
9									2.68	3.62
10										1.88

Таблица 2 Разстояния

Като приложим клъстерен анализ (например с пакета Statistica), стратегия single link ще получим лявата дендрограма на фиг.**??**.

При внимателно сравнение на двете дендрограми се вижда, че те представят една и съща клъстерна процедура. В какво е разликата? Например, обектите 8 и 2 са по-близки от 11 и 7. Също така обектите 3 и 2 са по-близки от 2 и 4. С други думи сред множеството дендрограми (представящи една и съща агломерация) съществува една,

за която обектите са наредени така върху оста x, че сумата от разстоянията между два съседни обекта да е минимална. Съществува такава пермутация $\pi : i \to (i)$, че

$$\sum_{i=2}^{n} d_{(i-1),(i)} \to min$$

.

Така поставена задачата много пролича на класическата задача за търговския пътник - намиране на най - късото разстояние за обхождане на няколко града. Тук обаче има и съществени разлики. Първо отсъствува един преход - този към началото: $d_{(n),1}$. Второ минимума се взима не по всичките $n!$ пермутации, а само по тези 2^{n-2}, които запазват типа на дендрограмата - редът на обединяване на обектите. Във всеки случай тази задача може да се разглежда като едно решение на задача за скалиране, както ще видим по-нататък.

17.1.1 Променливи и наблюдения

Това е метод за едновременно разглеждане както на променливите, така и на наблюденията. Той може да се използува само ако притежаваме пълната матрица X, а не само матриците на приликите. По идеология много прилича на Biplot или дуално скалиране.

За по-подробна информация вж. (Hartigan 1975)

17.2 Клъстерен анализ на средни

Нека сега си поставим задачата да разделим множеството от всички обекти $1, 2, \ldots, n$ на фиксиран брой k непресичащи се клъстери. Да означим множеството от клъстери с K. Нека $I, J \in K$ са два клъстера и означим по същия начин и множествата индекси на обекти от тях. Тогава

$$d(I, J) = \sum_{i \in I, j \in J} d_{i,j},$$

е естествено да се тълкува като разстоянието между два клъстера, а $d(I, I)$ - като разстоянието вътре в клъстера I. За всеки обект i, да означим с $I(i)$ клъстера, който го съдържа.

17.2.1 Метод на пълните разстояния

Нека си поставим задачата да минимизираме сумата от всички вътрешно - клъстерни разстояния A или, което е същото, да максимизираме сумата от всички между клъстерни разстояния B. Наистина $A + B = \sum_{i<j} d_{i,j}$.

Нека разгледаме един обект $i \in I(i)$. Да проверим кога има смисъл да го прехвърлим в клъстера $J \neq I$, т.е. кога това ще увеличи B (и намали A). Очевидно, трябва $d(i, I(i)) > d(i, J)$. Наистина, при преминаването на i в J сумата B ще нарастне с

$d(i, I(i))$ и ще намалее с $d(i, J)$. Така получаваме необходими и достатъчни условия за оптимална клъстеризация на k клъстера:

$$\forall i : d(i, I(i)) \leq \min_{J \neq I} d(i, J). \tag{17.1}$$

В доклада си (Akoka 1992) авторът предлага естествен алгоритъм за решаване на тази задача и доказва, че той достига абсолютния минимум на A.

17.2.2 Метод на средните разстояния

В повечето програми се използува аналогичен алгоритъм, т.н. K-means clustering, в който обаче се разглеждат не сумарни, а средни разстояния:

$$d(i, J) = \frac{1}{J} \sum_{j \in J} d(i, j), \quad \text{или} \quad d(i, J) = ||x_i - \overline{x}_J||. \tag{17.2}$$

На този метод може да се гледа като метод за максимизиране на F-статистиката на критерия за проверка на хипотезата за еднаквост на k средни стойности.

17.3 Многомерно скалиране MDS

Многомерното скалиране е метод за представяне на наблюдаваните обекти като точки в пространство с понижена размерност, така че разстоянията между тях по възможност да се запазят. Той намира широко приложение в социалните науки, психологията и статистиката.

Да предположим,че имаме n обекта в множеството O със различия $\{\delta_{rs}\}$ между обектите r и s, определени на $O \times O$. Целта на метричното MDS е да се намери множество от точки в Евклидовото пространство E представящи обектите така, че разстоянията между точките $\{d_{rs}\}$ да са $d_{rs} \approx f(\delta_{rs})$. Тук f е непрекъсната монотонна функция, възможно зависеща от параметри.

По (Cox and Cox 1994) нека ϕ е произволно преобразование от O в E. Така да означим $\phi(r) = x_r$ $(r \in O, x_r \in E)$, $\tilde{X} = \{x_r : r \in O\}$ образа и с d_{rs} разстоянието между x_r и x_s.

(Schoenberg 1935) описва метод за реконструиране на евклидовите координати x_r в p-мерно Евклидово пространство E от зададените разстояния $d_{rs} = (x_r - x_s)^T (x_r - x_s)$.

Тъй като решението е неопределено, можем до сложим средната стойност в началото на координатната система. $\sum_{r=1}^{n} x_{ri} = 0, i = 1, \cdots, p$. Така получаваме:

$$d_{rs}^2 = x_r^T x_r + x_s^T x_s - 2 x_r^T x_s$$

$$\frac{1}{n} \sum_{r=1}^{n} d_{rs}^2 = \frac{1}{n} \sum_{r=1}^{n} x_r^T x_r + x_s^T x_s$$

$$\frac{1}{n} \sum_{s=1}^{n} d_{rs}^2 = x_r^T x_r + \frac{1}{n} \sum_{s=1}^{n} x_s^T x_s$$

$$\frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = \frac{2}{n}\sum_{r=1}^{n} x_r^T x_r.$$

Така получаваме:

$$x_r^T x_r = -\frac{1}{2}\left(d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right). \tag{17.3}$$

Да означим $A = \left\{-\frac{1}{2}d_{rs}^2\right\}$ и $H = I_n - \frac{1}{2}ee^T$, където $e = (1, \cdots, 1)^T$. Така, от (17.3) получаваме $XX^T = HAH$ и $rank\left(XX^T\right) = rank\left(X\right) = p$. Ако XX^T положително полу-определена с ранг p, то матрицата X може да бъде намерена от разложението $XX^T = V\Lambda V^T$, където $\Lambda = diag\left(\lambda_1, \cdots, \lambda_n\right)$ е диагоналната матрица от собствените стойности $\lambda_i; i = 1, \cdots, n$ и V са собствените вектори.

Затова проекциите на точките $x_r, r = 1, \cdots, n$ върху осите се намират лесно $X = V_1\Lambda_1^{\frac{1}{2}}$. Тук естествено се използуват само ненулевите собствени стойности.

Основната постановка, сега наричана класическо скалиране е дадена в (Young and Hauseholder 1938) и по-късно и независимо преоткрита от (Gower 1966). Нека пак означим с Δ матрицата от разстоянията между обектите. Тогава тази класическа постановка води до следната минимизационна задача:

$$\min_X tr(J(\Delta^{(q)} - D^{(q)}(X))J)^2 \tag{17.4}$$

Тук $\Delta^{(q)} = (\delta_{i,j}^q)_{1\le i,j\le n}$ е матрицата от q-тите степени на входните разстояния, Тук $D^{(q)} = (d_{i,j}^q)_{1\le i,j\le n}$ е матрицата от q-тите степени на Евклидовите разстояния между представителите (редовете на матрицата $X \in R^{n\times k}$. С $J = I_n - \frac{1}{n}1_{n\times n}$ сме означили центриращата матрица. Аналитично решение на тази задача е дадено в (Mathar 1985).

Представянето (17.4) страда от недостатъка че апроксимира по-добре корелациите от колкото разстоянията между обектите. Затова по-късно се разглеждат нови два метода.

$$\min_X tr(\Delta^{(2)} - D^{(2)}(X))^2, \tag{17.5}$$

$$\min_X \sum_{i,j} w_{i,j}(\delta_{i,j} - d_{i,j}(X))^2. \tag{17.6}$$

(17.5) се нарича squared distance scaling, а (17.6) - least square scaling. С $w_{i,j}$ сме означили неотрицателни тегла. При $w_{i,j} = 0$ съответните разстояния може да отсъствуват, което прави този метод най-удобен.

Общ недостатък на всички изброени методи е, че те не гарантират запазването на съотношенията на наредба между сходствата. От това че $\delta_{i,j} \ge \delta_{k,l}$ не следва, че $d_{i,j} \ge d_{k,l}$ за получената конфигурация.

Затова на практика по-често се използуват алгоритми, които минимизират функционалите (17.4 - 17.6) при това допълнително ограничение. Особено място тук заема алгоритъма на (Kruskal 1964), реализиран в най-популярните програми. Той минимизира функционала (17.6) като постепенно наглася една случайно генерирана начална конфигурация.

Тема 18

Методи за оценка на плътности

Една от основните задачи на статистиците е да опишат разпределението на наличните данни. След като по някакъв начин е избрано разпределението на извадката, може да се премине към анализ с цел правене на статистически изводи. При определянето на разпределението на извадката най-общо може да се подходи по два начина. Може да се избере *параметричен подход*, като се приеме, че въпросното разпределение принадлежи на семейство от разпределения, които са определени с точност до един или няколко неизвестни параметъра. Тогава въпросът за определяне на разпределението се свежда до оценяването на тези неизвестни параметри по наличните данни, като се използват стандартни методи за оценка. Алтернативният вариант е да се избере подход, при който няма предварително постулирана форма на разпределението и съответно то ще се определя изцяло на основата на данните от извадката. Именно този подход най-общо може да бъде наречен *непараметричен*. Пример за използването на непараметричен подход е построяването на емпирична функция на разпределение за дадена извадка и правене на изводи на базата на тази функция на разпределение.

Един клас методи, спадащи към групата на непараметричните методи, са *методите за оценка на плътности.*

Ще разгледаме:

- апроксимации на Еджуърт,
- криви на Пирсън,
- изглаждане на хистограми,
- ядрени оценки на плътности.

18.1 Апроксимации на Еджуърт

Разлагането на Еджуърт може да се разглежда като асимптотично приближение на плътност или на функция на разпределение. Основната идея се състои в значително усилване на централната гранична теорема за независими и еднакво разпределени случайни величини, предполагайки че съществуват техните моменти от висок порядък.

Теорема 18.1 *Нека разгледаме центрираната и нормирана сума:*

$$S_n = \frac{\sum_{k=1}^{n} \xi_k - nE\xi}{\sqrt{nD\xi}}$$

на независими и еднакво разпределени случайни величини $\xi_1,\ \xi_2,\ \ldots,\ \xi_n$, *такива че* $E\xi^3 < \infty$ *имащи плътност. Тогава равномерно относно* x *имаме*

$$f_{S_n}(x) - \phi(x)\left[1 + \frac{\mu_3}{6\sigma^3\sqrt{n}}(x^3 - 3x)\right] \underset{n\to\infty}{\longrightarrow} = o\left(\frac{1}{\sqrt{n}}\right)$$

.

В горния запис $\mu_3 = E\xi^3$, $\sigma = \sqrt{D\xi}$ и $\phi(x)$ е плътността на стандартно нормално разпределена случайна величина.

Доказателство: Ясно е, че за да построим приближение на плътността $f_{S_n}(x)$ според горния запис, трябва да познаваме единствено първите три начални момента. Съществува обобщение на горния резултат ако допуснем, че $E\xi^r < \infty$, но той е извън целите на този курс и няма да го формулираме. Ще продължим като илюстрираме основната идея на горната теорема с помощта на характеристични функции без да я доказваме и след това ще разгледаме един пример при малко по-общи предположения – използвайки апроксимацията на Еджуърт при допускане, че $E\xi^4 < \infty$.

Характеристичната функция на сумата $\phi_{S_n}(t)$ може да се запише чрез характеристичната функция на ξ по следния начин:

$$\phi_{S_n}(t) = e^{\frac{-it\sqrt{n}E\xi}{\sqrt{D\xi}}}\phi_\xi\left(\frac{t}{\sqrt{nD\xi)}}\right)^n.$$

Като използваме развитието в ред на Тейлър:

$$\ln(\phi_\xi(t)) = \sum_{j=1}^{\infty} \frac{\kappa_j (it)^j}{j!}$$

за характеристичната функция на сумата $\phi_{S_n}(t)$ получаваме:

$$\phi_{S_n}(t) = e^{\frac{-it\sqrt{n}E\xi}{\sqrt{D\xi}}}\phi_\xi\left(\frac{t}{\sqrt{nD\xi)}}\right)^n = exp\left[-\frac{t^2}{2} + n\sum_{j=3}^{\infty}\frac{\kappa_j}{\sigma^j n^{\frac{j}{2}} j!}(it)^j\right]$$

Използвайки горното развитие в ред на Тейлър, и връзката между характеристична функция и начални моменти, може да се покаже как $\kappa_1, \kappa_2, \ldots$ зависят от моментите на случайната величина, например $\kappa_1 = \mu_1$, $\kappa_2 = \sigma^2$ и т.н. Следователно можем да апроксимираме плътността $\phi_{S_n}(t)$ използвайки първите няколко члена на сумата в степента, което е еквивалентно на това да използваме само първите няколко момента на разпределението. По този начин, разписвайки само първите два члена на сумата, получаваме:

$$\phi_{S_n}(t) = \exp\left(-\frac{t^2}{2}\right)\exp\left(\frac{\kappa_3(it)^3}{6\sigma^3 n^{\frac{1}{2}}} + \frac{\kappa_4(it)^4}{24\sigma^4 n} + o(n^{-\frac{3}{2}})\right).$$

Като използваме обратното преобразуване на Фурие, за да получим плътност от апроксимираната характеристична функция, получаваме че плътността $f_{S_n}(x)$ на сумата от случайни величини може да се апроксимира така:

$$f_{S_n}(x) = n(x)\left[1 + \frac{\kappa_3 H_3}{6\sigma^3 n^{\frac{1}{2}}} + \frac{\kappa_4 H_4}{24\sigma^4 n} + \frac{\kappa_3^2 H_6}{72\sigma^6 n} + o(n^{-\frac{3}{2}})\right],$$

където H_n са полиноми на Ермит и се дефинират:

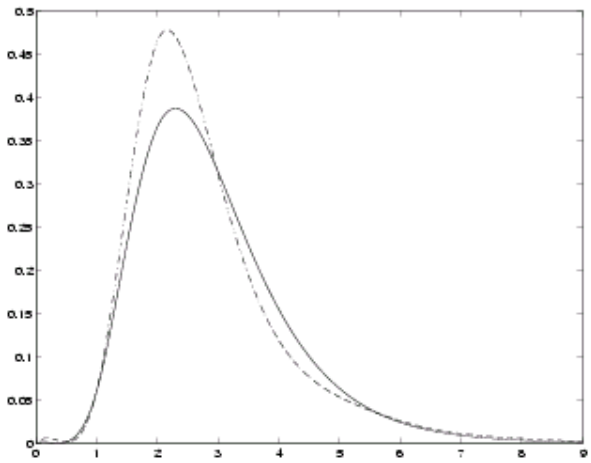
$$H_n(x) = (-1)^n e^{\frac{1}{2}x^2} \frac{d^n}{dx^n} e^{-\frac{1}{2}x^2}. \square$$

Нека разгледаме следния пример, за да илюстрираме ползата от горната апроксимация.

Пример 18.1 *Нека $\xi_1, \xi_2, ..., \xi_n, ...$ са независими експоненциално разпределени случайни величини с параметри $\lambda_i = i$. Тогава сумата $S_n = \sum_{j=1}^{n} \xi_j$ има плътност*

$$S_n(x) = n \sum_{k=1}^{n} (-1)^{k-1} \binom{n-1}{k-1} e^{-kx}$$

В случаи като този не е разумно да се използва направо централната гранична теорема и горната плътност да се апроксимира с нормална, защото дисперсиите на случайните величини се различават много. По-добри резултати дава апроксимацията на Еджуърт.



Фиг. 18.1: Еджуърт

При $n = 10$ фиг. 18.1 илюстрира апроксимацията (пунктир) заедно с теоретичната плътност (плътна линия). От формата на плътността на $S_{10}(x)$ ясно се вижда, че нормалното разпределение не може добре да я апроксимира. При изчисленията сме използвали представянето за $f_{S_n}(x)$ както е записано по-горе.

Ако използваме повече членове при апроксимирането на сумата в представянето за $\phi_{S_n}(t)$ естествено ще получим по-точно приближение, но изчисленията значително се усложняват.

Апроксимации на Еджуърт могат да се направят не само на плътности, но и на функции на разпределение. Като използваме свойствата на полиномите на Ермит, интегрирайки представянето за $f_{S_n}(x)$, получаваме:

$$F_{S_n}(x) = N(x) - n(x)\left[1 + \frac{\kappa_3 H_2}{6\sigma^3 n^{\frac{1}{2}}} + \frac{\kappa_4 H_3}{24\sigma^4 n} + \frac{\kappa_3^2 H_5}{72\sigma^6 n} + o(n^{-\frac{3}{2}})\right],$$

където $N(x)$ е функция на разпределение на стандартно нормално разпределение. По начинът, по който получихме израза за $F_{S_n}(x)$, читателят може да се заблуди, че за неговото съществуване е нужно случайната величина S_n да притежават плътност. Това далеч не е вярно – горният запис е в сила дори когато разпределението F няма плътност.

18.2 Криви на Пирсън

За изясняване на кривите на Пирсън ще напомним следният пример, който по естествен начин води до идеята за тези криви.

Пример 18.2 *(Хипергеометрично разпределение)*

Разглеждаме урна, в която има R червени топки и B черни топки, съответно общият брой топки в урната е $R + B$. По случаен начин от тази урна се теглят n топки по схема без връщане. Разпределението на броя червени топки в извадката е хипергеометрично и се дава от формулата

$$P(r) = \frac{\binom{R}{r}\binom{B}{n-r}}{\binom{R+B}{n}},$$

където $P(r)$ означава вероятността на събитието $\{$*падат се* r *червени топки*$\}$. Тук допълнително се налага изискването $n < R + B$ и $0 < r < min\ (R, n)$.

Ако случайната величина ξ е хипергеометрично разпределена, то е в сила

$$\mathbb{E}\xi = \frac{nR}{R+B}, \qquad \mathbb{D}\xi = \frac{nRB}{(R+B)^2}\left(1 - \frac{n-1}{R+B-1}\right).$$

При различни стойности на параметрите хипергеометричното разпределение се доближава до някои от често срещаните в практиката разпределения. Например при голям обем на общата съвкупност и неголяма извадка, хипергеометричното разпределение е близко до биномно с параметри n и $p = R/(R+B)$. Аналогично с подходящ избор на параметрите може да получим приближения на поасоново и нормално разпределение.

Ако за удобство означим хипергеометричната вероятност $P(r)$ с p_r, може да се покаже, че разликата $\Delta p_r = p_{r+1} - p_r$ удовлетворява диференчното уравнение

$$\frac{1}{p_r}\ \Delta p_r = -\left\{\frac{-(nR - B + n - 1) + (R + B + 2)r}{(B - n + 1) + (B - n + 2)r + r^2}\right\}.$$

Пирсън е използвал факта, че хипергеометричната вероятност има такова представяне, както и това, че за различни стойности на параметрите R, B и n се получават споменатите удобни приближения на други разпределения, за да конструира непрекъснат аналог на горното диференчно уравнение. Той разглежда система от вероятностни плътности $f(x)$, които удовлетворяват диференциалното уравнение

$$\frac{1}{f(x)}\ .\frac{df}{dx} = -\frac{x+b}{c_0 + c_1 x + c_2 x^2}.$$

При така избраната форма на диференциалното уравнение плътност $f(x)$, която го удовлетворява, ще има следните свойства:

1. Случайната величина с такава плътност приема стойности в определени граници. Извън тези граници плътността е нула.

2. В интервала, в който се изменят стойностите на случайната величина, плътността започва да нараства от нулата, достига своя (единствен) максимум и след това намалява до нула.

Горното диференциално уравнение може да бъде интерпретирано по следния начин: *скоростта на изменение на функцията $f(x)$, описана с първата ѝ производна, трябва да бъде равна на нула в три точки: в началото и края на интервала на изменение, и в точката на максимум.* Ясно е, че ако $f(x)$ е плътност с посочените свойства, то $\frac{df(x)}{dx} = 0$ за $f(x) = 0$ (т.е. извън областта, в която приема стойности случайната величина) и за $x = -b$ (в точката на максимум).

Константите, които влизат в диференциалното уравнение, могат да бъдат напълно характеризирани с помощта на първите четири момента на съответното разпределение. Ако с

$$r_j = \frac{\mu_j}{\sigma^j}$$

означаваме j-тия основен момент на разпределението, като μ_j е j-тия централен момент, а σ е стандартното отклонение, то имаме

$$c_0 = -\sigma^2 \frac{s+1}{s-2}, \quad c_1 = -b = -\frac{\sigma r_3}{2} \cdot \frac{s+2}{s-2}, \quad c_2 = \frac{1}{s-2},$$

където

$$s = \frac{6(r_4 - r_3^2 - 1)}{3r_3^2 - 2r_4 + 6}.$$

Това означава, че работим с разпределения, които се определят напълно от първите си четири момента. При това положение можем да използваме метода на моментите и да заместим теоретичните моменти с емпиричните, за да получим оценка за плътността на разпределението, от което е генерирана извадката.

Общият интеграл на диференциалното уравнение, което стои в основата на кривите на Пирсън, има вида

$$y = y_0 e^{v(x)},$$

където

$$v(x) = \int \frac{x+b}{c_0 + c_1 x + c_2 x^2}\, dx.$$

Стойностите на този интеграл зависят от стойностите на израза $c_0 + c_1 x + c_2 x^2$ в знаменателя. При тази постановка задачата става еквивалентна на разглеждане на уравнението

$$c_0 + c_1 x + c_2 x^2 = 0.$$

Корените R_1 и R_2 на това уравнение зависят от дискриминантата му

$$D = c_1^2 - 4c_0c_2 = c_1^2\left(1 - \frac{4c_0c_2}{c_1^2}\right).$$

Ако означим

$$\kappa = \frac{c_1^2}{4c_0c_2},$$

то дискриминантата може да се запише като

$$D = c_1^2\left(1 - \frac{1}{\kappa}\right).$$

Нека още

$$t = \sqrt{r_3^2(s+2)^2 + 16(s+1)} = 4\sqrt{(s+1)(1-\kappa)}.$$

Тогава за корените на квадратното уравнение получаваме

$$R_1 = \frac{\sigma}{4}(1-t), \qquad R_2 = \frac{\sigma}{4}(1+t).$$

Сега да забележим, че корените R_1 и R_2, а следователно и решението на диференциалното уравнение, се определят от величината κ. Това означава, че κ може да бъде използвана като критерий за различаване на решенията или, еквивалентно, за определяне на типа на разпределението, който получаваме с тази процедура. Критерият κ може да бъде записан по следния начин:

$$\kappa = \frac{r_3^2(r_4+3)^2}{4(4r_4 - 3r_3^2)(2r_4 - 3r_3^2 - 6)} = -\frac{r_3^2(s+2)^2}{16(s+1)}.$$

За да изясним значението на критерия κ ще разложим тричлена $c_0 + c_1x + c_2x^2$:

$$c_0 + c_1x + c_2x^2 = c_2\left[\left(x + \frac{c_1}{2c_2}\right)^2 - \left(\frac{c_1^2}{4c_2^2} - \frac{c_0}{c_2}\right)\right]$$

$$= c_2\left[\left(x + \frac{c_1}{2c_2}\right)^2 - \frac{4c_0^2}{c_1^2}\frac{c_1^2}{4c_0c_2}\left(\frac{c_1^2}{4c_0c_2} - 1\right)\right] =$$

$$= c_2\left[\left(x + \frac{c_1}{2c_2}\right)^2 - \frac{4c_0^2}{c_1^2}\kappa(\kappa - 1)\right].$$

Тогава квадратният тричлен ще се разлага на два реални множителя ако е изпълнено условието $\kappa(\kappa - 1) > 0$.

Последното е в сила за $\kappa \in (-\infty,\ 0) \cup (1,\ \infty)$. Освен това в този случай при положителни κ корените ще бъдат с еднакъв знак, а при $\kappa < 0$ знаците им ще бъдат различни. С помощта на този интервал на изменение на стойностите на κ можем да класифицираме кривите на Пирсън. Най-общо, има 12 типа криви, като 7 от тях са основни, а останалите се получават като частни случаи от тях. Основните типове са показани на таблица 18.1.

Тип	Определящи параметри	Плътност	Допълнителни параметри	Ограничения за x
I	$\kappa < 0$	$\tilde{n}_0\left(1+\frac{x}{l_1}\right)^{q_1}\left(1-\frac{x}{l_2}\right)^{q_2}$	$q_{1,2} = \frac{1}{2}\{(s-2) \mp s(s+2)\frac{r_3}{t}\}$ $l_1 = \frac{q_1 l}{s-2},\ l_2 = \frac{q_2 l}{s-2},\ l = l_1 + l_2$ $\tilde{n}_0 = \frac{n}{l}\cdot\frac{q_1^{q_1} q_2^{q_2}}{(s-2)^{s-2}}\cdot\frac{\Gamma(s)}{\Gamma(q_1+1)\Gamma(q_2+1)}$	$-l_1 < x < l_2$
II	$\kappa = 0,$ $r_3 = 0,$ $r_4 < 3$	$\tilde{n}_0\left(1-\frac{x^2}{l'^2}\right)^q$	$q = \frac{5r_4-9}{2(3-r_4)}$ $l' = \sigma\sqrt{\frac{2r_4}{3-r_4}}$ $\tilde{n}_0 = \frac{n}{2^{2q+1}l'}\cdot\frac{\Gamma(2q+2)}{\{\Gamma(q+1)\}^2}$	$x < \lvert l' \rvert$
III	$\kappa = \pm\infty$	$\tilde{n}_0(1+\frac{x}{l_1})^p e^{-\frac{px}{l_1}}$	$p = \frac{4}{r_3^2} - 1$ $l_1 = \sigma\left(\frac{2}{r_3} - \frac{r_3}{2}\right)$ $\tilde{n}_0 = \frac{n}{l_1}\cdot\frac{p+1}{e^p\Gamma(p+1)}$	$-l_1 < x$
IV	$0 < \kappa < 1$	$\tilde{n}_0(1+\frac{x^2}{l^2})^{-q} e^{-\nu\ arctg\frac{x}{l}}$	$r = -s,\ q = \frac{r+2}{2},\ \tilde{n}_0 = \frac{n}{l.F(r,\nu)}$ $\nu = -\frac{r(r-2)r_3}{\sqrt{16(r-1)-r_3^2(r-2)^2}}$ $l = \frac{\sigma}{4}\sqrt{16(r-1)-r_3^2(r-2)^2}$	$-\infty < x < \infty$
V	$\kappa = 1$	$\tilde{n}_0 x^{-p} e^{-\frac{\gamma}{x}}$	$p = 4 + \frac{8+4\sqrt{4+r_3^2}}{r_3^2}$ $\gamma = \sigma(p-2)\sqrt{p-3}$ $\tilde{n}_0 = \frac{n\gamma^{p-1}}{\Gamma(p-1)}$	$0 \le x < \infty$
VI	$1 < \kappa < \infty$	$\tilde{n}_0 x^{-q_1}(x-l)^{q_2}$	$q_{1,2} = \frac{1}{2}\{s(s+2)\frac{r_3}{t} \mp (s-2)\}$ $l = \frac{\sigma t}{2}$ $\tilde{n}_0 = \frac{n l^{q_1-q_2-1}\Gamma(q_1)}{\Gamma(q_1-q_2-1)\Gamma(q_2+1)}$	$l < x < \infty$
VII	$\kappa = 0,$ $r_3 = 0$ $r_4 > 3$	$\tilde{n}_0(1+\frac{x^2}{l^2})^{-q}$	$q = \frac{5r_4-9}{2(r_4-3)} > 0$ $l = \sigma\sqrt{\frac{2r_4}{r_4-3}}$ $\tilde{n}_0 = \frac{n\Gamma(q)}{l\Gamma(q-\frac{1}{2})\Gamma(\frac{1}{2})}$	$-\infty < x < \infty$

Таблица 18.1: Основни типове криви на Пирсън

18.3 Хистограми

Хистограмата може да се разглежда като най-простата форма на непараметрична оценка на плътност. Когато класът от разпределения, характеризирани напълно от първите си четири момента, се счита за недостатъчно "богат", можем да използваме чисти непараметрични методи, какъвто е построяването на хистограма.

18.3.1 Изглаждане на хистограми

Обикновено построяването на самата хистограма е само предварителен етап в тази процедура за получаване на оценка на плътности. След като имаме хистограмата, трябва да получим от нея функцията на плътността. Именно това е изглаждането на хистограмата. Практически изглаждане на хистограмата означава, че през стълбовете на хистограмата се прекарва някаква крива, която се приема за приближение на плътността. В най-простия случай това е начупена крива, която минава през средата на всеки от стълбовете на хистограмата (за крайните стълбове се взема отстъп от половин стъпка т.е. изместваме се встрани с половин ширина на стълб от хистограмата).

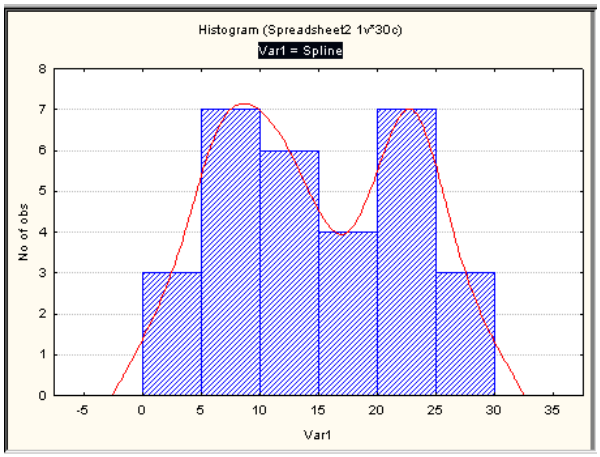
По-удобно разбира се е да работим с гладки плътности. Това се постига като вместо отсечки, свързващи стълбовете, за целта се използват *сплайн-функции* (най-често

кубични сплайни). Сплайн-функциите са полиноми върху отделни интервали, които приемат равни стойности на границата между тези интервали и освен това удовлетворяват допълнителни условия за гладкост на свързването. Интуитивно казано, при прехода между два интервала стойностите трябва да се "слепват"и преходът да е достатъчно гладък, като последното зависи от реда на сплайна.

При построяването на хистограми трябва да се отчитат два основни фактора: избор на интервал, в който се изменят регистрираните стойности на наблюдаваната променлива и брой на подразделенията на този интервал (брой на стълбовете в хистограмата). Изборът на интервал на изменение обикновено е стандартна процедура и се прави така, че да не се изпускат наблюдения [1]. Ако например данните се изменят в интервала от 50 до 100, то очевидно вземането на по-тесен интервал за хистограмата ще доведе до изпускане на наблюдения и съответно до изкривени резултати.

За разлика от избора на интервал на изменение, изборът на броя на подразделенията е по-сложна задача, тъй като той силно влияе върху формата на хистограмата и съответно на изглаждащата крива. Ако изберем да построим хистограмата с твърде малко стълбове, това ще скрие част от характеристиките на извадката – например можем да получим унимодална плътност вместо бимодална. Обратно, ако изкуствено увеличим броя на стълбовете, ще засилим влиянието на случайността и ще подчертаваме тривиални характеристики на извадката с локална природа (например ще запазим локални екстремуми, които би трябвало да се отстранят в процеса на изглаждане). Поради това оптималният избор на брой стълбове изисква повишено внимание, а и опит от страна на статистика, когато се определя експертно. Съществуват и формални процедури за определянето на броя стълбове, на които тук няма да се спираме.

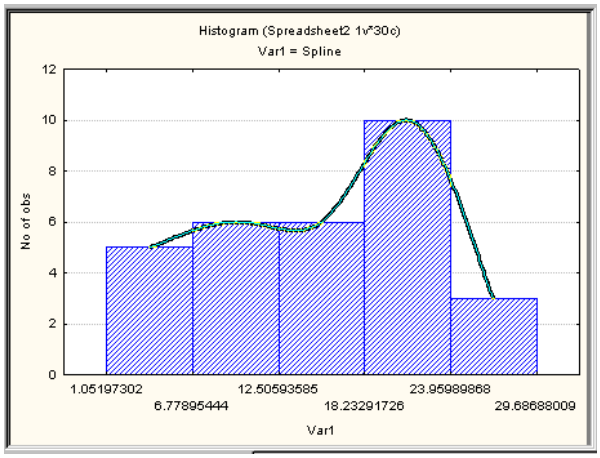
За илюстриране на казаното дотук ще разгледаме пример, в който от една и съща извадка получаваме различни хистограми и изглаждащи криви (при използването на сплайн-функция за изглаждане на хистограмата) в зависимост от избора на броя стълбове. Изкуствено са генерирани 30 случайни числа от равномерно разпределение върху интервала [0, 30]. За тази извадка построяваме и изглаждаме три различни хистограми.



Фиг. 18.2: 10 стълба

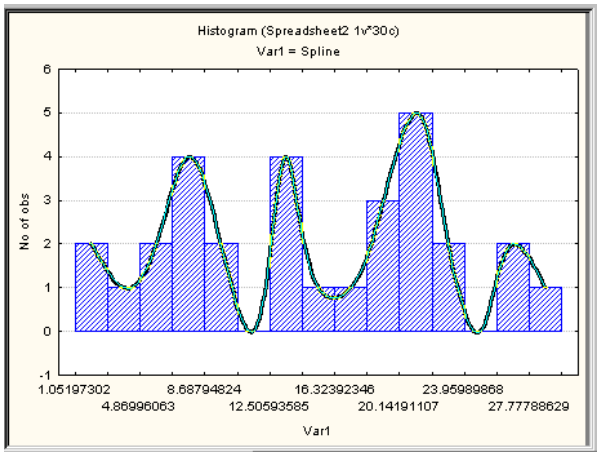
На фигура 18.2 е показана хистограма за извадката при разделяне на интервала на 10 части, съответно с 10 стълба, заедно с изглаждащия сплайн. Тя в някакъв смисъл дава "средно"разделяне, което елиминира локалните особености в данните без да губим много информация.

[1] В някои софтуерни пакети като Statistica това по подразбиране се прави автоматично.



Фиг. 18.3: 5 стълба

Ако построим хистограмата като зададем разделянето да е на 5 стълба, получаваме по-груба структура и различна форма на сплайна. Това разбира се дължи на загубата на информация в следствие на агрегирането в малък брой стълбове. Подобна ситуация е показана на фигура 18.3.



Фиг. 18.4: 15 стълба

Накрая, ако зададем много фино разделяне на интервала, това ще подчертае локалните характеристики на извадката. Последното обаче крие рискове от изпъкване на някои слабо релевантни свойства, които по принцип е трябвало да бъдат филтрирани при изглаждането. Фигура 18.4 дава хистограмата за извадката при разделяне на 15 стълба.

18.3.2 Оптимална хистограма

При калибрирането на хистограми традиционните методи от типа на максималното правдоподобие не работят. Затова се прибягва до непараметрични методи от типа на минималното разстояние до неизвестната функция $f(x)$.

Един популярен критерий е квадратичната грешка:

$$\int (\widehat{f}_h(x) - f(x))^2 dx, \tag{18.1}$$

където широчината на стълбчето на хистограмата се дава от индекса h. Това разстояние се разлага на три члена, които ще разгледаме поотделно:

$$\int \widehat{f}_h(x)^2 dx - 2\int \widehat{f}_h(x) f(x) dx + \int f(x)^2 dx \tag{18.2}$$

1. Първият член се пресмята лесно и точно когато променяме началото и/или широчината.

2. Третия член се определя от незвестната плътност и е неуправляем - за нас той е константа. Следователно може да се изпусне.

3. Следователно вторият член е най - важен

(Rudemo 1982) е предложил да се приложи за него leave-one-out оценка. Той предлага да се конструира нова хистограма $\widehat{f}_{h,-i}$ като едно от наблюденията ($i-$тото) се отстрани и хистограмата се пресмята наново. Това се повтаря за всичките n наблюдения и резултатите се усредняват.

Така получаваме:

$$\int \widehat{f}_h(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{h,-i}(x_i). \tag{18.3}$$

За хистограма с честоти c_k, този израз се пресмята точно:

$$\frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2 h}\sum_k c_k^2. \tag{18.4}$$

от Rice Virtual Labs ((Lane 1998)).

Observe that the bin counts must be recomputed after any change in the bin width h and/or the bin location parameter. In our particular implementation, we have chosen 20 bin shifts and 30 bin widths. The largest is guaranteed to be too large (see the topic of oversmoothing in (**?**). If you look carefully, you will see the bin widths available are not equally spaced. This is important because it is relative changes (rather than absolute changes) in the bin widths that are comparable across the range. The bin widths available, if plotted on a logarithmic scale, would be equally spaced.

18.4 Ядрени оценки

Както видяхме, построяването и изглаждането на хистограми не е лишено от недостатъци, особено предвид силната зависимост на крайния резултат от избора на броя на интервалите. Тези проблеми донякъде се преодоляват с въвеждането на специален клас оценки, наречени *ядрени оценки*. Идеята на тези оценки накратко може да бъде представена по следния начин: За всяко от наличните наблюдения вземаме по една плътност, наречена *ядро*, която поставяме така, че да е центрирана върху съответното наблюдение. После по подходящ начин "разливаме"тези плътности и ги смесваме така, че да се получи една обща плътност, която ще служи за оценка на неизвестната плътност на разпределението, от което са данните.

18.4.1 Обща постановка

Нека $x_1, x_2, \ldots, x_n$ са наблюдения над случайна величина с неизвестна плътност f (както обикновено, можем да считаме, че всяко едно от тях е реализация от една от n на брой независими и еднакво разпределени случайни величини). Тогава стандартната ядрена оценка се определя като

$$\hat{f}(x) = \frac{1}{nh}\ \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

Тук с $K(\cdot)$ сме означили ядрото, а h е константа, [2] която определя гладкостта на оценката и обикновено се изисква да удовлетворява определени условия. Тривиално се вижда, че стандартната ядрена оценка е неотрицателна и се интегрира до единица т.е. че изпълнява изискванията за плътност.

Често се оказва, че използването на един фиксиран параметър за гладкост е недостатъчно и затова се работи с променлив параметър, който е функция на броя на наблюденията n. Тогава горната формула се модифицира по следния начин:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right).$$

Стандартни условия върху h_n са те да са положителни и $h_n \to 0$ и $nh_n \to \infty$. Искаме също и ядрата да изпълняват определени условия, които ще бъдат прецизирани по-долу за разглежданите специални случаи. За последната оценка има и други изрази, които са сходни като идея. При всички случаи статистическата задача се свежда до оценяване на константите и/или ядрото.

Друга алтернатива за построяване на ядрени оценки е да се използват толкова константи, колкото са наблюденията т.е. по една за всяко ядро. По този начин се осигурява по-добро локално изглаждане на плътността и ядрената оценка има вида

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right).$$

Този подход изисква да се оценят n константи h_i, но дава по-голяма гъвкавост при конструирането на оценката. Именно той е в основата на филтрираните ядрени оценки, които ще разгледаме по-долу.

Ядрените оценки се обобщават за метрични пространства с произволна природа по следния начин. Нека Υ е такова пространство. За него ядрената оценка е

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{\rho(x,\ x_i)}{h_n}\right), \qquad x \in \Upsilon.$$

Тук $x_1, x_2, \ldots, x_n$ са независими и еднакво разпределени случайни величини със значения в Υ, $\rho : \Upsilon \times \Upsilon \to [0,\ \infty)$ е метрика в Υ, $K : [0,\ \infty) \to \mathbb{R}^1$ е ядрото, удовлетворяващо условия за регулярност, а $h_n > 0$ е такова, че $h_n \to 0,\ \ nh_n \to \infty$.

След тези общи бележки върху ядрените оценки за илюстрация ще разгледаме два специални случая: *ядрата на Розенблат-Парзен* и *филтрираните ядрени оценки.*

18.4.2 Ядра на Розенблат-Парзен

Ядрена оценка на Розенблат-Парзен се нарича оценката:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right).$$

[2] Английският термин е *bandwidth*

където h_n имат свойствата, описани в предишния параграф. При достатъчно слаби ограничения върху f и K, така получената оценка на плътността е асимптотично неизместена, състоятелна и асимптотично нормална. Освен това:

$$Ef_n(x) = f(x) + O({h_n}^2)$$

$$Df_n(x) = \frac{f(x)\int K^2(u)du}{h_n n} + o((h_n n)^{-1})$$

Ако допуснем, че K е функция с ограничена вариация, f е равномерно непрекъсната и редицата h_n е подбрана така, че $n^{-1}{h_n}^{-2}\ln(n) \to 0$, за сходимостта на ядрената оценка към истинската плътност $f(x)$ може да се получи условието:

$$P\Big(\lim_{n\to\infty}\sup_x |f_n(x) - f(x)| = 0\Big) = 1$$

Обикновено за практически цели за h_n се използва

$$h_n = cn^{-\frac{1}{5}}$$

където c е подходящо избрана константа. Що се отнася до ядрените функции $K(x)$, най-често използвани са следните функции:

1. Правоъгълна

$$K(x) = \begin{cases} \frac{1}{2}, & |x| \le 1 \\ 0, & |x| > 1 \end{cases}$$

2. Триъгълна

$$K(x) = \begin{cases} \frac{1}{\sqrt{6}} - \frac{|x|}{6}, & |x| \le \sqrt{6} \\ 0, & |x| > \sqrt{6} \end{cases}$$

3. Гаусова

$$K(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

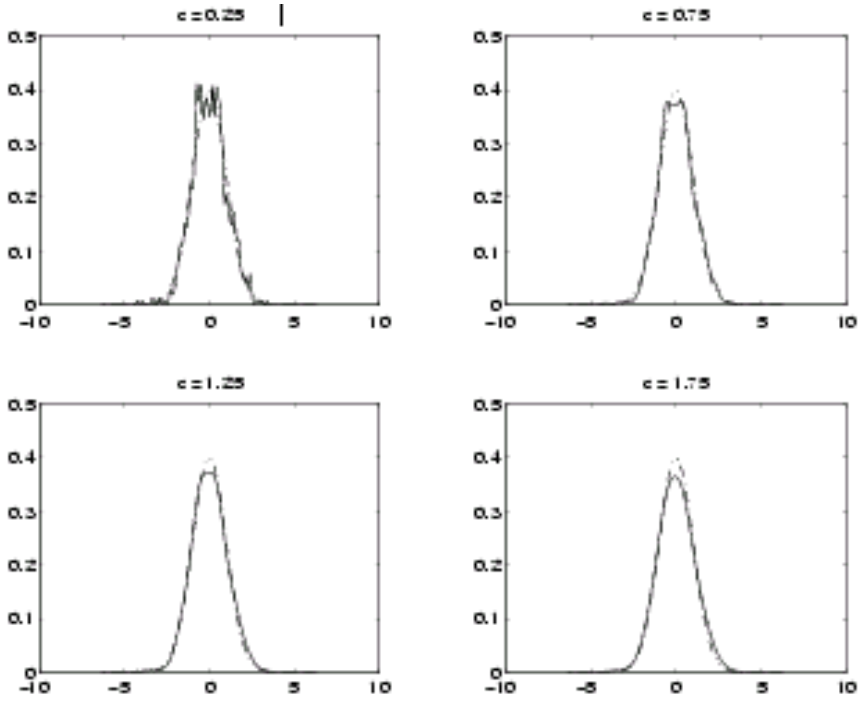
4. "Оптимална" (Ядро на Епанечников)

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), & |x| \le \sqrt{5} \\ 0, & |x| > \sqrt{5} \end{cases}$$

5. Двойно претеглена

$$K(x) = \begin{cases} \frac{15}{16\sqrt{7}}\left(1 - \frac{x^2}{7}\right)^2, & |x| \le \sqrt{7} \\ 0, & |x| > \sqrt{7} \end{cases}$$

На фиг. 18.5 сме илюстрирали ядрена оценка на плътността на извадка от стандартно нормално разпределение (плътна линия) и теоретичната плътност - $N(0,1)$. Броят на наблюденията n е 1000 и използваната ядрена функция е Гаусова. Използвали сме различни константи c - изглаждащото действие е очевидно.



Фиг. 18.5: Ядрена оценка

18.4.3 Филтрирани ядрени оценки

Филтрираните ядрени оценки се основават на идея, подобна на тази в Бейсовия подход към статистиката. По-горе отбелязахме, че е възможно да се използват различни изглаждащи константи h_i, по една за всяко наблюдение. Това осигурява по-добро локално изглаждане около отделните наблюдения, но процедурата си остава изцяло зависима от данните в извадката. Ако обаче разполагаме с априорна информация за локалната гладкост на плътността, желателно е да имаме метод, който може да отчете и нея при конструирането на оценка за плътността. Освен това стандартните ядрени оценки обикновено не дават много добри резултати в опашките на плътността, където имаме малко наблюдения, съответно оценките на изглаждащите параметри нямат много добри статистически качества. Тези проблеми до известна степен се разрешават с използването на филтрирани ядрени оценки.

При филтрираните ядрени оценки отново имаме константите h_i, всяка от които е свързана с отделна област от носителя на плътността. Към всяка от тези константи асоцииране по една функция, която се нарича *филтрираща функция*. Тези функции играят ролята на тегла, определящи в каква степен съответната константа ще се използва за всяко от наблюденията. Тогава филтрираната ядрена оценка може да бъде построена като комбинация от ядрените оценки с използването на константите h_i, но при филтриране на данните с филтриращите функции.

По-точно, нека разгледаме набор от функции $\{\rho_j\}_{j=1}^m$, където $0 \leq \rho_j(x) \leq 1, \ \forall x$ и $\sum_{j=1}^m \rho_j(x) = 1, \ \forall x$. Тези функции са филтриращите функции и могат да бъдат интерпретирани като апостериорни вероятности, които отчитат априорната информация за локалната гладкост на плътността. Тогава, ако константите h_j изпълняват

посочените горе условия, филтрираната ядрена оценка за филтъра $\{\rho_j\}_{j=1}^m$ се дефинира като

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\rho_j(x_i)}{h_j}K\left(\frac{x-x_i}{h_j}\right). \tag{18.5}$$

Идеята на филтрираната ядрена оценка е следната: при зададена смес от краен брой елементи

$$g(x) = \sum_{j=1}^{m}\pi_j g_j(x)$$

и данни x_i с неизвестна плътност $f(x)$, ядрената оценка, филтрирана със сместа g, се задава с уравнение (18.5), като в случая

$$\rho_j(x) = \frac{\pi_j g_j(x)}{g(x)}.$$

Тогава за всеки компонент на сместа g бихме могли да подберем подходяща в някакъв смисъл стойност на h и по този начин да изменяме константите в зависимост от дисперсиите на отделните компоненти във филтриращата смес. Практически това се прави като първо оценяваме някаква смес, изчисляваме за нея оптималните стойности на изглаждащите константи и оттам построяваме филтрираната оценка.

Изглаждащите константи в общия случай се получават като решение на минимизационна задача, чието решение не е известно в явен вид и изисква използването на числени процедури за приближено решаване.

До алтернативна формулировка на филтрираните ядрени оценки се стига, ако разгледаме смес от ядрени оценки, за която апостериорните вероятности $\rho_j(x)$ се използват като теглови коефициенти в сместа. Този подход позволява да се включва информация за носителя на плътността. Тогава оценката става

$$\tilde{f}(x) = \sum_{j=1}^{m}\rho_j(x)\left(\frac{1}{nh_j}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h_j}\right)\right), \tag{18.6}$$

което може да бъде записано още като

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\rho_j(x)}{h_j}K\left(\frac{x-x_i}{h_j}\right). \tag{18.7}$$

Информацията за носителя на f се отчита с помощта на условието $\rho_j(x) = 0$ за онези x, за които $f(x) = 0$. За да бъде гарантирано, че оценката наистина е плътност, трябва да е изпълнено

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{h_j}\int\rho_j(x)K\left(\frac{x-x_i}{h_j}\right)\,dx = 1.$$

Тъй като теглата не са фиксирани, а са функции на x, имаме потенциално различна смес за различните x, което гарантира отчитането на локалните особености

на плътността. Ще обърнем специално внимание на читателя, че докато в оценката (18.5) филтрираме по отношение на наблюдението т.е. на точката, в която е центрирано съответното ядро, в (18.7) филтрирането е по отношение на точката, в която искаме да оценим неизвестната плътност.

Накрая без доказателства ще отбележим някои асимптотични свойства на горните оценки. Ако са изпълнени условията за h_j и ядрата $K(\cdot)$ имат плътност с нулева средна и крайни втори моменти, то $\hat{f}(x)$ и $\tilde{f}(x)$ са състоятелни в слаб смисъл. Ако освен това съществуват вторите производни на f и на ρ_j, и втората производна на f е в L_2, то $\hat{f}(x)$ и $\tilde{f}(x)$ са състоятелни в смисъл на средноквадратична сходимост.

Тема 19

Probit и Survival анализи

Във тази глава ще представим накратко основните методи за моделиране (предсказване) на вероятността на сбъдване на някакво събитие като функция на други количествени фактори.

19.1 Logit и Probit анализи

19.1.1 Постановка

Да предположим, че оценяваме въздействието на концентрацията на някакъв препарат срещу насекоми - вредители. Прилагаме препарата си при различни концентрации $x_1 < x_2 < \cdots < x_n$ върху някакви групи от по $n_1, n_2, ..., n_n$ насекоми. В резултат от действието му $k_1, k_2, ..., k_n$ от насекомите умират. Искаме по тези резултати да оценим въздействието на концентрацията x върху вероятността за летален изход $p(x)$. Ясно е, че сл.в. k_i са биномни с разпределение $B(n_i, p(x_i))$.

Ще въведем за удобство следното преобразование на вероятността (то се нарича логистична трансформация):

$$u(p) = \ln(p/(1-p)). \tag{19.1}$$

Преобразованието е взаимно еднозначно и $\infty < u < \infty$.

Сега можем да разгледаме параметричен линеен модел за u:

$$u(p(x)) = \ln \frac{p(x)}{1-p(x)} = \beta_1 + \beta_2 x \tag{19.2}$$

и получаваме регресионен модел за вероятността:

$$p(x) = \frac{1}{1 + exp - (\beta_1 + \beta_2 x)} \tag{19.3}$$

За да бъде функцията нарастваща, трябва $\beta_2 > 0$.

Определение 19.1 *Логистично разпределение наричаме разпределение с ф.р. и плътност, съответно:*

$$F(x) = \frac{1}{1 + exp(-x)}, \qquad f(x) = \frac{exp(-x)}{(1 + exp(-x))^2}, \tag{19.4}$$

Плътността (19.4) е симетрична. Възможна е и друга параметризация:

$$\mu = -\frac{\beta_1}{\beta_2}, \quad \sigma = \frac{1}{\beta_2}.$$

Тогава регресионния модел (19.3) се превръща в

$$p(x) = F(\frac{x-\mu}{\sigma}). \tag{19.5}$$

Тук $F(x)$ може да бъде произволна ф.р. Например, когато поставим в (19.5) нормалната ф.р. $\Phi(x)$, наричаме получения модел Probit анализ.

19.1.2 Оценяване

За оценка на параметрите на модела (19.3) може да се използува метода на максимално правдоподобие. За съжаление, обаче, получените уравнения са трансцедентни. Затова се използува следния метод, който при достатъчно много наблюдения дава не по- лоши резултати.

Да разгледаме биномно разпределение $B(n,p)$ и означим с $\hat{p} = k/n$. Да означим $g(p) = \ln p/(1-p)$. При големи n имаме:

$$n^{1/2}(g(\hat{p}) - g(p)) \to N(0,\sigma) \tag{19.6}$$

$$\sigma^2 = (\frac{1}{p} + \frac{1}{1-p})^2 p(1-p) = \frac{1}{p(1-p)}. \tag{19.7}$$

Тук използувахме израза за производната на $g(p)$.

Следователно, ако означим с $Y_i = g(\hat{p}(x_i))$ получаваме

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

При това ϵ_i са независими, центрирани, нормални и имат дисперсия $\sigma_i^2/n_i = (n_i p(x_i)(1-p(x_i)))^{-1}$. Така получаваме, че оценките за коефициентите се получават по формулите:

$$\hat{\beta}_2 = \frac{\sum w_i(Y_i - \tilde{Y}(x_i - \tilde{x})}{\sum w_i(x_i - \tilde{x})^2}, \qquad \hat{\beta}_1 = \tilde{Y} - \hat{\beta}_2\tilde{x} \tag{19.8}$$

$$\tilde{Y} = \frac{\sum w_i Y_i}{\sum w_i}, \quad \tilde{x} = \frac{\sum w_i x_i}{\sum w_i}, \quad w_i = \frac{1}{\sigma_i^2}$$

19.2 Survival анализ

Да предположим, че изучаваме пациенти болни от (изобщо казано) смъртоносна болест. През времето на изучаването някои от тях умират, други напускат изследването, появяват се и нови пациенти. В края на изследването се интересуваме от влиянието на различни фактори върху времето за преживяване в дни (survival time) на пациентите. Наблюденията за пациенти, картона на които съдържа непълна информация, наричаме *цензурирани*. Например, " Лицето А преживя поне 2 месеца преди да замине от страната".

19.2.1 Life - tables

Интервалът време на наблюдения се разделя на подинтервали. За всеки от тях се записват числата:

- b_k -брой на пациентите в началото на интервала;
- e_k -брой на пациентите починали през това време;
- c_k -брой на пациентите напуснали изследването.

С всяко от тези числа е свързана и съответна честота. Така получената таблица може да се разглежда като разширена честотна таблица. От горните числа могат да се пресметнат ред допълнителни статистики:

- рискови случаи - $r_k = b_k - c_k/2$;
- честота на неудачните - $f_k = e_k/r_k$;
- честота на оцелелите - $1 - f_k$;

Броевете на умрели e_k през различните интервали се предполагат независими сл.в. (при условие фиксирани b_k).

Определение 19.2 *Наричаме функция на оцеляването (survival function) $S(t)$ наричаме произведението от вероятностите за оцеляване до момента t.*

Така извадъчната функция на оцеляване може да се определи като

$$S(t) = \prod_{k=1}^{t_k<t} (1 - f_k)$$

19.2.2 Моделиране

За анализ най-често се използват следните модели:

Модел на Кокс за пропорционалния риск

$$h(t, z_1, z_2, .., z_m) = h_0(t) * exp(b_1 * z_1 + ... + b_m * z_m) \tag{19.9}$$

Членът $h_0(t)$ се нарича начален (baseline) риск и чрез разделяне с него моделът може да се линеаризира:

$$log[h(t, z)/h_0(t)] = b_1 * z_1 + ... + b_m * z_m.$$

$h_0(t)$ се интерпретира като риска на един индивид при отсъствие ($z_k = 0, k = 1, \ldots, m$) на останалите фактори.

Моделът на Кокс може да се интерпретира като непараметричен - не се предполага нищо за разпределението на времената за оцеляване. Така за два индивида при еднакви стойности на факторите риска си остава същия (пропорционално на началния). Това предположение в много случаи си остава съмнително.

Например, след операционния риск зависи от възрастта на пациента, но след преминаването на първите дни тази зависимост би трябвало да намалява. Затова се въвежда следния модел:

Пропорционален риск с фактори зависещи от времето

$$h(t,z) = h_0(t) * exp(b_1 * z_1(t) + ... + b_m * z_m(t)) \tag{19.10}$$

При ускорени изпитания, например, за пробив на изолация на електрически кабел, напрежението бавно се повишава до настъпване на пробив. В този случай разбира се влиянието на фактора несъмнено зависи от времето. Точната формула на зависимостта трябва да се изясни от физични съображения.

Моделите (19.9) и (19.10) са от един и същи тип и лесно могат да бъдат сравнявани. Да допуснем, че в модел от вида (19.9) въведем един допълнителен член зависещ от времето. Когато коефициента b пред него е значим, трябва да отхвърлим хипотезата за "пропорционалност на риска".

Експоненциална регресия

Тук се предполага, че времената за оцеляване са подчинени на експоненциално разпределение. Това е вече параметричен модел. За параметъра на експоненциалното разпределение (м.о.) е в сила следния модел:

$$t(z) = exp(a + b_1 * z_1 + ... + b_m * z_m). \tag{19.11}$$

Нормална регресия и логнормална регресия

Тук се предполага, че времената за оцеляване са подчинени на нормално (или логнормално) разпределение.

$$t(z) = a + b_1 * z_1 + ... + b_m * z_m, \tag{19.12}$$

$t(z)$ - времето за оцеляване. Ако вместо $t(z)$ в (19.12) поставим $\ln(t)$, регресията става логнормална.

Тези два модела не могат директно да бъдат сравнявани.

Оценката на всички модели се прави по метода на максимално правдоподобие. За последните два той се превръща в обикновена регресия, но с тегла които се преизчисляват на всяка стъпка - (IRLS) Iterated Reweighted Least Squares.

Тема 20

Времени редове

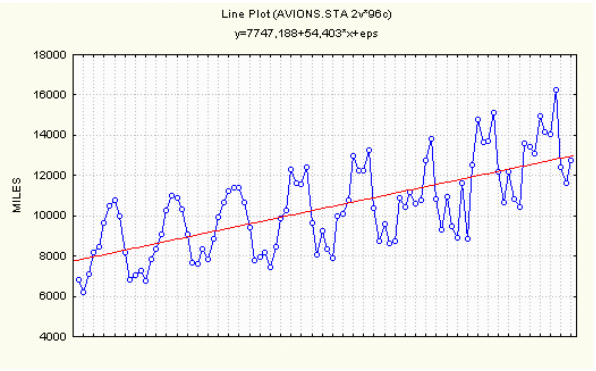
Много статистически данни се получават чрез наблюдаване на дадено явление последователно във времето. В този случай предположението за статистическа независимост на наблюденията в извадката може да се окаже невярно.Това води до необходимостта от изучаване на нови, по-сложни статистически модели. От друга страна, в практиката е изключително полезно умението да прогнозираме бъдещите прояви на "Н.В.Случай" точно тогава, когато той е проявявал определен характер в миналото.

20.1 Отстраняване на тренда

Прието е съвкупност от наблюдения, проведени през равни интервали във времето и върху един и същ обект или явление, да се нарича временен ред. Вероятно първият математически модел на временен ред е следният:

$$X(t) = f(t) + e(t) \tag{20.1}$$

Тук $X(t)$ е наблюдение или измерване, f е детерминирана функция на времето, например полином: $f(t) = a.t + b$ или тригонометрична функция: $f(t) = a.sin(w.t + b)$. За грешките $e(t)$ се предполага, както и в регресионния анализ, че имат нормално разпределение и са независими помежду си.



Фиг. 20.1: Британски авиолинии

Така поставена задачата за оценка на параметрите на тренда съвпада със задачата на регресионния анализ. В практиката много рядко се оказва, че такъв модел е адекватен - грешките се оказват зависими. Защо е така?

За разлика от класическия регресионен анализ тук строим модел за поведението на един и същ обект. За стохастичните модели е характерно, че в грешката се включва

влиянието на голям брой неотчетени фактори. При временните редове част от тях са присъщи на обекта, т.е. те влияят през цялото време и се отразяват на всички наблюдения. Това е главната причина за зависимостта.

Да погледнем например данните от пример от фиг. 20.1. Веднага се вижда нарастването на прелетените разстояния. На фигурата е показана и регресионна линия, наложена върху данните. Към линейното нарастване, обаче е наложена и сезонна компонента. Би трябвало тогава да търсим тренд във формата:

$$f(t) = b + at + \sum_{k=1}^{m} c_k sin(2\pi kwt) + d_k cos(2\pi kwt) \tag{20.2}$$

Тук $1/w = 12$ е дължината на основния период. Ако го направим обаче, ще получим остатъци, които хич няма да изглеждат добре.

Във временните редове се строят модели, които отчитат тази особеност. Така при оценяване на f се налага да се използуват методи, отчитащи взаимната корелация на грешките. Това налага подробно да се изучава зависимостта във времето, различните й форми и прояви. На това е посветена теорията на случайните процеси.

Съществуват два подхода в изследването на временните редове. Първият от тях е възникнал при изучаването на периодични явления, какъвто беше и примерът погоре (след отстраняване на тренда). Моделът, който стои в основата му, е наслагване на периодични колебания. Главната задача е да се определят амплитудите и честотите на тези колебания. Прието е този тип задачи да се нарича спектрален анализ. Без да даваме точно определение на понятието спектър, тук ще отбележим само, че една съществена негова характеристика е представянето на амплитудата на колебанията като функция от честотите. Особено пълно този подход е представен в (Дженкинс и Ваттс 1972).

Вторият подход е основан на понятието авторегресия - като предиктори се използуват предходните наблюдения. Основната цел тук е прогнозирането на неизвестното бъдещо наблюдение. Виж, например, класическата книга на Бокс и Дженкинс (Бокс и Дженкинс 1974). Тези два подхода са тясно свързани и е трудно да бъде намерена точна граница между тях.

20.2 Едномерен спектрален анализ

Синоними на това понятие са "хармоничен анализ" и "честотен анализ". Ще се позовем на книгата на (Бриллинджер 1980), за да покажем широкото използуване на спектралния анализ в човешкото знание.

Изучаването на отделни честотни компоненти във физиката вероятно за първи път е осъществил И.Нютон през 1664 г., когато е разложил слънчевата светлина с помощта на триъгълна призма. Спектроскопията, т.е. спектралният анализ, се използува сега при определяне наличието на химически елементи.

20.2.1 Понятия

Спектърът на мощността е обект на изследване в турбулентността и хидромеханиката, в акустиката и геофизиката, в медицината и психологията. Популярността на понятието спектър е обяснима, тъй като повечето природни явления имат периодичен характер. В (Бриллинджер 1980) могат да се намерят и много други интересни примери.

Период и честота

Нека сме наблюдавали временния ред от месечните полети в течение на 8 години или 96 месеца. Минималната единица за време е 1 месец. Съответно колебанията, които можем да наблюдаваме, са с периоди в диапазона от 2 до 96 месеца. Ако има такива с други периоди ние или няма да ги регистрираме изобщо, или грешно ще ги отнесем в горния интервал. Например, цикличността на температурата в рамките на денонощието не присъствува в месечните и записи.

Както знаем, известен е и друг начин за определяне дължината на периода, - броят на колебанията за единица време или честота. В системата СИ е приета единица за честота 1/сек или Херц (Hz). Тя би била удобна за временни редове, чиито измервания са през интервали, измерени в секунди. В нашия пример максималната наблюдаема честота е 1/2 [1/месец], а минималната - 1/120 [1/месец].

Тренд

Тук е необходимо да направим малко отклонение. За означаване на определени дългосрочни тенденции в реда се използува понятието тренд. Когато те имат периодичност с известен период (обикновено свързан с календара) говорим за сезонен тренд. Думата сезон тук и по-нататък трябва да се разбира като събирателно понятие. За месечни данни това ще бъдат месеците, за ежедневни - седемте дни на седмицата или 365-те дни в годината, за ежечасни - 24-те часа на денонощието и т.н.

Понятието тренд е до голяма степен условно. Това, което в рамките на един човешки живот изглежда дългосрочен тренд, може да се окаже просто случайна флуктуация или периодично явление в геологични мащаби.

Трендът може да се разглежда като детерминирана функция на времето, но често към него се причислява и част от нискочестотните колебания. Аналогично е положението и при сезонния тренд.

Автокорелация

Ако в реда съществува периодичност, то корелацията на наблюденията, проведени на разстояние, кратно на един период, следва да е голяма и положителна. В нашия пример очевидно ще присъстват сезонните колебания на пътуванията. Това предполага, че корелационните коефициенти на наблюдения с отместване, кратно на 12 ще са близки до 1. Ако се оценят всички възможни корелации с различни отмествания, по тях е възможно да се открият изразени периодичности.

Много по-удобна е техниката, наречена "трансформация на Фурие". Сега няма да се спираме на подробности (виж следващите два параграфа), но с нейна помощ се проявяват всички съставящи колебания. Всяко колебание се характеризира с три числа: честота, амплитуда и фаза, и се описва с тригонометричната функция $A.sin(W.t + B)$.

Периодограма

Амплитудата A показва максималното отклонение от равновесното положение, W е честота на колебание, а фазата B - отклонението от нулата в началния момент.

Алгоритмите за дискретно преобразование на Фурие позволяват изчисляването на амплитудите и фазите на колебания с честоти, кратни на минималната. Наборът от така получени амплитуди се нарича периодограма и е основа за статистическите изводи в спектралния анализ.

За съжаление, в периодограмата е отразена твърде силно случайността на наблюденията. Оценките на отделните амплитуди в спектъра на реда - теоретичния набор от амплитуди, са неизместени и стават почти независими с увеличаване броя на наблюденията, но дисперсията им не намалява.

20.2.2 Пример за спектрален анализ

Корелационният коефициент на наблюдения през равно разстояние във времето (k единици) се нарича автокорелация за "лаг" k. Стационарен временен ред се нарича ред, чиито автокорелации зависят само от лага, но не и от избора на наблюденията, или с други думи - ред, за описанието на който е безразлично откога се наблюдава.

Естествено и математическото очакване на наблюденията в един стационарен временен ред е постоянно. Това изискване не позволява наличие на тренд, например линеен, в реда. Предполагаме също така, че връзката между наблюденията /корелацията/ намалява с отдалечаването им едно от друго.

Като пример за извършване на спектрален анализ с помощта на пакета Statistika пак ще разгледаме данните от фиг. 20.1. От графиката се установява наличието на два вида тренд в реда:

1. линеен, непериодичен, изразяващ се в обща тенденция на нарастване
2. периодичен (сезонен) - колебания с период около 12 месеца.

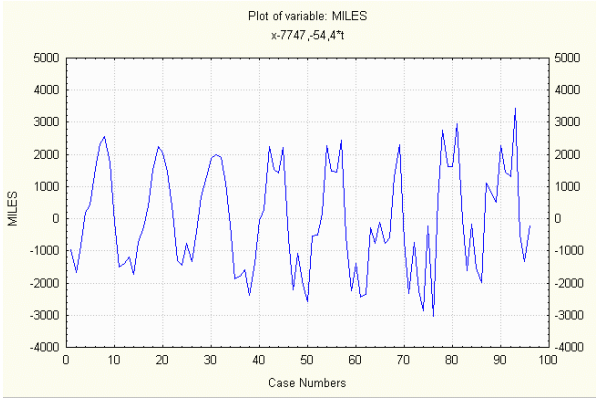
Въоръжени с тази информация първо ще отстраним с процедурата за трансформации линейния тренд.

Линеен тренд

Спектралният анализ на временни редове почти винаги трябва да започне с тази стъпка. Като минимум, преди изчисляването на автокорелационната функция и спектъра е необходимо да се извърши центриране на реда. В противен случай ненулевата средна стойност ще затрудни интерпретацията на споменатите функции. Нормирането на реда не е задължителна операция, но се препоръчва за да се избегне

работата с големи числа. Тези две операции се извършват с функцията СТАНДАРТИЗИРАНЕ. На графиката на реда естествено тази операция не си личи.

Още по-лошо се отразява на спектъра наличието на променлива средна стойност, както е при разглеждания ред. Несезонната компонента, която изглежда линейна, напълно би доминирала в автокорелацията и спектъра.

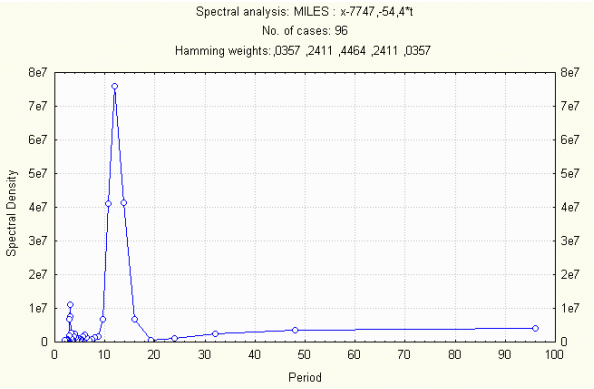


Фиг. 20.2: Данните след отстраняване на тренда

Нека отстраним линейния тренд. Това може да се направи по два начина - чрез ДИФЕРЕНЦИРАНЕ или с помощта на ЛИНЕЕН ТРЕНД. Нека използуваме втората функция, като по-естествена и стандартизираме получения ред.

Графиката (фиг.20.2) на получения ред показва, че общо взето несезонната компонента е премахната успешно.

Сезонен тренд



Фиг. 20.3: Спектър

Сега следва да анализираме периодичността с трансформация на Фурие. Пресмятаме периодограма. Тъй като получената периодограма не е достатъчно гладка функция, е по-удобно да я изгладим. Така получаваме по-добра представа за спектъра.

На фигурата веднага се вижда огромната стойност на главната периодика - 12 месеца.

В разглеждания пример, и редът, и корелационната му функция (фиг.20.3) показват годишна периодичност. В спектъра обаче, информацията е в известна степен по-подробна - наблюдаваме и втори по-малък връх, съответстващ на период около три месеца. По принцип той може да се дължи на изтичане на мощност от големия връх. В нашия случай това изглежда малко вероятно. За борба с това явление е предвидена функцията ПОДТИСКАНЕ НА РЕДА (tapering), която умножава данните (краищата на реда) с косинусов или друг прозорец. Неговата графика за различни стойности на параметъра може да се получи чрез прилагането на ПОДТИСКАНЕ

НА РЕДА към ред, съставен от единици.

Сезонно диференциране

Обичаен метод за отстраняване на сезонен тренд е сезонното диференциране, което беше дефинирано в предния параграф чрез трансформацията $y(t) = x(t) - x(t-s)$. Удобно е да въведем оператора преместване назад B, който действува така: $Bx(t) = x(t-1)$. С негова помощ сезонното диференциране се записва по следния начин.

$$y(t) = (1 - B^s)x(t) \tag{20.3}$$

Операторът B се нарича още лагов, а операторът B^s - сезонен лагов. По-сложни оператори във временните редове са линейните филтри. Например, операторната форма на филтъра

$$y(t) = x(t) - \sqrt{3}x(t-1) + x(t-2),$$

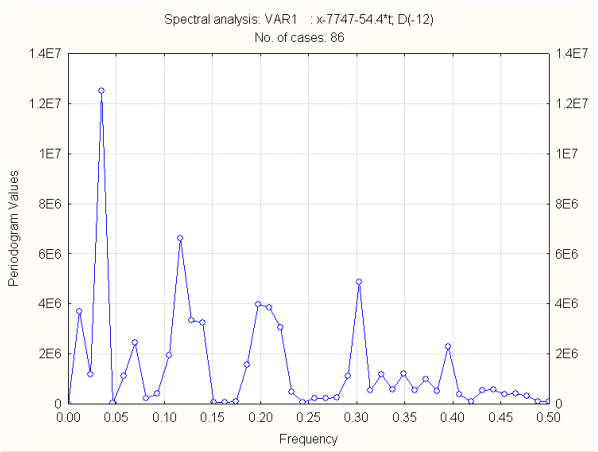
ще бъде

$$y(t) = (1 - \sqrt{3}B + B^2)x(t). \tag{20.4}$$

Не е трудно да се забележи, че формално погледнато, операторите на линейните филтри се представят като полиноми на лаговия оператор B и могат да се разлагат на множители. Това представяне на филтрите е полезно поради това, че последователното прилагане на два филтъра е еквивалентно на прилагането на филтър определен от произведението на техните представяния. Лесно може да се провери, че един от множителите на (20.3) при $s = 12$ е (20.4), т.е.

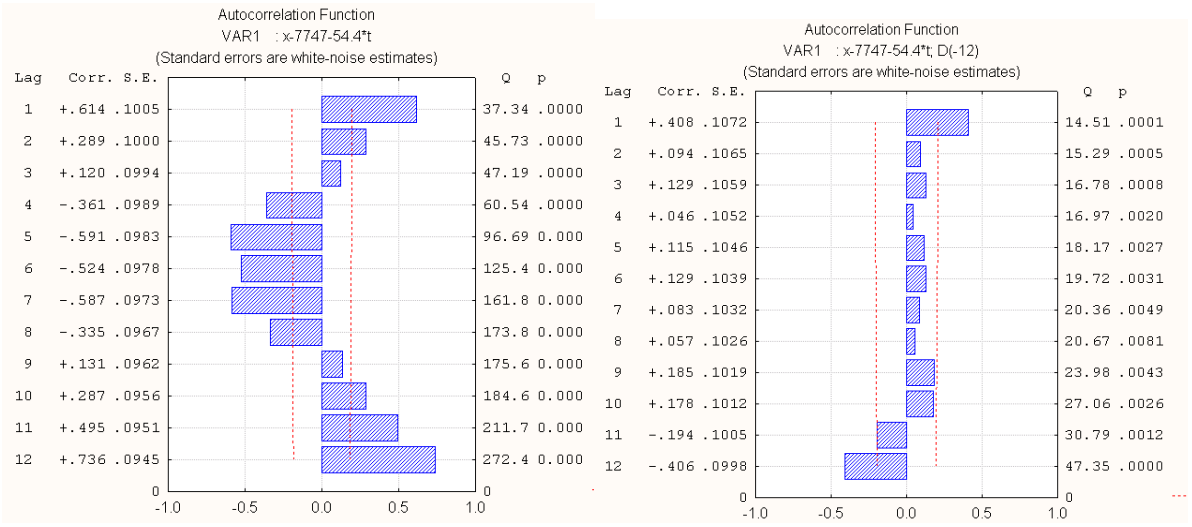
$$1 - B^{12} = (1 - \sqrt{3}B + B^2)(1 + B + B^2)... \tag{20.5}$$

В нашия пример е уместно да използуваме оператора на СЕЗОННО ДИФЕРЕНЦИРАНЕ (20.5).



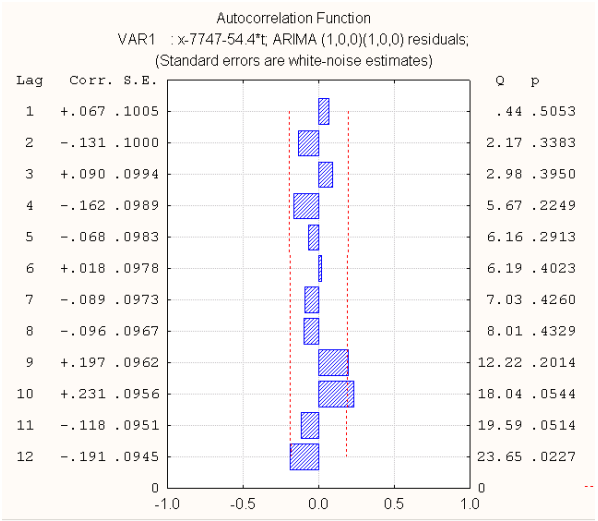
Фиг. 20.4:

Качественият резултат се вижда в периодограмата (фиг.20.4) на получения ред - на всички кратни на $1/12$ честоти мощността е сведена до нула. Възможно е използваният оператор $1 - B^{12}$ да е твърде груб - само една-две от тези честоти изглеждаха съществени. Това, че сме прекалили, се потвърждава и от разглеждане на автокорелационните функции преди и след диференцирането:



Фиг. 20.5: Автокорелации преди и след диференцирането Можем да опитаме да оценим коефициента пред оператора B^{12}. Така наречената ARIMA.

```
Variable: VAR1    : x-7747-54.4*t
Transformations:
          Model: (1,0,0)(1,0,0)  Seasonal lag: 12
 No. of obs.:  96      Initial SS= 2427E5 Final SS= 6794E4(28.00%)  MS= 7305E2
 Parameters (p/Ps-Autoregressive, q/Qs-Moving aver.); highlight: p<.05
             Const.  p(1)  Ps(1)
  Estimate: -60.46 .51266 .82209
  Std.Err.: 427.53 .09288 .07657
```



Фиг. 20.6: Автокорелации на остатъците

Най-сетне като че ли всичко е в ред??.

20.3 ТЕОРИЯ

20.3.1 Автоковариация

Нека $\xi(t), t = 0, 1, 2, ...,$ е безкрайна редица от случайни величини. Тя се нарича стационарна в широк смисъл, ако са изпълнени равенствата:

$$\mathbf{E}\,\xi(t) = 0, \quad \mathbf{E}\,\xi(t).\xi(t+s) = C(s). \tag{20.6}$$

Функцията $C(s)$ се нарича автоковариационна функция.

20.3.2 Спектър

Спектрална плътност (или спектър) на $\xi(t)$ наричаме функцията:

$$f(w) = (2\pi)^{-1} \sum_s C(s).exp(-i.w.s). \tag{20.7}$$

Тя е определена при условие, че $\sum ||C(s)|| < \infty$. и е реална, неотрицателна, симетрична и периодична с период 2π. Верна е и следната формула за обръщане (теорема на Бохнер-Хинчин):

$$C(s) = \int_{-\pi}^{\pi} f(w).exp(i.w.s).dw \tag{20.8}$$

20.3.3 Дискретна трансформация на Фурие

Нека числовата редица $x(t), t = 0, 1, 2, ..., T-1$, е крайна. Тогава комплексната функция:

$$F(w) = \sum_{t=0}^{T-1} x(t).exp(-i.w.t). \tag{20.9}$$

се нарича Дискретна Трансформация на Фурие (ДТФ) и се разглежда само в точките $w = 2\pi.s/T, s = 0, 1, 2, ..., T-1$. Формулата за обръщане на ДТФ е

$$x(t) = T^{-1} \sum_w F(w).exp(i.w.t). \tag{20.10}$$

20.3.4 Бърза трансформация на Фурие

Бързата трансформация на Фурие (БТФ) е алгоритъм за пресмятане на ДТФ, основан на твърдението: Нека R и S са взаимно прости числа и $T = R.S$. Тогава

$$F(w) = \sum_a \sum_b y(sR + rS).exp(-2\pi i(ar/R + bs/S), \tag{20.11}$$

където $y(t) = x(t \mod T), w = 2\pi c/T, a = c \mod R, b = c \mod S, a = 0, 1, 2, ..., R-1, b = 0, 1, 2, ..., S-1$. Аналогично твърдение е вярно и в по-общ случай, когато T се

разлага на повече от два прости множителя, и позволява пресмятанията на ДТФ да се реализират итеративно със значително съкращаване на броя на операциите (вж (Cooley and J.W.Tukey 1965)).

20.3.5 Асимптотични свойства на ДТФ

Ако редицата $x(t), t = 0, ..., T-1$, се състои от наблюдения върху стационарен временен ред и T клони към безкрайност, то стойностите на F в различните честоти (не допълващи се до π) са асимптотично независими и нормално разпределени (в комплексната равнина), с нулева средна стойност и нарастваща дисперсия $2\pi T f(w)$.

20.3.6 Периодограма

Периодограма се нарича функцията

$$(2\pi T)^{-1}||F(w)||^2. \tag{20.12}$$

Тя е неизместена оценка на спектралната плътност $f(w)$, но дисперсията ѝ е постоянна при $T \to \infty$. Състоятелна оценка на спектралната плътност се получава чрез усредняване с подходящи тегла $c(k)$ на нарастващ (с T) брой съседни стойности на периодограмата:

$$G(w_k) = \sum_{j=-m}^{m} c(j)F(w_{k-j}) \tag{20.13}$$

В СТАТЛАБ теглата са равни, т.е. $c(k) = 1/(2m+1)$.

20.3.7 Подтискане на реда.

За отслабване на явлението изтичане на мощност преди БТФ данните се преобразуват по формулата

$$y(t) = c(t).x(t), t = 0, 1, ..., T-1, \tag{20.14}$$

където $c(t)$ е подходяща функция, наречена прозорец. В СТАТЛАБ се използува косинусов прозорец (с параметър $k < T/2$):

$$c(t) = \begin{cases} (1 - cos(\pi(t/k))/2 & t < k, \\ 1 & k < t < T-k, \\ (1 - \cos(\pi(T-1-t)/k)) & t > T-k-1. \end{cases} \tag{20.15}$$

Литература

Akaike, H. (1973). A new look at the statistical model identification. *IEEE Trans. Aut. Contr. 19*, 716–723.

Akoka, J. (1992). A new algorithm for generation of clusters. In G. L. C. Serge Joly (Ed.), *Distancia 92*. Universite de Rennes 2.

Cooley, J. and J.W.Tukey (1965). An algorithm for the machine calculation of complex fourier series. *Math.Comp. 19*, 297–301.

Cox, T. and M. Cox (1994). *Multidimensianal Scaling*. N.Y.: Chapman & Hall.

C.R.Rao (1965). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Dixon, J. (Ed.) (1981). *BMDP Statistical Software - 81*. Los Angeles: University of California.

Einslein, K., A. Ralston, and H. S. Wilf (Eds.) (1977). *Statistical Methods for Digital Computers*. New York: John Wiley & Sons.

Fletcher, R. (1971). A modified marquardt subroutine for nonlinear least squares. Technical Report R6799, AERE.

Gower, J. (1966). Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika 53*, 325 – 338.

Harman, G. (1972). *Contemporary factor analysis*. Moskwa: Statistika. (russian translation).

Hartigan, J. (1975). *Clustering Algorithms*. Joh Wiley.

Jennrich, R. I. (1977). Stepwise discriminant analysis. See Einslein, Ralston, and Wilf (1977), pp. 76–95.

Kowalik, J. and M. Osborn (1968). *Method for unconstrained optimisation*. American Elsevier Publishing Company.

Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika 29*, 115 – 129.

Lane, D. M. (1984-1998). *Rice Virtual Labs in Statistics*.

Mathar, R. (1985). The best eucledian fit to to a given distance matrix in prescribed dimensions. *Linear Algebra Applic. 67*, 1 – 6.

Nelder, J. and R. Meed (1965). A simplex method for function minimisation. *Computer J. 7*, 308–313.

Pollard, J. (1982). *Guide in computational methods of statistics*. Moskwa: Finansi i statistika. (russian translation).

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimates. *Scandinavian Journal of Statistics 9*, 65–78.

Schoenberg, I. (1935). Remarks to Mourice Frechet's article "Sur la Definition axiomatique d'une classe d'espaces vectoriels distances applicables vectoriellement sur l'espace de Hillbert"
. *Ann.Math. 36*, 724 – 732.

Torgerson, W. (1958). *Theory and Method of Scaling*. New York: John Wiley & Sons.

T.W.Anderson (1958). *Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Wilks, S. (1967). *Mathematical statistics*. Moskwa: Nauka. (russian translation).

Yamaoka, K., Y. Tanigawara, T. Nakagawa, and T. Uno (1981). A pharmacokinetic program (multi) for microcomputer. *J. Pharm. 4*, 879–885.

Yan, W. and H. Wale (1974). *Factor analysis and its applications*. Sofia: Tehnika. (bulgarian translation).

Young, G. and A. Hauseholder (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika 3*, 19 – 22.

Бокс, Г. и Г. Дженкинс (1974). *Анализ временных рядов. Прогноз и управление*, Volume 1,2. Москва: Мир.

Бриллинджер, Д. (1980). *Временные ряды. Обработка данных и теория*. Москва: Мир.

Въндев, Д. и П. Матеев (1988). *Статистика с Правец 82*. София: Наука и Искуство.

Г.П.Климов (1975). *Приложна математическа статистика*. София: Наука и изкуство.

Демиденко, Е. З. (1989). *Оптимизация и регрессия*. Москва: Наука, Гл. ред. физ.-мат. лит.

Дженкинс, Г. и Д. Ваттс (1972). *Спектральный анализ и его приложения*, Volume 1,2. Москва: Мир.

Химмельблау, Д. (1973). *Анализ процессов статистическими методами*. Москва: Мир.

Химмельблау, Д. (1975). *Прикладное нелинейное програмирование*. Москва: Мир.