Bodové odhady parametrů a výstupů

26. listopadu 2013

Máme rozdělení s neznámými parametry a chceme odhadnout jeden nebo několik příštích výstupů. Již víme, že úplnou informaci v této situaci nese sdružené rozdělení pravděpodobnosti a víme také, jak toto sdružené rozdělení zkonstruovat. Umíme se také vyrovnat s několika technickými problémy, na které můžeme narazit.

V tomto materiálu si ukážeme, jak ze sdruženého rozdělení zkonstruovat intervalové a bodové odhady parametrů i výstupů.

Intervalové odhady

Pokud bychom chtěli zachovat co nejvíce informace, zvláště o vzájemné závislosti veličin, mohli bychom z definičního oboru sdruženého rozdělení vyříznout oblast s např. 95% nejpravděpodobnějších bodů. Tím bychom dostali mnohorozměrný útvar, jehož tvar bychom těžko popisovali. Proto touto cestou nepůjdeme a omezíme se na intervalové odhady vždy pro jedinou neznámou.

Klíčové je tedy spočítat marginální rozdělení pro jednotlivé neznámé. To již umíme. Vezmeme sdružené rozdělení a zintegrujeme (určitý integrál přes celý definiční obor) přes všechny ostatní neznámé (tj. vyjma té, jejíž marginální rozdělení počítám).

V případě diskrétního rozdělení místo integrování sčítám.

Když již marginální rozdělení znám, pak buď integrací nebo pomocí tabulek uříznu na krajích příslušné procento a zbyde mi příslušný interval spolehlivosti pro příslušnou neznámou.

Hotovo.

Nevýhodou tohoto postupu je ona integrace. To je často ruční práce, ke které navíc potřebujeme mít sdružené rozdělení napsané v nějakém jednoduchém tvaru, což není vždy splněno.

Proto se často místo intervalových odhadů používají bodové odhady, které sice ztrácí ještě více informace, ale jsou podstatně jednodušší.

Bodové odhady

Při bodových odhadech se snažíme najít buď střední hodnoty příslušných veličin nebo souřadnice nejvyššího bodu sdruženého rozdělení. První přístup je správnější, druhý jednodušší. Rozdíl ve výsledcích je často zanedbatelný.

Často si můžeme mnoho práce ušetřit, protože pro nejčastější typy modelů jsou odvozeny jednoduché algoritmy, jak získat bodové odhady i bez hledání maxima sdruženého rozdělení.

Lineární regresní model s normálním šumem

Tento model má tvar: $y_n = \theta \cdot \psi + e_n$, kde θ je vektor parametrů $\theta = (b_0, a_1, b_1, a_2, b_2, \ldots k)$, ψ je vektor vstupů $\psi = (x_n, y_{n-1}, x_{n-1}, y_{n-1}, x_{n-2}, y_{n-2}, \ldots 1)$ a e_n je šum s normálním rozdělením s nulovou střední hodnotou a rozptylem R. Tedy $e_n \sim N(0, R)$.

Odhad parametrů

Neznámé parametry jsou tady jednak ve vektoru θ, jednak je to neznámý rozptyl šumu R.

Jejich bodové odhady snadno získáme pomocí datové matice a počítadla:

Nejprve vyrobíme datový vektor, který bude mít na prvním místě výstup y_n a na dalších místech členy vektoru vstupů:

$$\Psi = (y_n, x_n, y_{n-1}, x_{n-1}, y_{n-1}, x_{n-2}, y_{n-2}, \ldots 1).$$

Datovou matici vyrobíme tak, že vynásobíme datový vektor nejprve ve sloupcové podobě a pak v řádkové podobě:

$$D = \Psi' \cdot \Psi.$$

Všechny datové matice, do kterých postupně dosazujeme data, sečteme a získáme informační matici V:

$$V = \sum_{i=1}^{\kappa} D_i,$$

kde κ je počítadlo datových matic.

Dále matici V rozdělíme na čtyři části tak, že oddělíme první řádek a první sloupec:

V_y	*	*	*
*	*	*	*
v	*	M	*
*	*	*	*

Číslo vlevo nahoře pojmenuji V_y. Sloupečkový vektor pojmenuji v (jako vektor) a zbylou matici po odříznutí prvního řádku a sloupce pojmenuji M (jako matice).

Bodový odhad pro vektor θ spočtu dle vzorce:

$$\hat{\theta} = \left(M^{-1}\right) \cdot v.$$

Výsledkem je sloupcový vektor bodových odhadů pro jednotlivé parametry vektoru θ.

Bodový odhad pro rozptyl spočtu dle vzorce:

$$\hat{R} = \frac{V_y - v' \cdot \left(M^{-1}\right) \cdot v}{\kappa}.$$

Příklad

Načtěte data z odkazu „Data ke cvičení" na našich stránkách a pokuste se tato data modelovat lineárním regresním modelem čtvrtého řádu s normálním šumem a bez konstanty. Proveďte bodové

odhady parametrů a rozptylu šumu.

Máme tedy model: $y_n = A \cdot y_{n-1} + B \cdot y_{n-2} + C \cdot y_{n-3} + D \cdot y_{n-4} + e_n$, kde $e_n \sim N(0, R)$. Pokusíme se odhadnout parametry A, B, C, D a R:

```
clear all;
clc;
load DataCviceni;

V=zeros(5,5);        %Pocatecni nastaveni inf. matice
K=0;                    %a pocitadla

for n=5:200
    P=[y(n),y(n-1),y(n-2),y(n-3),y(n-4)];   %Datovy vektor
    D=P'*P;                                 %Datova matice
    V=V+D;                                  %Informacni matice
    K=K+1;                                  %Pocitadlo
end;
                        %Rozdeleni inf. matice na casti:
Vy=V(1,1);                %Cislo
M=V(2:end,2:end);        %Matice
v=V(2:end,1);            %Vektor
                        %Bodovy odhad
Theta=(M^-1)*v            %Linearnich parametru
R=(Vy-v'*(M^-1)*v)/K    %Rozptylu
```

Získáme tyto výsledky:

$$\hat{\theta} = \begin{pmatrix} 1,6267 \\ -0,8001 \\ 0,1939 \\ -0,0203 \end{pmatrix},$$

$$\hat{R} = 109,1426.$$

Odhad výstupů

Když mám bodové odhady parametrů, mohu je považovat za správné hodnoty parametrů a s nimi počítat.

Při výpočtu střední hodnoty budu za šum dosazovat jeho nejpravděpodobnější hodnotu a tou je nula. (Gaussovka se střední hodnotou nula má nejvyšší hodnotu v nule.)

Spočtu tedy jednokrokový bodový odhad výstupu $\widehat{y_{n+1}}$ takto:

$$\widehat{y_{n+1}} = \hat{A} \cdot y_n + \hat{B} \cdot y_{n-1} + \hat{C} \cdot y_{n-3} + \hat{D} \cdot y_{n-3} + 0 = \hat{\theta} \cdot \psi,$$

kde za parametry dosazuji jejich bodové odhady.

V tomto případě mohu velmi snadno získat i intervalový odhad. Neurčitost výstupu je totiž dána neurčitostí parametrů a rozptylem šumu. Rozptyl šumu zůstává stále týž, kdežto rozptyly parametrů velmi rychle klesají, takže jsou brzy proti rozptylu šumu zcela zanedbatelné. Když rozptyl parametrů zcela zanedbám, bude gaussovka pro výstup odpovídat gaussovce pro šum, jen bude posunutá z nuly do hodnoty $\widehat{y_{n+1}}$:

$$y_{n+1} \sim N\left(\widehat{y_{n+1}},\, R\right).$$

Mohu pak s tímto rozdělením pracovat a dělat např. intervalové odhady.

Příklad

Navazujeme na předchozí příklad a snažíme se odhadnout 95%-ní interval spolehlivosti pro 201. výstup.

```
clear all;
clc;
load DataCviceni;

Theta=[1.6267, -0.8001, 0.1939, -0.0203];
Psi=[y(200), y(199), y(198), y(197)];
R=109.1426;

EY201=Theta*Psi';  %Stredni hodnota

DolniMez=normal_inv(0.025,EY201,R) %Uriznu 2.5% vlevo
HorniMez=normal_inv(0.975,EY201,R) %Uriznu 2.5% vpravo
```

Získáme 95%-ní interval spolehlivosti: $\langle 282,2;\ 323,2\rangle$.

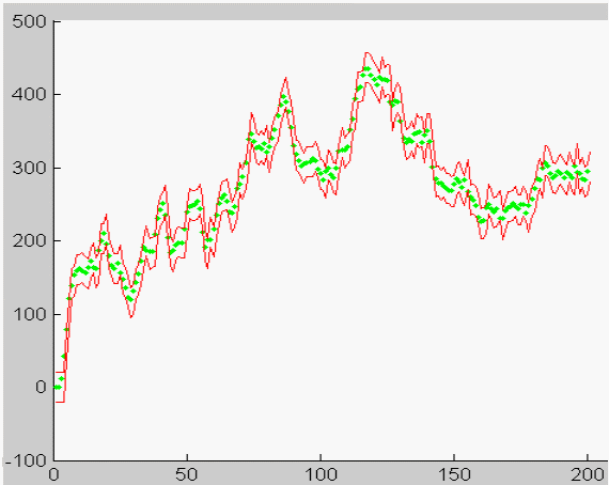
Funkce normal_inv je ze statistického balíčku, který si můžete stáhnout ze stránek Ivana Nagyho. Nápovědu najdete příkazem hlp, bod 3.

Další příklad

V minulém příkladě jsme odhadovali jen 201. výstup. Vytvořte program, který bude vykreslovat 95%-ní interval spolehlivosti pro všechny výstupy pátým počínaje. Parametry uvažujte ty, které jsme získali bodovým odhadem.

Projedeme data a pro každou hodnotu n od tří dále provedeme bodový odhad, který nazveme z. Pak spočteme příslušný 97,5%-ní kvantil normálního rozdělení, který nám udává, kolikanásobek směrodatné odchylky musíme vzít, abychom dostali právě 95%-ní interval. (Tj. uřízli 2,5% na každé straně.) Tento příslušný násobek pak jednou přičteme, jednou odečteme, čímž získáme horní i dolní hranici intervalu.

```
clear all;
clc;
load DataCviceni;

Theta=[1.6267, -0.8001, 0.1939, -0.0203];
R=109.1426;

z(1)=0;
z(2)=0;
z(3)=0;
z(4)=0;

for n=4:200
    Psi=[y(n), y(n-1), y(n-2), y(n-3)];
    z(n+1)=Theta*Psi';  %Stredni hodnota
end;

Smo=sqrt(R);
Kvantil=normal_inv(0.975,0,1);
z1=z+Kvantil*Smo;
z2=z-Kvantil*Smo;

hold on;
plot(y,'g.');
plot(z1,'r-');
plot(z2,'r-');
hold off;
```



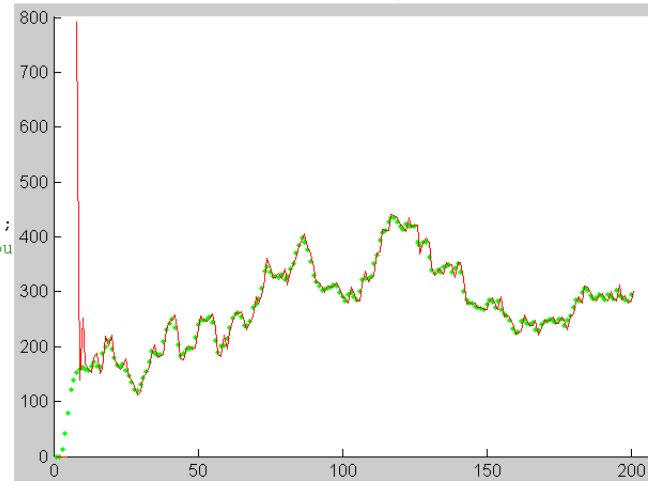
Vidíme, že naprostá většina (cca 95%) dat (zeleně) opravdu sedí v odhadovaném intervalu (červeně).

První čtyři z jsme navolili jen proto, aby se vektor z snadno vykreslil jedničkou počínaje.

Ještě další příklad

V minulém příkladu bylo poněkud nepřirozené, že jsme průběžně odhadovali výstupy, ale parametry jsme už měli určeny ze všech dat. V reálných případech odhadujeme výstupy i parametry z těch dat, která máme v té chvíli k dispozici.
Vytvořte program, který bude odhadovat výstupy pátým počínaje. Ať vykreslí graf dat i jejich bodových odhadů. Parametry odhadujte průběžně.

```
clear all;
clc;
load DataCviceni;

z(1)=0;     %Parametry mohu odhadovat az od 6.
z(2)=0;     %Proto prvnich pet z nejak zadam
z(3)=0;
z(4)=0;
z(5)=0;

V=zeros(5,5);
K=0;

for n=5:200
    P=[y(n), y(n-1), y(n-2), y(n-3), y(n-4)]; %Datovy vektor
    D=P'*P; %Datova matice
    V=V+D;  %Informacni matice
    K=K+1;  %Pocitadlo

    Vy=V(1,1);  %Rozsekani matice
    v=V(2:end,1);
    M=V(2:end,2:end);

    Theta=(M^-1)*v; %Odhad parametru
    Psi=[y(n), y(n-1), y(n-2), y(n-3)];
    z(n+1)=Theta'*Psi';  %Odhad vystupu
end;

hold on;
plot(y,'g.');
plot(z,'r-');
hold off;
```



Vidíme, že program sice zpočátku odhadoval zcela špatně (šestou a sedmou hodnotu kvůli singulární informační matici dokonce nespočetl vůbec), ale od nějaké dvacáté hodnoty je už odhad velice dobrý. Odhady parametrů se už totiž ustálily a pak už se jen trochu zpřesňovaly.

Všimněte si, že vykreslovat např. 95%-ní interval spolehlivosti jsme v tomto případě nemohli, protože nepřesnost odhadu je zpočátku dána zejména nepřesností parametrů a nikoli jen šumem. Interval spolehlivosti bychom museli počítat přes marginální rozdělení.

Vícekrokový odhad

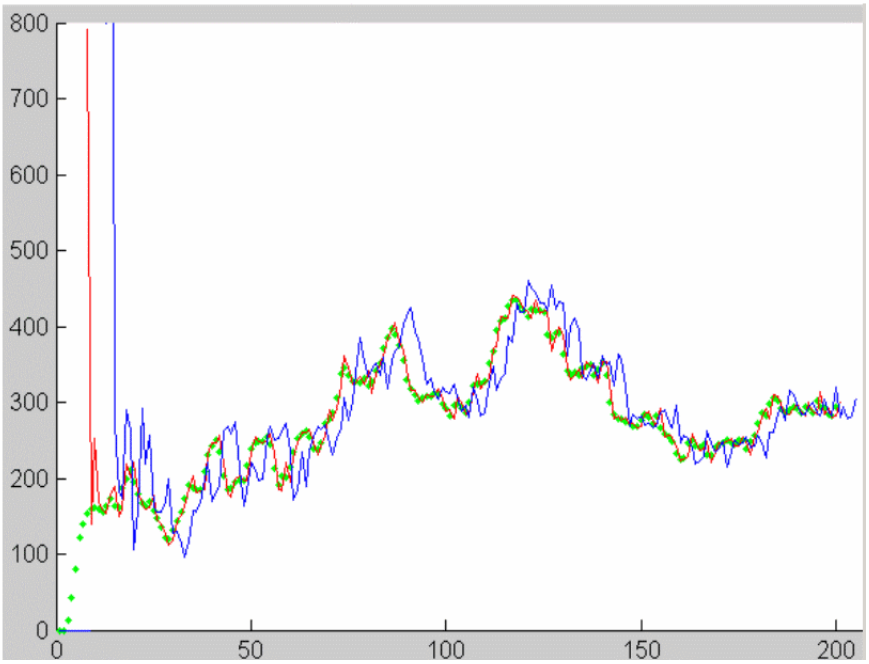
My již umíme, alespoň pro gaussovský šum, ze známých hodnot a známého modelu vyrobit rozdělení pro hodnoty o několik kroků dopředu. Tím se zde již zabývat nebudeme a soustředíme se jen na výpočet středních hodnot výstupu na několik kroků dopředu.

To je velmi jednoduché, protože do odhadu na dva kroky dopředu jen dosadíme odhad jednokrokový, do odhadu na tři kroky dopředu dosadíme odhad dvoukrokový, atd.

Příklad

Do grafu z minulého příkladu vykreslete ještě odhad na pět kroků dopředu.

```
clear all;
clc;
load DataCviceni;

z(1)=0;      %Parametry mohu odhadovat az od 6.
z(2)=0;      %Proto prvnich pet z nejak zadam
z(3)=0;
z(4)=0;
z(5)=0;

V=zeros(5,5);
K=0;

for n=5:200
    P=[y(n), y(n-1), y(n-2), y(n-3), y(n-4)]; %Datovy vektor
    D=P'*P; %Datova matice
    V=V+D;  %Informacni matice
    K=K+1;  %Pocitadlo

    Vy=V(1,1);  %Rozsekani matice
    v=V(2:end,1);
    M=V(2:end,2:end);

    Theta=(M^-1)*v; %Odhad parametru
    Psi=[y(n), y(n-1), y(n-2), y(n-3)];
    z1(n+1)=Theta'*Psi';  %1-krokovy odhad vystupu

    Psi=[z1(n+1), y(n), y(n-1), y(n-2)];
    z2(n+2)=Theta'*Psi';  %2-krokovy odhad vystupu

    Psi=[z2(n+2), z1(n+1), y(n), y(n-1)];
    z3(n+3)=Theta'*Psi';  %3-krokovy odhad vystupu

    Psi=[z3(n+3), z2(n+2), z1(n+1), y(n)];
    z4(n+4)=Theta'*Psi';  %4-krokovy odhad vystupu

    Psi=[z4(n+4), z3(n+3), z2(n+2), z1(n+1)];
    z5(n+5)=Theta'*Psi';  %5-krokovy odhad vystupu
end;

hold on;
plot(y,'g.');
plot(z1,'r-');
plot(z5,'b-');
hold off;
```



Vidíme, že pětikrokový odhad je mnohem méně přesný než jednokrokový.
(Pokud byste chtěli tento obrázek vykreslit, je třeba nastavit rozsah osy y na 0 až 800.)

Diskrétní model

Odhad parametrů

Bodový odhad parametrů v diskrétním modelu je velmi jednoduchý. Nejprve si vytvoříme tabulku četností. Ta je obdobou tabulky podmíněných pravděpodobností.

Procházíme data a za každý případ, který nastal, připíšeme do příslušného políčka jedničku.

Poté každý řádek tabulky znormujeme, aby jeho součet byl jedna.

Příklad

Proveďte bodové odhady parametrů $\theta_{1|1}$, $\theta_{1|2}$, $\theta_{2|1}$ a $\theta_{2|2}$ v diskrétním modelu prvního řádu.
Data y: 1, 1, 2, 1, 2, 2, 2, 1, 2, 1, 1, 2.

Diskrétnímu modelu prvního řádu přísluší následující tabulka podmíněných pravděpodobností:

y_{n-1}	$P(y_n = 1)$	$P(y_n = 2)$
1	$\theta_{1\|1}$	$\theta_{2\|1}$
2	$\theta_{1\|2}$	$\theta_{2\|2}$

Vytvořme podobnou tabulku pro četnosti:

y_{n-1}	$P(y_n = 1)$	$P(y_n = 2)$
1	2	4
2	3	2

Dvojka vlevo nahoře znamená, že jednička následovala po jedničce dvakrát. Čtyřka vpravo nahoře znamená, že dvojka následovala po jedničce čtyřikrát. Atp.

Tabulku znormujeme, aby součet v každém řádku byl 1:

y_{n-1}	$P(y_n = 1)$	$P(y_n = 2)$
1	$\hat{\theta}_{1\|1} = \frac{2}{6}$	$\hat{\theta}_{2\|1} = \frac{4}{6}$
2	$\hat{\theta}_{1\|2} = \frac{3}{5}$	$\hat{\theta}_{2\|2} = \frac{2}{5}$

To jsou hledané bodové odhady parametrů.

Odhad výstupů

Odhad výstupů se v diskrétním případě provádí dvěma způsoby. Jednak odhadneme pro daný případ nejpravděpodobnější hodnotu, tedy na příslušném řádku tabulky najdeme největší pravděpodobnost, jednak můžeme spočíst střední hodnotu dle vzorce: $\bar{y}_n = 1 \cdot P(1) + 2 \cdot P(2) + \ldots$ Tato střední hodnota nám při dvou možnostech výstupu ukazuje, jak moc je která možnost pravděpodobná.

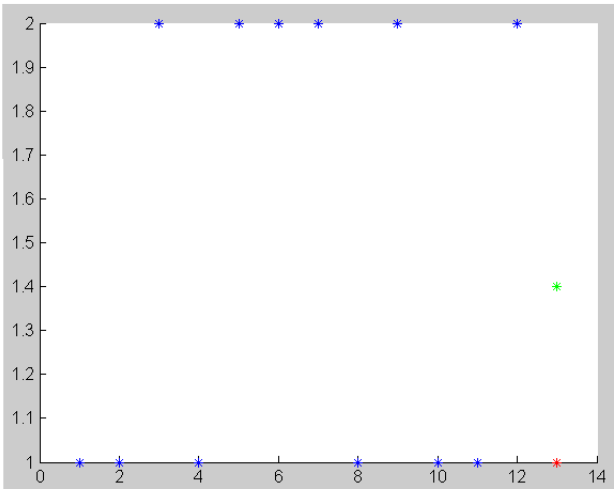
Příklad

Odhadněte další hodnotu ve výše uvedené řadě. Udejte nejpravděpodobnější hodnotu i střední hodnotu výstupu. Vykreslete graf s daty i odhadem.

Poslední v řadě je dvojka. Aktuální tedy bude druhý řádek tabulky (viz řádek 19). Na tomto řádku má největší pravděpodobnost jednička (viz řádek 20). Jednička tedy bude nejpravděpodobnějším výstupem. Střední hodnota je: $\mu = \frac{3}{5} \cdot 1 + \frac{2}{5} \cdot 2 = \frac{7}{5}$, což je hodnota blíže k jedničce než ke dvojce.

Nyní v Matlabu:

```
1   clear all;
2   clc;
3
4   y=[1,1,2,1,2,2,2,1,2,1,1,2];
5
6   Tab=zeros(2,2);
7
8   %Odhad parametru
9   for n=2:12
10      i=y(n-1);    %Urceni radku
11      j=y(n);      %Urceni sloupce
12      Tab(i,j)=Tab(i,j)+1;
13  end;
14  for Radek=1:2
15      Tab(Radek,:)=Tab(Radek,:)/sum(Tab(Radek,:));
16  end;
17
18  %Odhad vystupu
19  i=y(n); %Ktery radek je aktualni
20  [M,j]=max(Tab(i,:));  %j je nejpravd. hodn.
21  jj=1*Tab(i,1)+2*Tab(i,2);
22
23  hold on;
24  plot(y,'b*');
25  plot(13,j,'r*');
26  plot(13,jj,'g*');
27  hold off;
28
```



Vidíme, že nejpravděpodobnější hodnota je 1 (červeně). Střední hodnota (zeleně) nám ukazuje, že pravděpodobnost je ve prospěch jedničky vychýlena jen lehce.

Logistický model

Zde musíme hledat maximum sdruženého rozdělení numericky. Odhad výstupů je pak stejný jako u diskrétního modelu.

Logistický model zde podrobněji řešit nebudeme.