1 Vyrovnávání dat

Naše pozorování jsou dána tabulkou čísel $\begin{array}{|c|c|c|c|}\hline x_1 & x_2 & \cdots & x_n \\ \hline y_1 & y_2 & \cdots & y_n \\ \hline \end{array}$, kde x_i často bývají časové údaje, a my chceme data proložit nějakou hladkou funkcí, která by vystihovala hlavní vlastnosti dat, ale ignorovala malé fluktuace a nepřesnosti.

1.1 Metoda nejmenších čtverců

Zde volíme tvar hladké funkce předem, například jako přímku $x \mapsto a + bx$ nebo parabolu $x \mapsto a + bx + cx^2$. Parametry křivky $a, b, c, \ldots$ určíme jako ty, které minimalizují součet čtverců odchylek mezi křivkou a daty.

Předpokládejme tedy, že chceme data proložit funkcí tvaru

$$x \mapsto a_1 f_1(x) + \cdots + a_k f_k(x), \quad k \leq n,$$

kde $a_1, \ldots a_n$ jsou neznámé koeficienty, které potřebujeme nalézt. Vyrovnaná data budeme značit

$$\hat{y}_i = a_1 f_1(x_i) + \cdots + a_k f_k(x_i), \quad i = 1, \cdots, n.$$

Pokud použijeme vektorový zápis $y := (y_1, \cdots, y_n)^T \in \mathbb{R}^n$, resp. $\hat{y} := (\hat{y}_1, \cdots, \hat{y}_n)^T \in \mathbb{R}^n$, můžeme definiční rovnost pro vyrovnaná data přepsat do maticového tvaru

$$\hat{y} = Fa, \quad \text{kde} \quad F = (f_j(x_i))_{i,j=1}^{n,k}, \quad \text{a kde} \quad a = (a_1, \cdots, a_k)^T \in \mathbb{R}^k.$$

Rozepsáno po složkách:

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} f_1(x_1) & \cdots & f_k(x_1) \\ \vdots & \cdots & \vdots \\ f_1(x_n) & \cdots & f_k(x_n) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix}.$$

Koeficienty a budeme hledat jako ty, které minimalizují reziduální součet čtverců $S(a) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Ten je možné vyjádřit ve tvaru

$$S(a) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} a_j f_j(x_i)\right)^2 = (y - \hat{y})^T(y - \hat{y}) = (y - Fa)^T(y - Fa).$$

Předpokládejme pro jednoduchost, že matice F má plnou hodnost, tj. $h(F) = k$. Potom po zderivování $S(a)$ podle a dostaneme

$$\frac{\partial}{\partial a} S(a) = -2(F^T y - F^T F a),$$

a položíme-li derivaci rovnou 0, zjistíme, že $\hat{a}$ řeší soustavu k lineárních rovnic o k neznámých

$$F^T F a = F^T y.$$

Této soustavě říkáme soustava normálních rovnic. Řešení $\hat{a}$ musí být bodem globálního minima funkce $S(a)$, protože ta je konvexní v a.

Protože F má plnou hodnost, matice F^TF (čtvercová $k \times k$) má také plnou hodnost a existuje její inverze. Proto

$$\hat{a} = (F^TF)^{-1}F^Ty, \tag{1}$$

a vyrovnaná data obdržíme ze vztahu

$$\hat{y} = F\hat{a} = Hy, \quad \text{kde} \quad H := F(F^TF)^{-1}F^T.$$

Matici H se říká projekční matice.

Poznámka Obecně není nutné předpokládat, že x_i jsou čísla, mohou to být např. dvojice či l-tice čísel.

Příklad: Prokládání přímkou $x \mapsto a_1 + a_2x$

V tomto případě $f_1(x) = 1, f_2(x) = x$

$$F = (\mathbf{1}, x) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \text{ kde } \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ a kde } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Pak

$$F^TF = \begin{pmatrix} \mathbf{1}^T \\ x^T \end{pmatrix}(\mathbf{1}, x) = \begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^Tx \\ x^T\mathbf{1} & x^Tx \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Příslušná inverze je pak tvaru

$$(F^TF)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - n^2\overline{x}^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Matice F^Ty je tvaru

$$\begin{pmatrix} \mathbf{1}^T \\ x^T \end{pmatrix} y = \begin{pmatrix} \mathbf{1}^Ty \\ x^Ty \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{pmatrix}.$$

Příslušné koeficienty $\hat{a}_1, \hat{a}_2$ pak získáme

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \frac{1}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \begin{pmatrix} \overline{y}\sum_{i=1}^n x_i^2 - \overline{x}\sum_{i=1}^n x_iy_i \\ \sum_{i=1}^n x_iy_i - n\overline{x}\,\overline{y} \end{pmatrix}$$

Odtud plyne

$$\hat{a}_2 = \frac{\sum_{i=1}^n x_iy_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2},$$

a

$$\hat{a}_1 = \frac{\overline{y}\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right) - \overline{x}\left(\sum_{i=1}^n x_iy_i - n\overline{x}\,\overline{y}\right)}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \overline{y} - \hat{a}_2\overline{x}.$$

□

Pravděpodobnostní interpretace: regresní model

Předpokládáme, že posloupnost $y := (y_1, \cdots, y_n)^T$ je realizací náhodného vektoru $Y := (Y_1, \cdots, Y_n)^T$ vyhovující lineárnímu modelu

$$Y = Fa + e, \tag{2}$$

kde náhodný vektor $e = (e_1, \cdots, e_n)^T$ má složky s nulovou střední hodnotou a var $e = \sigma^2 I$, kde I je $n \times n$-rozměrná jednotková matice a $\sigma^2 > 0$.

Poznámka: Zápis (2) říká, že vektor Y má střední hodnotu Fa a varianční matici $\sigma^2 I$. Právě tato lineární závislost[1] střední hodnoty $EY = (EY_1, \cdots, EY_n)^T$ na vektoru parametrů $a = (a_1, \cdots, a_k)^T$ je důvodem, proč se modelu říká lineární.

Ze základní přednášky z matematické statistiky víme, že za předpokladů regresního modelu mají odhadnuté koeficienty a data vyrovnaná metodou nejmenších čtverců některé pěkné vlastnosti. Platí například následující věta.

Věta 1 *Odhad* (1) *parametru a v regresním modelu je nestranný, jeho rozptyl je roven $\sigma^2(F^TF)^{-1}$. Jsou-li navíc (y_i, e_i) nezávislé a stejně rozdělené, pak $\hat{a}$ je konzistentní odhad vektoru a.*

Spliny – po částech parametrické vyrovnávání

Zatím jsme vyrovnávali data pomocí předem dané křivky, závislé na několika málo parametrech (např. parabola). Nyní budeme vyrovnávat pomocí křivek, které jsou flexibilnější. V jednotlivých předem zvolených intervalech je křivka definovaná různě, ale jako celek si stále zachovává určitou hladkost, vyjádřenou pomocí spojitosti derivací až do určeného řádu.

Definice 1 *Buď dána posloupnost bodů $\{u_j\}_{j=1}^m$, ty nazveme uzly. Splinem řádu k nazveme funkci f, která je v každém intervalu $[u_j, u_{j+1}]$ polynomem stupně k a která má v celém definičním oboru spojité derivace až do řádu $k-1$ včetně.*

Po částech parametrické vyrovnávání si ukážeme na případu kubického splinu (splinu řádu 3). Zaměříme se na případ dvoubloukového splinu.

Nejprve zvolíme index $k \in \{1, \cdots, n\}$ a tím i odpovídající uzel x_k. Vyrovnané hodnoty $\hat{y}_i$ budeme definovat jako $f(x_i)$ pro $i \in \{1, \cdots, n\}$, kde

$$\begin{aligned} f(x) &= c_0 + c_1 x + c_2 x^2 + c_3 x^3, && \text{pokud } x \le x_k \\ &= c_0 + c_1 x + c_2 x^2 + c_3 x^3 + d(x - x_k)^3, && \text{pokud } x > x_k. \end{aligned}$$

Funkce f má tím pádem spojitou druhou derivaci (i v bodě x_k – ověření přenecháváme čtenáři). Hodnoty parametrů $c_0, \cdots, c_3, d$ volíme tak, abychom minimalizovali hodnotu

$$\sum_{i=1}^{k} \left(y_i - c_0 - c_1 x_i - c_2 x_i^2 - c_3 x_i^3\right)^2 + \sum_{i=k+1}^{n} \left(y_i - c_0 - c_1 x_i - c_2 x_i^2 - c_3 x_i^3 - d(x_i - x_k)^3\right)^2.$$

[1] Jiná lineární závislost na vektoru $a = (a_1, \cdots, a_k)^T$ než závislost tvaru $a \mapsto Fa$ kde F matice s k sloupci neexistuje (pokud má být výsledkem této závislosti konečně rozměrný vektor čísel).

Zřejmě jde o metodu nejmenších čtverců s maticí

$$F = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{k+1} & x_{k+1}^2 & x_{k+1}^3 & (x_{k+1}-x_k)^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n-x_k)^3 \end{pmatrix}.$$

Zbývá tedy řešit soustavu normálních rovnic

$$\left(\begin{array}{ccccc|c} n & \sum x_i & \sum x_i^2 & \sum x_i^3 & \sum (x_i-x_k)_+^3 & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \sum x_i(x_i-x_k)_+^3 & \sum x_i y_i \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \sum x_i^5 & \sum x_i^2(x_i-x_k)_+^3 & \sum x_i^2 y_i \\ \sum x_i^3 & \sum x_i^4 & \sum x_i^5 & \sum x_i^6 & \sum x_i^3(x_i-x_k)_+^3 & \sum x_i^3 y_i \\ \sum (x_i-x_k)_+^3 & \sum x_i(x_i-x_k)_+^3 & \sum x_i^2(x_i-x_k)_+^3 & \sum x_i^3(x_i-x_k)_+^3 & \sum ((x_i-x_k)_+)^6 & \sum (x_i-x_k)_+^3 y_i \end{array}\right),$$

kde $(x_i - x_k)_+ = I_{(x_i-x_k)>0}(x_i - x_k)$ značí kladnou část čísla $(x_i - x_k)$.

Pro obecný p-obloukový kubický spline postupujeme obdobně, volíme $p-1$ uzlů a $p+3$ parametrů $c_0, c_1, c_2, c_3, d_1, \dots d_{p-1}$.

1.2 Klouzavé průměry

Klouzavé průměry (KP) jsou schopny postihnout trend v datech, tedy směr a míru pohybu pozorovaných hodnot, bez toho, že bychom měli pro trend nějaký specifický model. Data vyrovnávají pouze lokálně, tj. v daném bodě se vyrovnaná hodnota počítá pouze z několika okolních hodnot, nikoli z celé pozorované řady.

Na základě dat $y = (y_1, \cdots, y_n)^T$ získáme vyrovnané hodnoty $\hat{y}_i$ předpisem

$$\hat{y}_i := \sum_{j:=-r}^{r} a_j y_{i+j}, \text{ pro } i = r+1, \cdots, n-r,$$

kde váhy $(a_{-r}, \cdots, a_r)$ splňují $\sum_{j=-r}^{r} a_j = 1$. Číslo $2r+1$ nazveme délkou klouzavého průměru.

Klouzavé aritmetické průměry

Všechny váhy $a_{-r}, \cdots, a_r$ jsou stejné a jsou rovny hodnotě $\frac{1}{2r+1}$.

Vyrovnaná hodnota je tedy opravdu jen prostý průměr z $2r+1$ okolních hodnot. Výhodou klouzavých aritmetických průměrů je, že všechny váhy jsou nezáporné, nevýhodou, že není vyrovnán počáteční a koncový úsek dat.

Konstrukce KP vyrovnáváním úseků polynomy

Váhy klouzavých průměrů volíme tak, že aproximujeme $2r+1$ členů řady $y_{t-r}, \dots, y_t, \dots y_{t+r}$ vhodným polynomem (řádu k) a hodnotu tohoto polynomu v bodě t použijeme jako vyrovnanou hodnotu $\hat{y}_t$. Číslo k nazýváme řád klouzavého průměru.

K tomu, abychom dospěli k příslušným vahám pro klouzavé průměry délky $2r+1$ a řádu k, budeme uvažovat model

$$y_{t+u} = c_0 + c_1 u + c_2 u^2 + \cdots + c_k u^k \qquad \text{pro} \qquad u = -r, \cdots, r.$$

Maticově to lze zapsat ve tvaru

$$\mathbf{y}_t = \begin{pmatrix} y_{t-r} \\ \vdots \\ y_{t+r} \end{pmatrix} = \begin{pmatrix} 1 & -r & (-r)^2 & \cdots & (-r)^k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & r & r^2 & \cdots & r^k \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_k \end{pmatrix}.$$

Skutečné vyrovnané hodnoty pak obdržíme dosazením $u := 0$,

$$\hat{y}_t = c_0 + c_1 u + c_2 u^2 + \cdots + c_k u^k|_{u=0} = c_0. \tag{3}$$

Parametry $c_0, \cdots, c_k$ získáme metodou nejmenších čtverců.

Příklad: Kubický polynom ($k = 3$) délky 5 ($r = 2$). Uvažovaný model je tvaru

$$y_{t+u} = c_0 + c_1 u + c_2 u^2 + c_3 u^3.$$

K volbě parametrů c_0, c_1, c_2, c_3 použijeme metodu nejmenších čtverců. Nejprve sestavíme matici

$$F = \begin{pmatrix} 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{pmatrix}.$$

Parametr $c = (c_0, \cdots, c_3)^T$ dostaneme jako řešení soustavy normálních rovnic

$$F^T F c = F^T \mathbf{y}_t, \text{ kde } \mathbf{y}_t = (y_{t-2}, y_{t-1}, y_t, y_{t+1}, y_{t+2})^T$$

pro $t = 3, \cdots, n-2$. Projekční matici H obdržíme pomocí vzorce

$$H = F(F^T F)^{-1} F^T = \begin{pmatrix} \frac{69}{70} & \frac{2}{35} & \frac{-3}{35} & \frac{2}{35} & \frac{-1}{70} \\ \frac{2}{35} & \frac{27}{35} & \frac{12}{35} & \frac{-8}{35} & \frac{2}{35} \\ \frac{-3}{35} & \frac{12}{35} & \frac{17}{35} & \frac{12}{35} & \frac{-3}{35} \\ \frac{2}{35} & \frac{-8}{35} & \frac{12}{35} & \frac{27}{35} & \frac{2}{35} \\ \frac{-1}{70} & \frac{2}{35} & \frac{-3}{35} & \frac{2}{35} & \frac{69}{70} \end{pmatrix}. \tag{4}$$

Vyrovnaná hodnota $\hat{y}_t = c_0$ je rovna součinu prostředního řádku H a sloupcového vektoru $\mathbf{y}_t$,

$$\hat{y}_t = \frac{1}{35}(-3, 12, 17, 12, -3)\mathbf{y}_t, \quad \text{pro} \quad t = 3, \cdots, n-2.$$

□

Pro klouzavé průměry zkonstruované vyrovnáváním úseků polynomy obecně platí, že

– součet vah je roven 1,

– váhy jsou symetrické kolem prostřední hodnoty,

– pro sudé k je klouzavý průměr řádu k a $k+1$ stejný.

Počáteční a koncové úseky řady (tedy $t \leq r$ a $t \geq n - r + 1$) nemůžeme proložit pomocí vzorce (3), protože nemáme k dispozici odpovídající $\mathbf{y}_t$. Použijeme proto prokládání „prvním a posledním možným polynomem“,

$$\hat{y}_{1+r+u} = c_0 + c_1 u + c_2 u^2 + \cdots + c_k u^k, \quad u = -r, \cdots, r,$$

a

$$\hat{y}_{n-r+u} = c_0 + c_1 u + c_2 u^2 + \cdots + c_k u^k, \quad u = -r, \cdots, r. \tag{5}$$

Koeficienty spočítáme opět metodou nejmenších čtverců.

Dosazením $u = r + 1$ do rovnice (5) s odhadnutými koeficienty dostaneme předpověď na jeden krok dopředu. Tento postup lze ovšem použít jen pro krátkodobé předpovědi.

Poznámka: Počáteční, koncové a předpovědní klouzavé průměry se liší pro řády k a $k + 1$.

Příklad - pokračování: Hodnoty y_{n-1}, y_n lze vyrovnat prokládáním polynomu

$$\hat{y}_{n-2+u} = c_0 + c_1 u + c_2 u^2 + c_3 u^3, \quad u = -2, -1, 0, 1, 2.$$

Uvědomme si, že metodou nejmenších čtverců dojdeme ke stejné projekční matici H jako v (4) a tedy

$$\hat{y}_{n-1} = \frac{1}{35}(2, -8, 12, 27, 2)\mathbf{y}_{n-2}, \quad \hat{y}_n = \frac{1}{70}(-1, 4, -6, 4, 69)\mathbf{y}_{n-2}.$$

Obdobně dojdeme ke vzorcům pro začátek řady

$$\hat{y}_2 = \frac{1}{35}(2, 27, 12, -8, 2)\mathbf{y}_3, \quad \hat{y}_1 = \frac{1}{70}(69, 4, -6, 4, -1)\mathbf{y}_3.$$

Předpověď o jeden krok dopředu získáme použitím hodnoty $u = 3$. Příslušné koeficienty lze získat následujícím způsobem

$$(1, 3, 3^2, 3^3)(F^T F)^{-1} F^T = \left(-\frac{4}{5}, \frac{11}{5}, -\frac{4}{5}, -\frac{14}{5}, \frac{16}{5}\right).$$

Předpověď o jeden krok dopředu pak vychází ve tvaru

$$\hat{y}_{n+1} = \frac{1}{5}(-4, 11, -4, -14, 16)\mathbf{y}_{n-2}.$$

□

1.3 Whittacker-Hendersonova metoda

Při vyhlazování jsou většinou v protikladu dva požadavky:

– dosáhnout dobré shody mezi původními a vyrovnanými daty (tedy co nejmenší průměrná kvadratická odchylka vyrovnaných hodnot $\hat{y}_i$ od původních hodnot y_i),

– vyrovnaná data by měla být co nejvíce „hladká“ (tedy co nejmenší variace vyrovnaných hodnot).

Whittacker-Hendersonova metoda umožňuje hledat kompromis mezi těmito dvěma požadavky tím, že jim přiděluje rozdílnou váhu.

Vyrovnané hodnoty určíme tak, aby minimalizovaly hodnotu

$$M = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + h \sum_{i=r+1}^{n} (\Delta^r \hat{y}_i)^2, \tag{6}$$

kde $h > 0$ je parametr, který neodhadujeme, ale volíme. Symbol

$$\Delta^r \hat{y}_i = \Delta^{r-1} \hat{y}_i - \Delta^{r-1} \hat{y}_{i-1}$$

označuje rekurzivně definovanou r-tou zpětnou diferenci posloupnosti $\hat{y}_i$, kde

$$\Delta^0 \hat{y}_i = \hat{y}_i \qquad \text{a} \qquad \Delta^1 \hat{y}_i = \Delta \hat{y}_i = \hat{y}_i - \hat{y}_{i-1}.$$

Tuto zpětnou diferenci lze zapsat pomocí binomické formule

$$\Delta^r \hat{y}_i = \begin{pmatrix} r \\ 0 \end{pmatrix} \hat{y}_i - \begin{pmatrix} r \\ 1 \end{pmatrix} \hat{y}_{i-1} + \cdots + (-1)^r \begin{pmatrix} r \\ r \end{pmatrix} \hat{y}_{i-r}.$$

První člen ve vzorci (6) tedy měří shodu mezi y_i a $\hat{y}_i$ a druhý člen hladkost vyrovnání pomocí diferencí r-tého řádu (r se většinou volí 2, 3 nebo 4). Parametr $h > 0$ volíme podle toho, jestli klademe větší důraz na shodu nebo na hladkost.

Abychom byli schopni vyjádřit hodnotu M maticově, definujeme matici K rozměru $(n-r) \times n$,

$$K = \begin{pmatrix} (-1)^r & \cdots & -\begin{pmatrix} r \\ 1 \end{pmatrix} & 1 & \cdots & 0 \\ \vdots & (-1)^r & \cdots & -\begin{pmatrix} r \\ 1 \end{pmatrix} & \cdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & (-1)^r & \cdots & -\begin{pmatrix} r \\ 1 \end{pmatrix} & 1 \end{pmatrix}.$$

Pak lze psát

$$M = (\hat{y} - y)^T (\hat{y} - y) + h \cdot \hat{y}^T K^T K \hat{y}.$$

Matice druhých derivací M podle $\hat{y}_i$ je $2(I + hK^T K)$, což je pozitivně definitní matice, tedy bod, ve kterém $\nabla_{\hat{y}} M(\hat{y}) = 0$, je bod lokálního minima funkce M. Derivováním podle $\hat{y}$ dostaneme

$$\left(\frac{\partial M}{\partial \hat{y}_1}, \ldots, \frac{\partial M}{\partial \hat{y}_n} \right) = \nabla_{\hat{y}} M(\hat{y}) = 2\hat{y} - 2y + 2hK^T K \hat{y}.$$

Bod $\hat{y} = (I + hK^T K)^{-1} y$ je tak bodem minima M na $\mathbb{R}^n$.