

Wstęp do metod numerycznych

Aproksymacja i zagadnienie najmniejszych kwadratów

P. F. Góra

<http://th-www.if.uj.edu.pl/zfs/gora/>

2011/12

Aproksymacja

Termin *aproksymacja* występuje w dwu znaczeniach:

Aproksymacja punktowa: Mając N punktów, staramy się znaleźć funkcję *należącą do znanej kategorii*, która będzie przebiegać możliwie “najbliżej” tych punktów. Podkreślam, że funkcja jest **znana** co do swego kształtu (np. wielomian ustalonego stopnia, kombinacja funkcji trygonometrycznych, funkcja opisująca jakiś rozkład prawdopodobieństwa itp), a tylko nieznane są jej parametry.

Aproksymacja ciągła: Mając ustaloną funkcję $g(x)$, której sposób obliczania jest trudny, skonstruować inną funkcję, która będzie w pewnym sensie bliska funkcji wyjściowej, a jednocześnie obliczeniowo prostsza.

I. Aproksymacja punktowa

Interpolacja punktowa najczęściej kojarzy się z **dopasowaniem funkcji do danych doświadczalnych**. Mamy N par punktów $\{(x_i, y_i)\}_{i=1}^N$, gdzie x_i jest dokładną wartością argumentu, y_i zmierzoną (lub obliczoną na jakimś wcześniejszym etapie) wartością funkcji. Skrajnym przypadkiem aproksymacji punktowej jest interpolacja — funkcja przechodzi przez *wszystkie* punkty doświadczalne, ale jest “trudną” funkcją: wielomianem wysokiego stopnia, funkcją sklejaną, skomplikowaną funkcją wymierną, my tymczasem chcemy mieć jakąś “prostą” funkcję, przechodzącą dostatecznie blisko wszystkich punktów.

Z teorii możemy wiedzieć, że zależność pomiędzy x a y *powinna mieć* charakter $y = y(x)$, jednak zmierzone (lub obliczone) wartości nie odpowiadają dokładnie wartościom teoretycznym, gdyż są obarczone błędami pomiarowymi (obliczeniowymi).

Liniowe zagadnienie najmniejszych kwadratów

Każdej zmierzonej (i obarczonej błędem) wartości y_i odpowiada wartość teoretyczna \tilde{y}_i , jaką zmienna y “powinna” przybrać dla danej wartości zmiennej x . **Przyjmujemy, że wartość teoretyczna jest kombinacją liniową pewnych znanych funkcji:**

$$\tilde{y}_i = a_1 \cdot f_1(x_i) + a_2 \cdot f_2(x_i) + \dots + a_s \cdot f_s(x_i) \quad (1)$$

Zespół *wszystkich* wartości teoretycznych możemy zatem przedstawić jako

$$\tilde{\mathbf{y}} = \mathbf{A}\mathbf{p}, \quad (2a)$$

gdzie

$$\mathbf{A} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & f_3(x_1) & \cdots & f_s(x_1) \\ f_1(x_2) & f_2(x_2) & f_3(x_2) & \cdots & f_s(x_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f_1(x_n) & f_2(x_n) & f_3(x_n) & \cdots & f_s(x_n) \end{bmatrix} \in \mathbb{R}^{n \times s}, \quad (2b)$$

$$\mathbf{p} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_s \end{bmatrix} \in \mathbb{R}^s. \quad (2c)$$

Przykład

Do n punktów pomiarowych $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ dopasowujemy wielomian drugiego stopnia $\tilde{y} = ax^2 + bx + c$. Wartości teoretyczne możemy zapisać jako

$$\tilde{\mathbf{y}} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \dots & \dots & \dots \\ x_n^2 & x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \quad (3)$$

Zauważmy, że dopasowywanie do danych wielomianu ustalonego stopnia (nie tylko linii prostej!) **jest zagadnieniem liniowym!**

Błędy pomiarowe

Różnica pomiędzy wartością zmierzona y_i a wartością teoretyczną \tilde{y}_i jest spowodowana błędem pomiarowym: $y_i - \tilde{y}_i = \xi_i$. Przyjmujemy, że liczby ξ_i , $\langle \xi_i \rangle = 0$, są liczbami losowymi, pochodzącymi z rozkładu normalnego (Gausa). Oznaczmy wektor wszystkich błędów pomiarowych przez $\xi = [\xi_1, \xi_2, \dots, \xi_n]^T \in \mathbb{R}^n$. Dalej, przyjmijmy, że łącznie wszystkie błędy tworzą n -wymiarowy rozkład Gausa o macierzy kowariancji G :

$$\langle \xi \xi^T \rangle = G, \quad (4)$$

gdzie $\langle \dots \rangle$ oznacza średniowanie po realizacjach zmiennych losowych. Macierz G jest symetryczna i dodatnio określona.

Metoda najmniejszych kwadratów

Twierdzenie 1. *Jeżeli błędy pomiarowe pochodzą z rozkładu Gaussa o macierzy kowariancji G , estymator największej wiarygodności odpowiada minimum formy kwadratowej*

$$Q = \frac{1}{2} \xi^T G^{-1} \xi. \quad (5)$$

Zauważmy, że ponieważ G jest symetryczna i dodatnio określona, także G^{-1} jest symetryczna i dodatnio określona, a zatem forma kwadratowa (5) z całą pewnością posiada minimum.

Obecność *odwrotności* macierzy kowariancji w wyrażeniu (5) oznacza, że pomiary obarczone *większym* błędem dają *mniej* wkład do Q .

Forma kwadratowa estymatorów

$$\begin{aligned} Q &= \frac{1}{2} \boldsymbol{\xi}^T \mathbf{G}^{-1} \boldsymbol{\xi} = \frac{1}{2} (\mathbf{y} - \tilde{\mathbf{y}})^T \mathbf{G}^{-1} (\mathbf{y} - \tilde{\mathbf{y}}) \\ &= \frac{1}{2} (\mathbf{y} - \mathbf{A}\mathbf{p})^T \mathbf{G}^{-1} (\mathbf{y} - \mathbf{A}\mathbf{p}) \\ &= \frac{1}{2} \left[\mathbf{y}^T \mathbf{G}^{-1} \mathbf{y} - (\mathbf{A}\mathbf{p})^T \mathbf{G}^{-1} \mathbf{y} - \mathbf{y}^T \mathbf{G}^{-1} \mathbf{A}\mathbf{p} + (\mathbf{A}\mathbf{p})^T \mathbf{G}^{-1} \mathbf{A}\mathbf{p} \right] \\ &= \frac{1}{2} \mathbf{p}^T \mathbf{A}^T \mathbf{G}^{-1} \mathbf{A}\mathbf{p} - \mathbf{p}^T \mathbf{A}^T \mathbf{G}^{-1} \mathbf{y} + \underbrace{\frac{1}{2} \mathbf{y}^T \mathbf{G}^{-1} \mathbf{y}}_{=\text{const}} \end{aligned} \quad (6)$$

W **liniowym** zagadnieniu najmniejszych kwadratów minimalizowana funkcja jest *formą kwadratową w parametrach*. Dzięki temu wiemy, że minimum istnieje i jest jednoznaczne. *Liniowość* oznacza tutaj, że funkcja “teoretyczna” zależy *liniowo* od parametrów, nie od argumentu!

Minimum formy kwadratowej

Aby znaleźć estymator, należy znaleźć taki wektor \mathbf{p} , że forma kwadratowa (6) przybiera najmniejszą możliwą wartość. Można to zrobić albo bezpośrednio, metodą zmiennej metryki lub gradientów sprzężonych, albo formalnie rozwiązując równanie $\nabla Q = 0$, gdzie różniczkujemy po składowych wektora \mathbf{p} . Otrzymujemy

$$\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A} \mathbf{p} = \mathbf{A}^T \mathbf{G}^{-1} \mathbf{y} \quad (7)$$

Tego równanie nie można “uprościć” pozbywając się członu $\mathbf{A}^T \mathbf{G}^{-1}$, bo jest to macierz niekwadratowa, dla której nie da się zdefiniować odwrotności. Natomiast samo równanie (7) jest dobrze określone, gdyż macierz tego równania $\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A}$ jest symetryczna i dodatnio określona.

Własności wektora estymatorów

Wektor \mathbf{p} obliczamy z (7) dla takich wartości pomiarów, jakie faktycznie mamy. Tak obliczony wektor \mathbf{p} jest wektorem estymatorów. Pamiętajmy, że pomiary są obarczone błędami losowymi, a więc także obliczone estymatory są, formalnie, *gaussowskimi liczbami losowymi*. Co można powiedzieć o tych liczbach? $\mathbf{y} = \mathbf{A}\mathbf{p}^* + \boldsymbol{\xi}$, gdzie \mathbf{p}^* jest zbudowany z “prawdziwych”, nieznanych wartości $[a_1, \dots, a_s]^T$ z równania (1). Widać, że

$$\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A} (\mathbf{p} - \mathbf{p}^*) = \mathbf{A}^T \mathbf{G}^{-1} \boldsymbol{\xi} \quad (8)$$

wobec czego

$$\langle \mathbf{p} \rangle = \mathbf{p}^*, \quad (9)$$

gdyż $\langle \boldsymbol{\xi} \rangle = 0$. Obliczone estymatory są przybliżeniem “prawdziwych” wartości parametrów w sensie równania (9).

Macierz kowariancji estymatorów wynosi

$$\begin{aligned} \mathbf{C}_p &= \langle (\mathbf{p} - \mathbf{p}^*)(\mathbf{p} - \mathbf{p}^*)^T \rangle \\ &= (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{G}^{-1} \langle \boldsymbol{\xi} \boldsymbol{\xi}^T \rangle \mathbf{G}^{-1} \mathbf{A} (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1}, \end{aligned} \quad (10)$$

gdzie skorzystaliśmy z symetrii macierzy \mathbf{G}^{-1} i macierzy $\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A}$. Korzystając z równania (4) otrzymujemy

$$\begin{aligned} \mathbf{C}_p &= (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{G}^{-1} \overbrace{\langle \mathbf{G} \mathbf{G}^{-1} \rangle}^{\mathbb{I}} \mathbf{A} (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1}, \\ &= (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1} \underbrace{\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A} (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1}}_{\mathbb{I}} \\ &= (\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A})^{-1}. \end{aligned} \quad (11)$$

Nadokreślony układ równań

Zamiast minimalizować formę kwadratową (5), moglibyśmy *zażądać*, aby równanie $y_i = \tilde{y}_i$ było *ściśle* spełnione dla wszystkich punktów pomiarowych (x_i, y_i) . Wobec równania (2a) oznacza to, że chcemy rozwiązać układ równań liniowych

$$\mathbf{A}\mathbf{p} = \mathbf{y}. \quad (12)$$

Jest to *nadokreślony* układ równań (s niewiadomych i n , $n > s$, równań) i, poza wyjątkowymi przypadkami, nie ma on ścisłego rozwiązania. Jak jednak wiemy z poprzednich wykładów, metoda *SVD* (*Singular Value Decomposition*) dostarcza przybliżonego rozwiązania takich układów, optymalnego w sensie najmniejszych kwadratów. Jeżeli macierz kowariancji \mathbf{G} jest proporcjonalna do macierzy jednostkowej, $\mathbf{G} = \sigma^2 \mathbb{I}$, co odpowiada pomiarom nieskorelowanym i obarczonym takimi samymi błędami i w praktyce zdarza się bardzo często, przybliżone rozwiązanie (12) uzyskane za pomocą *SVD* jest (w arytmetyce dokładnej) *tym samym* rozwiązaniem, które otrzymalibyśmy minimalizując formę kwadratową (6) lub rozwiązując układ równań (7).

Gdybyśmy, zamiast (12), zażądali spełnienia układu równań

$$\mathbf{G}^{-1}\mathbf{A}\mathbf{p} = \mathbf{G}^{-1}\mathbf{y}, \quad (13)$$

również nadokreślonego, ale uwzględniającego różne wagi poszczególnych pomiarów, rozwiązanie optymalne w sensie *SVD* byłoby równoważne rozwiązaniu równania (7).

Pomiary nieskorelowane

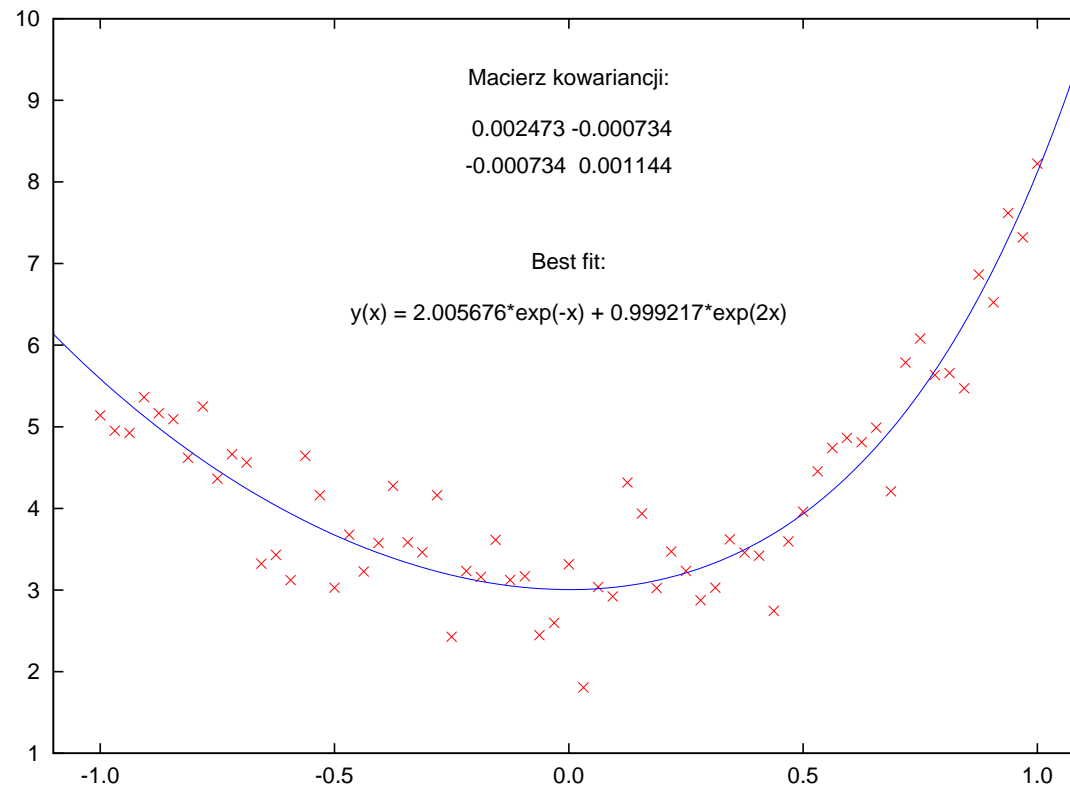
Na ogół (i na ogół z dobrym uzasadnieniem) zakłada się, że pomiary są niezależne, a ich wyniki nieskorelowane. Wówczas elementy pozadiagonalne macierzy G znikają, $G = \text{diag}\{\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2\}$. Minimalizowana forma kwadratowa (5) upraszcza się do

$$Q = \sum_{i=1}^n \frac{(a_1 f_1(x_i) + a_2 f_2(x_i) + \dots + a_s f_s(x_i) - y_i)^2}{\sigma_i^2}. \quad (14)$$

W dość częstym przypadku pomiarów nieskorelowanych i identycznych $\forall i = 1, \dots, n: \sigma_i^2 = \sigma^2$, a zatem $G = \sigma^2 \mathbb{I}$. W tym wypadku estymatory nie zależą od macierzy kowariancji pomiarów, gdyż macierz G wypada z równania (7), natomiast macierz kowariancji estymatorów upraszcza się do

$$C_p = \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1}. \quad (15)$$

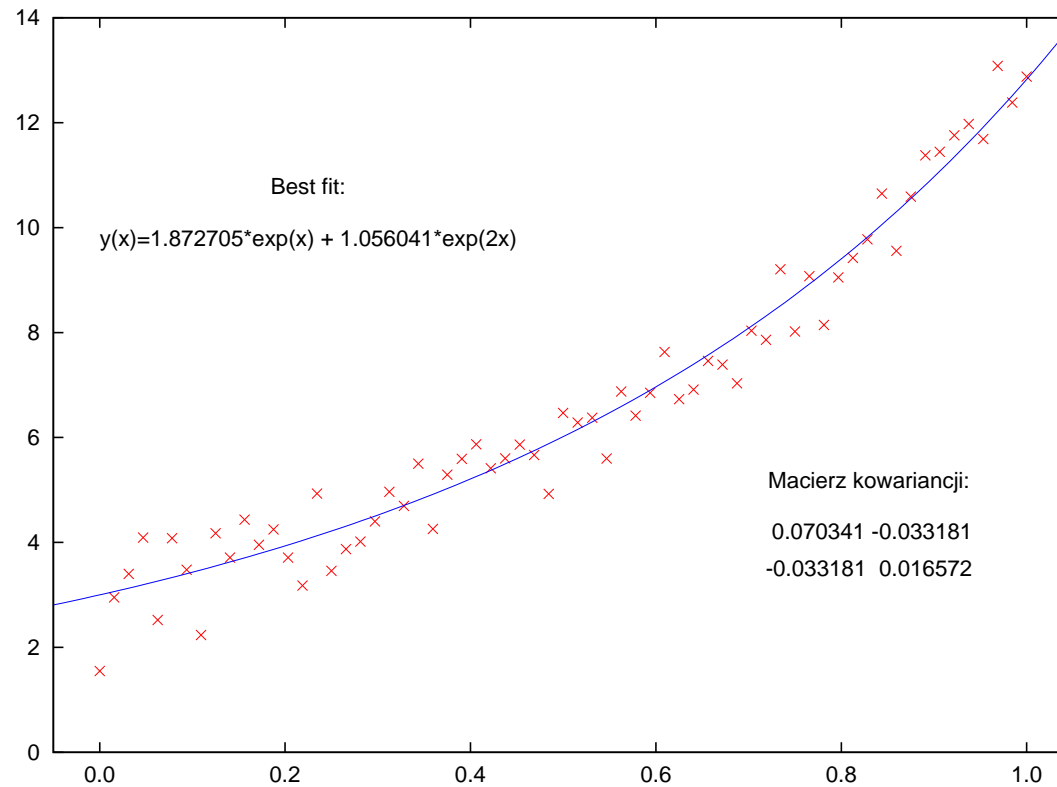
Przykład



Do zaznaczonych punktów dopasowano krzywą $y = ae^{-x} + be^{2x}$ za pomocą *liniowej* metody najmniejszych kwadratów. Przyjęto, że pomiary są identyczne i nieskorelowane, o stałym błędzie $\sigma^2 = 0.305226$.

W powyższym przykładzie współczynniki korelacji estymatorów są bardzo małe, gdyż, efektywnie, obie funkcje dopasowują się do innych zakresów danych (funkcja e^{2x} jest mała dla $x < 0$, funkcja e^{-x} jest mała dla $x > 0$). Jeżeli różne funkcje bazowe “konkurują” o te same dane, współczynnik korelacji jest, co do wartości bezwzględnej, większy. Ujemny współczynnik korelacji pomiędzy estymatorami oznacza, że *prawie* tak samo dobre dopasowanie można uzyskać zmniejszając jeden, zwiększając zaś drugi.

Przykład



Do zaznaczonych punktów dopasowano krzywą $y = ae^x + be^{2x}$ za pomocą *liniowej* metody najmniejszych kwadratów. Przyjęto, że pomiary są identyczne i nieskorelowane, o stałym błędzie $\sigma^2 = 0.8152$.

Kryterium Akaike

Czasami nie wiadomo ile funkcji bazowych $f_i(x)$ należy uwzględnić w dopasowaniu, czyli we wzorze (1). W szczególności, jeśli do danych doświadczalnych dopasowujemy wielomian, niekiedy — jeśli nie mamy dobrego modelu teoretycznego — nie wiemy, jaki stopień wielomianu wybrać. Jest jasne, że im wyższy stopień wielomianu, tym dopasowanie będzie “lepsze” (wielomian interpolacyjny będzie przechodził **dokładnie** przez wszystkie punkty!), ale zawsze staramy się dobrać model o jak najmniejszej liczbie parametrów.

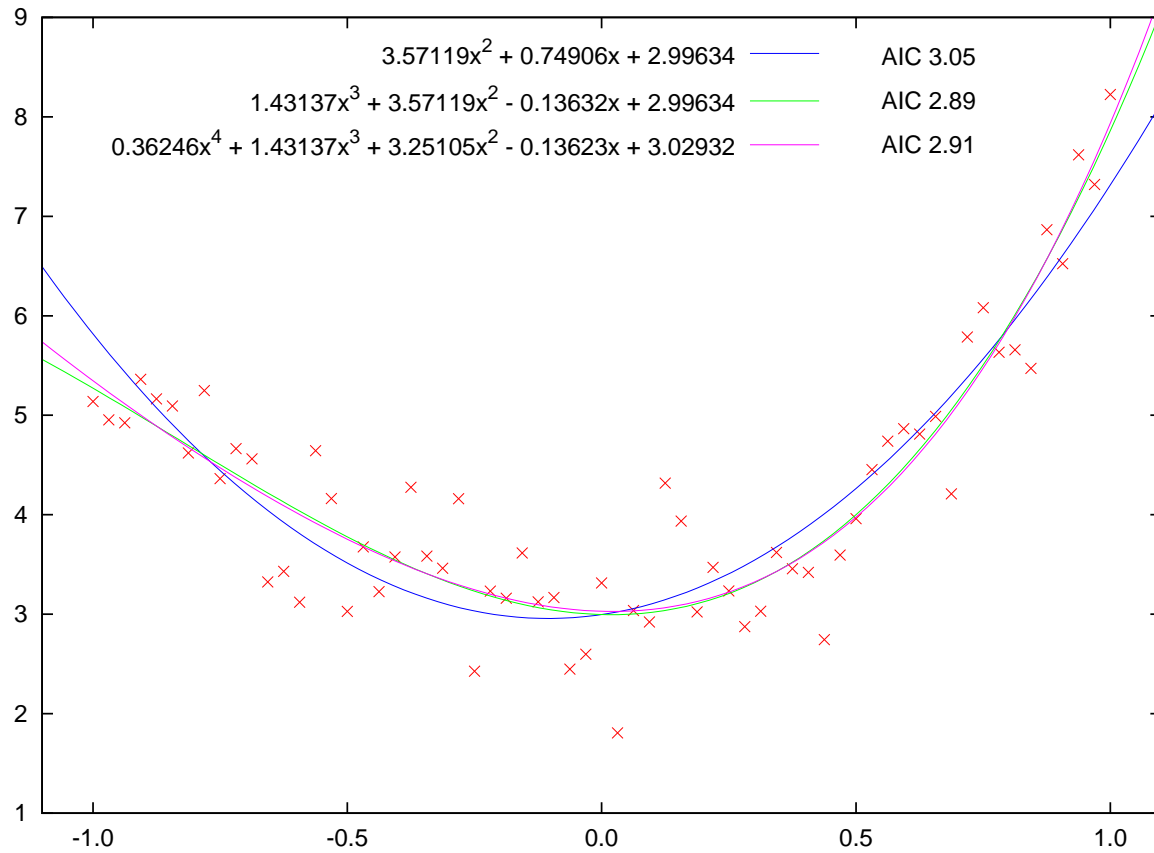
Jak zbalansować jak najlepsze dopasowanie z postulatem jak najmniejszej liczby parametrów?

Hirotsugu Akaike zaproponował kryterium, które nagradza za jak najlepsze dopasowanie, ale karze za zbyt wiele parametrów: Należy zminimalizować wielkość

$$AIC = \ln Q + \frac{2s}{N}, \quad (16)$$

gdzie Q jest wartością minimalizowanej formy kwadratowej (6) w minimum, zwaną **błędem rezydujalnym**, s liczbą parametrów, N liczbą punktów, do których dopasowujemy. *AIC* jest akronimem od Akaike Information Criterion.

Przykład



Wielomiany drugiego, trzeciego i czwartego stopnia dopasowane do tych samych danych

Nieliniowe zagadnienie najmniejszych kwadratów

Przypuśćmy, że dopasowywana do danych pomiarowych zależność teoretyczna zależy od parametrów w sposób nieliniowy,

$$\tilde{y}_i = f(x_i; \mathbf{p}) \quad (17)$$

gdzie $\mathbf{p} \in \mathbb{R}^s$ jest wektorem parametrów. Zakładamy, że $f(\cdot; \mathbf{p})$ jest *znana* funkcją, a tylko jej parametry są nieznane. *Na przykład* do danych doświadczalnych dopasowujemy funkcję Gaussa

$$y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \bar{x})^2}{2\sigma^2}\right). \quad (18)$$

Parametrami będą w tym wypadku \bar{x} oraz σ^2 . Widać, że funkcja (18) zależy od nich *nieliniowo*.

Zakładamy, że błędy pomiarowe są gaussowskie, o macierzy kowariancji \mathbf{G} . Wówczas tworzymy wektor $\mathbf{u} = [u_1, u_2, \dots, u_N]^T \in \mathbb{R}^N$, gdzie $u_i = y_i - \tilde{y}_i = y_i - f(x_i; \mathbf{p})$. Żądamy, aby funkcja

$$Q = \frac{1}{2} \mathbf{u}^T \mathbf{G}^{-1} \mathbf{u} \quad (19a)$$

osiągała minimum jako funkcja parametrów \mathbf{p} .

W najczęstszym przypadku pomiarów nieskorelowanych, obarczonych identycznymi błędami, funkcja (19a) redukuje się do postaci

$$Q = \text{const} \cdot \frac{1}{2} \sum_{i=1}^N (y_i - f(x_i; \mathbf{p}))^2 . \quad (19b)$$

Ani funkcja (19a), ani jej szczególna postać (19b), nie są formami kwadratowymi w parametrach!

Dodatnia określoność funkcji Q , $Q \geq 0$, w *praktyce* gwarantuje istnienie minimum. Nie da się jednak zagwarantować, że minimum jest tylko jedno.

Poza **bardzo nielicznymi** przypadkami, w których *łatwo* można rozwiązać układ równań $\nabla_p Q = 0$, minimum funkcji $Q(p)$ należy znaleźć **numerycznie**, przy pomocy metody Levenberga-Marquardta.

Pseudolinearyzacja

Czasami do znalezienia minimum Q stosuje się metodę *pseudolinearyzacji*. Przypuśćmy, że \mathbf{p}_n jest aktualnym przybliżeniem poszukiwanej wartości parametrów \mathbf{p} . Stawiamy hipotezę, iż “prawdziwe” wartości parametrów są małą poprawką w stosunku do \mathbf{p}_n : $\mathbf{p} \simeq \mathbf{p}_n + \delta\mathbf{p}$ i rozwijamy (17) w szereg Taylora do pierwszego rzędu:

$$\tilde{y}_i = f(x_i; \mathbf{p}_n + \delta\mathbf{p}) \simeq f(x_i; \mathbf{p}_n) + \left[\nabla_{\mathbf{p}} f |_{\mathbf{p}_n} \right]^T \delta\mathbf{p}. \quad (20)$$

Podstawiamy to rozwinięcie (dla uproszczenia zakładamy nieskorelowane, identyczne pomiary) do (19b).

Funkcja

$$Q = \frac{1}{2} \sum_{i=1}^N \left(y_i - f(x_i; \mathbf{p}_n) - [\nabla_{\mathbf{p}} f|_{\mathbf{p}_n}]^T \delta \mathbf{p} \right)^2 \quad (21)$$

jest formą kwadratową w poprawkach $\delta \mathbf{p}$. Po znalezieniu znanymi metodami wartości $\delta \mathbf{p}_{\min}$, odpowiadających (jedynemu) minimum (21), podstawiamy $\mathbf{p}_{n+1} = \mathbf{p}_n + \delta \mathbf{p}_{\min}$ i powtarzamy całą procedurę.

Taka procedura dość dobrze działa w wypadku nieliniowej metody najmniejszych kwadratów, choć **nie należy** jej polecać jako ogólnej metody minimalizacji. Pseudolinearyzacja ma tylko jedno niewątpliwe zastosowanie: Po znalezieniu **ostatecznych** wartości minimalizujących funkcję (19a), za pomocą pseudolinearyzacji wokół tego punktu znajdujemy macierz kowariancji estymatorów, będącą charakterystyką **liniową**.

Przykład

Przypuśmy, że do danych dopasowujemy funkcję Gaussa (18), zaś aktualnymi przybliżeniami parametrów są \bar{x}_n, σ_n^2 . Obliczamy

$$f(x; \bar{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \bar{x})^2}{2\sigma^2}\right), \quad (22a)$$

$$\left. \frac{\partial f}{\partial \bar{x}} \right|_{\bar{x}_n, \sigma_n^2} = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(x - \bar{x}_n)^2}{2\sigma_n^2}\right) \cdot \frac{x - \bar{x}_n}{\sigma_n^2}, \quad (22b)$$

$$\left. \frac{\partial f}{\partial (\sigma^2)} \right|_{\bar{x}_n, \sigma_n^2} = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(x - \bar{x}_n)^2}{2\sigma_n^2}\right) \cdot \left[\frac{(x - \bar{x}_n)^2}{4(\sigma_n^2)^2} - \frac{1}{2\sigma_n^2} \right] \quad (22c)$$

Wyrażenie

$$Q = \frac{1}{2} \sum_{i=1}^N \left(y_i - \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(x_i - \bar{x}_n)^2}{2\sigma_n^2}\right) \cdot \left\{ 1 + \frac{x_i - \bar{x}_n}{\sigma_n^2} \delta\bar{x} + \left[\frac{(x_i - \bar{x}_n)^2}{4(\sigma_n^2)^2} - \frac{1}{2\sigma_n^2} \right] \delta\sigma^2 \right\} \right)^2 \quad (23)$$

jest formą kwadratową w zmiennych $\delta\bar{x}$, $\delta\sigma^2$.

II. Przybliżenie Padé

Spośród wszystkich (licznych!) zagadnień aproksymacji ciągłej, omówimy tylko przybliżenia Padé.

Przypuśmy, że znamy wartości pewnej funkcji $g(z)$ i jej pochodnych do rzędu N w zerze* i na tej podstawie chcemy skonstruować przybliżenie funkcji $g(x)$ w pewnym przedziale zawierającym zero, tak, aby przybliżenie to zgadzało się z funkcją i jej N pochodnymi w zerze.

Najprostszym sposobem jest skonstruowanie rozwinięcia Maclaurina do rzędu N . Otrzymujemy w ten sposób wielomian, który co prawda spełnia wymagania w otoczeniu zera, ale przybliżenie wielomianowe zazwyczaj

*Jeżeli wartości te znamy w jakimś innym punkcie, możemy za pomocą prostej zmiany zmiennych sprowadzić ten punkt do zera.

szybko załamuje się już w niewielkiej odległości od zera. Lepsze byłoby przybliżenie wymierne.

Niech poszukiwane przybliżenie ma postać

$$R_{mk}(x) = \frac{P_m(x)}{Q_k(x)} = \frac{\sum_{j=0}^m a_j x^j}{\sum_{j=0}^k b_j x^j}, \quad (24)$$

przy czym $b_0 = 1$ i niech szeregiem Maclaurina aproksymowanej funkcji będzie

$$g(x) = \sum_{j=0}^{\infty} c_j x^j. \quad (25)$$

Przyjmijmy, że $m + k + 1 = N + 1$, czyli tyle, ile wyrazów zawiera szereg Maclaurina funkcji $g(x)$ do rzędu N . Obliczamy

$$g(x) - R_{mk}(x) = \frac{\left(\sum_{j=0}^{\infty} c_j x^j\right) \left(\sum_{j=0}^k b_j x^j\right) - \sum_{j=0}^m a_j x^j}{\sum_{j=0}^k b_j x^j}. \quad (26)$$

Funkcja $g(x)$ wraz z pochodnymi do rzędu N będzie się zgadzać z przybliżeniem (24), jeżeli w liczniku prawej strony równania (26) najniższy nieznikający wyraz będzie proporcjonalny do X^{N+1} . Otrzymujemy stąd warunki

$$\sum_{j=0}^k c_{N-s-j} b_j = 0 \quad s = 0, 1, \dots, N - m - 1; \quad c_j = 0 \text{ dla } j < 0 \quad (27a)$$

$$\sum_{j=0}^r c_{r-j} b_j = a_r \quad r = 0, 1, \dots, m; \quad b_j = 0 \text{ dla } j > k \quad (27b)$$

(27) stanowi układ $N+1$ równań liniowych na $N+1$ współczynników a_r , b_j .

Przykład

Przybliżeniem Padé $R_{22}(x)$ funkcji e^x jest

$$R_{22}(x) = \frac{12 + 6x + x^2}{12 - 6x + x^2}. \quad (28)$$

Błąd tego przybliżenia w przedziale $[-\frac{1}{2} \ln 10, \frac{1}{2} \ln 10]$ nie przekracza 0.01.

Przybliżenie Czebyszewa

Okazuje się, że lepsze przybliżenie uzyskuje się biorąc zamiast (24) iloraz wielomianów Czebyszewa:

$$C_{mk}(x) = \frac{\sum_{j=0}^m a_j T_j(x)}{\sum_{j=0}^k b_j T_j(x)}, \quad (29)$$

gdzie $T_j(x)$ jest j -tym wielomianem Czebyszewa. Dla $x \in (-1, 1)$ są one zdefiniowane jako

$$T_j(x) = \cos(j \arccos x) \quad (30)$$

i poprzez przedłużenie analityczne poza tym przedziałem. $T_j(x)$ jest wielomianem stopnia j , o najmniejszym wahaniu w przedziale $[-1, 1]$.

Niech funkcja $g(x)$ posiada rozwinięcie Czebyszewa

$$g(x) = \frac{1}{2}c_0 + \sum_{j=1}^{\infty} c_j T_j(x). \quad (31)$$

Współczynniki tego rozwinięcia otrzymujemy obliczając

$$c_j = \frac{2}{\pi} \int_{-1}^1 \frac{g(x)T_j(x)}{\sqrt{1-x^2}} dx. \quad (32)$$

Postępując jak poprzednio i korzystając z rozwinięcia (31), obliczamy

$$g(x) - C_{mk}(x) \quad (33)$$

żądając, aby w liczniku współczynniki przy $T_j(x)$ zniknęły tożsamościowo dla $j = 0, 1, \dots, N$. Otrzymujemy stąd układ równań

$$a_0 = \frac{1}{2} \sum_{i=0}^k b_i c_i \quad (34a)$$

$$a_r = \frac{1}{2} \sum_{i=0}^k b_i (c_{|r-i|} + c_{r+i}) \quad (34b)$$

($a_{r>m} \equiv 0$).

Przybliżenia Czebyszewa są lepsze od przybliżeń Padé, gdyż błąd tych pierwszych zachowuje się bardziej regularnie w całym przedziale.