

Về tính không chỉnh của bài toán hồi qui đa thức

NGUYỄN VĂN NHÂN

1. Mở đầu

Trong bài báo này, chúng tôi đề cập đến tính không chỉnh (ill-posedness) của bài toán hồi qui đa thức. Bài toán hồi qui đa thức có tầm áp dụng rất rộng trong lĩnh vực thống kê, đặc biệt là thống kê kinh tế. Chúng tôi sẽ đưa ra vài ví dụ để chứng tỏ rằng với một sai số rất bé trong số liệu thống kê có thể dẫn đến sai số rất lớn đối với hàm xu thế cần tìm. Sử dụng các hàm xu thế với sai số quá lớn như vậy để dự báo thì kết quả không tin cậy được. Các bài toán mà một sai số rất bé trong dữ kiện có thể dẫn đến một sai số rất lớn trong kết quả gọi là các bài toán không chỉnh (ill-posed). Người ta phát hiện có những bài toán như thế trong hầu hết các ngành khoa học như y học, quang học, địa vật lý, sinh học ... Đối với các bài toán không chỉnh, cần tiến hành chỉnh hóa chúng, nghĩa là xấp xỉ bài toán không chỉnh bằng một bài toán chỉnh và đánh giá sai số giữa lời giải của bài toán xấp xỉ và lời giải chính xác (mà ta không tìm được do không có dữ kiện chính xác) của bài toán ban đầu.

Trong vài thập niên gần đây, việc nghiên cứu tìm cách chỉnh hóa các bài toán không chỉnh đã thu hút sự quan tâm của nhiều nhà toán học trên khắp thế giới. Có những phương pháp chỉnh hóa áp dụng được cho một lớp khá rộng các bài toán không chỉnh như phương pháp Tikonov (xem [1]). Cũng có những phương pháp chỉnh hóa chỉ có thể áp dụng cho từng bài toán riêng lẻ (xem [2], [3]). Trong khuôn khổ bài báo này, chúng tôi chỉ chứng minh tính không chỉnh của bài toán hồi qui đa thức. Việc chỉnh hóa bài toán sẽ được trình bày trong một bài báo khác.

2. Bài toán hồi qui đa thức

Bài toán đặt ra là tìm đa thức bậc k :

$$y = f(x) = a_0 + a_1x + a_2x^2 + \dots + a_kx^k$$

xấp xỉ tốt nhất giữa các điểm dữ kiện $\{(x_i, y_i) / i = 0, 1, \dots, n\}$ theo

$$\text{nghĩa: } f(x_i) = y_i + \varepsilon_i, \quad i = 0, 1, \dots, n \quad (1)$$

$$\text{thỏa: } \sum_{i=0}^n \varepsilon_i^2 \rightarrow \min$$

Người ta thường giải các bài toán này bằng phương trình chuẩn tắc (normal equation) như sau (xem [4])

$$\text{Đặt: } X = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^k \\ 1 & x_1 & x_1^2 & \dots & x_1^k \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^k \end{bmatrix}$$

$$A = \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{bmatrix} \quad Y = \begin{bmatrix} y_0 \\ y_1 \\ \dots \\ y_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \dots \\ \varepsilon_k \end{bmatrix}$$

Hệ phương trình (1) viết dưới dạng phương trình ma trận:

$$XA = Y + \varepsilon$$

Nếu $\varepsilon \neq 0 \in \mathbb{R}^{n+1}$ thì phương trình $XA = Y$ vô nghiệm. Gọi $\text{Col } X$ là không gian vectơ con của $\mathbb{R}^{n+1}$ sinh bởi các vectơ cột của ma trận X . gọi $\hat{Y}$ là hình chiếu trực giao của Y theo tích vô hướng thông thường trong $\mathbb{R}^{n+1}$ xuống không gian vectơ $\text{Col } X$. Ta có ε đạt min nếu và chỉ nếu A là nghiệm của phương trình:

$$XA = \hat{Y} \quad (2)$$

$$\text{Ta có: } Y - \hat{Y} \perp \text{Col } X$$

$$\Leftrightarrow \langle Y - \hat{Y}, c_j \rangle = 0, \quad \forall j = \overline{1, k+1}$$

trong đó $\{c_j\}$ là các vectơ cột của X

Suy ra:

$$X^t(Y - \hat{Y}) = 0$$

$$\text{Hay } X^t X A = X^t Y \quad (3)$$

Phương trình (3) gọi là phương trình chuẩn tắc với ẩn là vectơ

$A \in \mathbb{R}^{k+1}$. Nghiệm A của phương trình chuẩn tắc cho bởi công thức:

$$A = (X^t X)^{-1} (X^t Y) \quad (4)$$

3. Vài ví dụ về tính không chỉnh của bài toán hồi qui đa thức

Ví dụ 1:

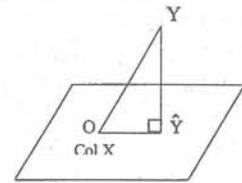
Tìm nhị thức bậc nhất $y = a_0 + a_1x$ xấp xỉ tốt nhất các điểm dữ kiện $\{(x_i, y_i)\}$, $i = 0, 1, 2$ với:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0,125 \\ 0,025 \end{bmatrix}, \quad Y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad (5)$$

Ta có:

$$X^t X = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0,125 & 0,25 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0,125 \\ 1 & 0,25 \end{bmatrix} = \begin{bmatrix} 3 & 0,375 \\ 0,375 & 0,078125 \end{bmatrix}$$

$$(X^t X)^{-1} = \frac{1}{6} \begin{bmatrix} 5 & -24 \\ -24 & 192 \end{bmatrix}$$



$$\text{và : } X^t Y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0,125 & 0,25 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$$

$$\text{Do đó : } A = \frac{1}{6} \begin{bmatrix} 5 & -24 \\ -24 & 192 \end{bmatrix} \begin{bmatrix} 6 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$$

Vậy, nhị thức bậc nhất xấp xỉ tốt nhất các điểm dữ kiện (5) là:

$$y = 1 + 8x \quad (6)$$

Bây giờ ta xét bài toán với các điểm dữ kiện bị nhiễu (các giá trị y_i bị nhiễu):

$$x = \begin{bmatrix} 0 \\ 0,125 \\ 0,25 \end{bmatrix}, \quad Y = \begin{bmatrix} 0,8 \\ 2,2 \\ 3,3 \end{bmatrix} \quad (7)$$

$$\text{Khi đó : } X^t Y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0,125 & 0,25 \end{bmatrix} \begin{bmatrix} 0,8 \\ 2,2 \\ 3,3 \end{bmatrix} = \begin{bmatrix} 6,3 \\ 1,1 \end{bmatrix}$$

$$\text{Và : } A = \frac{1}{6} \begin{bmatrix} 5 & -24 \\ -24 & 192 \end{bmatrix} \begin{bmatrix} 6,3 \\ 1,1 \end{bmatrix} = \begin{bmatrix} 0,85 \\ 10 \end{bmatrix}$$

Như vậy, với các sai số rất bé trong các giá trị y_i (0,2; 0,2; 0,3) ta thấy hàm xu thế trở thành:

$$y = 0,85 + 10x \quad (8)$$

Ta xét thêm trường hợp các giá trị x_i bị nhiễu:

$$x = \begin{bmatrix} 0 \\ 0,1 \\ 0,2 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad (9)$$

$$\text{Khi đó : } X^t Y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0,1 & 0,2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 & 0,3 \\ 0,3 & 0,05 \end{bmatrix}$$

$$(X^t X)^{-1} = \frac{1}{6} \begin{bmatrix} 5 & -30 \\ -30 & 300 \end{bmatrix}$$

$$X^t Y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0,1 & 0,2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 0,8 \end{bmatrix}$$

$$\text{Và : } A = \frac{1}{6} \begin{bmatrix} 5 & -30 \\ -30 & 300 \end{bmatrix} \begin{bmatrix} 6 \\ 0,8 \end{bmatrix} = \begin{bmatrix} 1 \\ 10 \end{bmatrix}$$

Như thế, với các sai số rất bé trong các giá trị x_i (0; -0,025; -0,05) ta thấy hàm xu thế trở thành

$$y = 1 + 10x \quad (10)$$

So sánh (6), (8) và (10), ta thấy hàm xu thế bị nhiễu khá lớn.

Ví dụ 2:

Trong ví dụ này, ta xét trường hợp thường gặp trong thống kê kinh tế là các số liệu thống kê được lấy theo từng tháng, hay từng quý, hay từng năm. Khi đó, các giá trị x_i không bị nhiễu.

Xét bài toán tìm tam thức bậc hai $y = a_0 + a_1 x + a_2 x^2$ xấp xỉ tốt nhất các điểm dữ kiện $\{(x_i; y_i)\}$, $i = 0, 1, 2, 3$ với:

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 \\ 4 \\ 10 \\ 18 \end{bmatrix} \quad (11)$$

Ta có:

$$X^t X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}$$

$$(X^t X)^{-1} = \begin{bmatrix} 7,75 & -6,75 & 1,25 \\ -6,75 & 6,45 & -1,25 \\ 1,25 & -1,25 & 0,25 \end{bmatrix}$$

$$X^t Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \\ 10 \\ 18 \end{bmatrix} = \begin{bmatrix} 32 \\ 110 \\ 394 \end{bmatrix}$$

$$\text{Suy ra : } A = \begin{bmatrix} 7,75 & -6,75 & 1,25 \\ -6,75 & 6,45 & -1,25 \\ 1,25 & -1,25 & 0,25 \end{bmatrix} \begin{bmatrix} 32 \\ 110 \\ 394 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

Vậy, hàm xu thế bậc hai là:

$$y = -2 + x + x^2 \quad (12)$$

Bây giờ, giả sử các số liệu thống kê $\{y_i\}$ bị nhiễu thành:

$$Y = \begin{bmatrix} 0,3 \\ 3,7 \\ 9,7 \\ 18,3 \end{bmatrix} \quad (13)$$

Khi đó, ta có:

$$X^t Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 0,3 \\ 3,7 \\ 9,7 \\ 18,3 \end{bmatrix} = \begin{bmatrix} 32 \\ 110 \\ 395,2 \end{bmatrix}$$

$$\text{Suy ra : } A = \begin{bmatrix} 7,75 & -6,75 & 1,25 \\ -6,75 & 6,45 & -1,25 \\ 1,25 & -1,25 & 0,25 \end{bmatrix} \begin{bmatrix} 32 \\ 110 \\ 395,2 \end{bmatrix} = \begin{bmatrix} -0,5 \\ -0,5 \\ 1,3 \end{bmatrix}$$

Như thế, với các sai số rất bé trong các giá trị y_i (0,3; -0,3; -0,3; 0,3) hàm xu thế bậc hai (12) bị nhiễu thành:

$$y = -0,5 - 0,5x + 1,3x^2 \quad (14)$$

Ta thử so sánh các giá trị dự báo y_i cho bởi các công thức (12) và (14)

Với $x_4 = 5$: (12) cho $y_4 = 28$, (14) cho $y_4 = 29,5$

Với $x_5 = 6$: (12) cho $y_5 = 40$, (14) cho $y_5 = 43,3$

Với $x_6 = 7$: (12) cho $y_6 = 54$, (14) cho $y_6 = 59,7$

(Xem tiếp trang 36)

BẢNG TÍNH VÀ PHÂN BỐ KHẤU HAO TSCĐ
Tháng 6/2004

CHỈ TIÊU	TOÀN DOANH NGHIỆP					BỘ PHẬN SỬ DỤNG					
						627		641		642	
	Nguyên giá	Thời gian sử dụng (năm)	Mức khấu hao trung bình hàng tháng	Số ngày tính (không tính) khấu hao/Số ngày trong tháng	Số khấu hao tính (không tính) trong tháng	Mức khấu hao trung bình hàng tháng	Số khấu hao tính (không tính) trong tháng	Mức khấu hao trung bình hàng tháng	Số khấu hao tính (không tính) trong tháng	Mức khấu hao trung bình hàng tháng	Số khấu hao tính (không tính) trong tháng
A	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Khấu hao trung bình hàng tháng của TSCĐ hiện có đầu tháng			37,5		37,5	19	19	4,5	4,5	14	14,0
2. KH TSCĐ tăng trong tháng			13,5		9,1	7,5	4,6	-	-	6,0	4,5
-Ngày 1: Nhà kho	540	15	3	30/ 30	3	-	-	-	-	3,0	3,0
-Ngày 7: MMTB sản xuất C	480	10	4	24/ 30	3,2	4	3,2	-	-	-	-
-Ngày 16: Xe đưa rước CNV	288	8	3	15/ 30	1,5	-	-	-	-	3,0	1,5
-Ngày 19: MMTB sản xuất D	420	10	3,5	12/ 30	1,4	3,5	1,4	-	-	-	-
3. KH TSCĐ giảm trong tháng			7,5		3,6	3	0,6	4,5	3,0	-	-
-Ngày 11: Cửa hàng	540	10	4,5	20/ 30	3	-	-	4,5	3,0	-	-
-Ngày 25: MMTB sản xuất A	360	10	3	6/ 30	0,6	3	0,6	-	-	-	-
4. KH TSCĐ tính trong tháng (4=1+2-3)					43		23		1,5		18,5
5. Khấu hao trung bình hàng tháng của TSCĐ hiện có cuối tháng (5=1+2-3)			43,5			23,5		-		20	

nghĩa là Bảng tính và phân bổ khấu hao TSCĐ trước đây không còn phù hợp nữa.

Chúng ta lần lượt xem xét tất cả các TSCĐ (cần tính khấu hao) được tính khấu hao như thế nào trong tháng? Căn cứ tình hình biến động, TSCĐ (cần tính khấu hao) trong doanh nghiệp được chia làm các nhóm sau:

Như vậy, tổng số khấu hao tính trong tháng của tất cả TSCĐ (cần tính khấu hao) được xác định như sau:

Với công thức tính khấu hao Bảng tính và phân bổ khấu hao TSCĐ được cải tiến như trên, doanh nghiệp có thể xác định Số khấu hao tính trong tháng của tất cả TSCĐ (cần tính khấu hao) dựa trên mức khấu hao trung bình hàng tháng của TSCĐ hiện có đầu tháng cộng thêm số khấu hao tính trong tháng của TSCĐ tăng trong tháng và trừ bớt đi số khấu hao không tính trong tháng của TSCĐ giảm, ngừng trong tháng.

Trên đây là những ý kiến nhằm góp phần thực hiện thống nhất qui định tại khoản 5 Điều 9 Mục III của Chế độ quản lý, sử dụng và tính khấu hao TSCĐ ban hành kèm theo

Quyết định 206 ■

(Tiếp theo trang 38)

Về tính không chính của bài toán hồi qui đa thức

4. Kết luận

Qua các ví dụ ở phần 3, ta thấy một sai số rất bé trong dữ kiện thống kê có thể dẫn đến sai số rất lớn trong hàm xu thế. Thông thường ta giả thiết các số liệu là chính xác. Điều này trong thực tế không thể có do sai số đo đạc hay do sai số làm tròn khi tính toán. Ngoài ra, ta thường nghĩ rằng một sai số rất bé trong dữ kiện cũng chỉ dẫn đến một sai số rất bé trong kết quả. Thực ra đó chỉ là tính chất của các bài toán chỉnh (well-posed). Tóm lại, bài toán hồi qui đa thức là bài toán không chỉnh và ta cần chỉnh hóa bài toán ■

Tài liệu tham khảo

- [1] Groetsch C.W, Inverse problems in the mathematica sciences, Vieweg, Germany (1993).
- [2] Ang D.D, Gorenflo R, Vy LK & Trong D.D, Moment theory and some inverse problems in potential theory and heat conduction, Springer Verlag, Germany (2002).
- [3] Huy N.V, Nhan N.V & Trong D.D, Reconstruction of analytic functions on the unit disc, Acta mathematica Vietnamica, Vol 27, N0 3 (2002) pp.307-320.
- [4] Greene W.H, Econometric analysis, MacMillan USA (1991).