직교함수를 이용한 최소자승법의 정밀도 향상에 관한 연구

A Study on the Improvement of the Accuracy for the Least-Squares Method Using Orthogonal Function

趙 元 詰* Cho, Won Cheol
李 在 浚** Lee, Jae Joon

Abstract

With increasing of computer use, a least squares method is now widely used in the regression analysis of various data.

Unreliable results of regression coefficients due to the floating point of computer and problems of ordinary least squares method are described in detail. To improve these problems, a least squares method using orthogonal function is developed. Also, comparison and analysis are performed through an example of numerical test, and re-orthogonalization method is used to increase the accuracy.

As an example of application, the optimum order of AR process for the time series of monthly flow at the Pyungchang station is determined using Akaike's FPE(Final Prediction Error) which decides optimum degree of AR process. The result shows the AR(2) process is optimum to the series at the station.

要 旨

computer 의 활용이 증대됨에 따라 각종 자료의 회귀분석에 최소자승법이 널리 사용되고 있다. computer 의 유효자리수에 따른 회귀계수의 불안정성과 표준최소자승법의 문제점을 기술하고, 이를 개선시키는 방안으로 직교함수를 이용한 최소자승법을 사용하였다. 또한 위의 두가지 방법의 결과를 수치검정예를 통하여 비교·분석하였으며 직교함수를 재직교화하여 정밀도를 향상시키는 기법도 다루었다.

적용예로 AR 과정의 적정차수를 결정하는 Akaike 의 FPE(final prediction error)를 이용하여 평창관측소의 월유량 시계열의 AR 과정 적정차수를 구하였으며, AR(2)가 적합한 것으로 선정되었다.

* 正會員·延世大學敎 工科大學 助敎授, 土木工學科
** 正會員·延世大學敎 大學院 博士課程

1. 서 론

어떤 학문이든지 그 연구활동에 있어서 관련된 변수들간에 상호 관련성을 찾으려고 할 때가 있다. 즉, 우리 주위에는 독립변수(independent variable)와 종속변수(dependent variable) 사이의 함수관계를 알아내려고 하는 수없이 많은 문제들이 있으며, 통계적 방법에서 널리 애용되고 있는 회귀분석(regression analysis)은 이러한 함수관계를 알아내는 데 매우 유익하게 쓰여지고 있다.

회귀(regression)라는 말은 영국의 우생학자 F. Galton(1822—1911)이 처음으로 사용하였으며, 누적오차를 최소로 하는 최적선을 구하는 기준으로서 제곱편차(squared deviation)를 사용하는 것이 편리하므로 최소자승법(least squares method)이라고도 한다.[1] 최소자승법에 의한 추정이나 근사화 기법은 각종 크기의 자료집단의 해석이나 조작에 널리 이용되어 왔으며, 다음과 같은 적용성을 갖고 있다.

i) 측정치가 없는 점에서의 값 추정
ii) 자료집단의 error 색출
iii) 대형 자료집단의 독특한 pattern 추출
iv) 스펙트럼의 추정

그러나 자료를 통계학적으로 해석하기 위해 이용하는 computer의 유효자릿수에 따른 절단오차는 그 결과치에 많은 영향을 주어 왔다. 그러므로 이 영향을 최대한으로 줄일 수 있는 방안으로 직교함수(orthogonal function)를 이용한 최소자승법이 개발되었는 바, 본 연구에서는 표준 최소자승법(ordinary least squares method)과 직교함수를 이용한 최소자승법의 비교·검토를 통해 후자의 우월성을 입증하고, 간단한 적용예를 보임으로서 수문자료 및 각종 자료의 통계적 회귀분석의 한 방법으로 사용될 수 있도록 한다.

2. 표준 최소자승법
(ordinary least squares method)

2.1 기본가정

변수 x와 y간에 직선회귀모형을 적합시킬 경우에는 일반적으로 다음과 같은 가정이 전제조건을 이루고 있다[2,3].

① 변수 x와 y사이에 존재하는 관련성은 주어진 x의 값에서 y의 기대치를 $\mu_{y,x}$라 할 때

$$\mu_{y,x}=\alpha+\beta\cdot x$$

와 같은 선형식으로 적절히 표현될 수 있다.

② 주어진 x의 값에서 변수 y는 정규분포를 이루며, 평균은 $\mu_{y,x}=\alpha+\beta x$로 x에 따라서 변하나 분산은 x에 관계없이 일정하다.

③ 독립변수 x는 오차없이 측정할 수 있는 수학변수(mathematical variable)이고, 종속변수 y는 측정오차를 수반하는 확률변수(random variable)이며 y의 측정오차들은 서로 독립(mutually independent)이다.

2.2 최소자승문제(least squares problem)의 일반예

다항식(polynomial)은 차수(order)의 조정에 따라 많은 형태를 가정할 수가 있으며 연속함수이기 때문에 내삽(interpolation)과 미분 또는 적분에 유용하게 사용할 수 있다.

본 절에서는 표준최소자승법의 설명을 간단히 하기 위하여 독립변수 x_n이 하나인 다항식의 경우에 국한하기로 한다.

먼저, 실험오차를 갖고 있는 실험측정자료 $Y_1, Y_2, \cdots, Y_N$을 $Y(x_1), Y(x_2), \cdots, Y(x_N)$으로 다음식과 같이 나타내 보자.

$$Y(x_n)=a_1f_1(x_n)+a_2f_2(x_n)+\cdots+a_Mf_M(x_n) \tag{1}$$

여기서, $f_1(x_n), \cdots, f_M(x_n)$은 각각의 $x_1, \cdots, x_N$에 대하여 알고 있는 값이며, 계수 $a_1, a_2, \cdots, a_M$은 미지수로 다음 식의 값이 최소가 되는 조건하에서 택해진다.

$$\sum_{n=1}^{N}\{Y_n-Y(x_n)\}^2\equiv\sum_{n=1}^{N}[Y_n-\{a_1f_1(x_n)+\cdots+a_Mf_M(x_n)\}]^2 \tag{2}$$

여기서, M은 적합함수의 수이다.

식 (2)에서 계수를 구하는 수학적 절차를 $Y(x_n)$이 두개의 함수만으로 구성된 간단한 예를 통해서 살펴보기로 한다.

$$Y(x_n)=a_1f_1(x_n)+a_2f_2(x_n) \tag{3}$$

식 (2)에서 잔차제곱의 합을 전개하면 식 (4)와 같이 된다.

$$\sum_{n=1}^{N}\{Y_n-Y(x_n)\}^2=\sum_{n=1}^{N}(Y_n)^2$$

$$+a_1^2\sum_{n=1}^{N}\{f_1(x_n)\}^2-2a_1\sum_{n=1}^{N}Y_n f_1(x_n)$$
$$+a_2^2\sum_{n=1}^{N}\{f_2(x_n)\}^2-2a_2\sum_{n=1}^{N}Y_n f_2(x_n)$$
$$+2a_1a_2\sum_{n=1}^{N}f_1(x_n)f_2(x_n) \qquad (4)$$

식 (4)의 우변에 있는 합은 모두 기지의 양이며 변수는 a_1과 a_2뿐으로 이 두개의 변수 a_1과 a_2는 식 (4)를 각각 a_1, a_2로 편미분하고 0으로 놓음으로써 얻어지는 식 (5)를 풀므로서 구할 수가 있다.

$$a_1\sum_{n=1}^{N}\{f_1(x_n)\}^2+a_2\sum_{n=1}^{N}f_1(x_n)f_2(x_n)=\sum_{n=1}^{N}f_1(x_n)Y_n$$
$$a_1\sum_{n=1}^{N}f_2(x_n)f_1(x_n)+a_2\sum_{n=1}^{N}\{f_2(x_n)\}^2=\sum_{n=1}^{N}f_2(x_n)Y_n \qquad (5)$$

이 식 (5)는 변수 a_1, a_2를 가진 두개의 선형 연립방정식이며, 두개이상(M개)의 함수가 사용되더라도 식 (4)에서의 잔차제곱합의 전개는 같은 구조를 갖으며, 이와같은 경우 M개의 계수

$$a_1\sum_{n=1}^{N}\{f_1(x_n)\}^2+a_2\sum_{n=1}^{N}f_1(x_n)f_2(x_n)+\cdots+a_M\sum_{n=1}^{N}f_1(x_n)f_M(x_n)=\sum_{n=1}^{N}f_1(x_n)Y_n$$
$$a_1\sum_{n=1}^{N}f_2(x_n)f_1(x_n)+a_2\sum_{n=1}^{N}\{f_2(x_n)\}^2+\cdots+a_M\sum_{n=1}^{N}f_2(x_n)f_M(x_n)=\sum_{n=1}^{N}(f_2(x_n)Y_n$$
$$\vdots \qquad\qquad \vdots$$
$$a_1\sum_{n=1}^{N}f_M(x_n)f_1(x_n)+a_2\sum_{n=1}^{N}f_M(x_n)f_2(x_n)+\cdots+a_M\sum_{n=1}^{N}\{f_M(x_n)\}^2=\sum_{n=1}^{N}f_M(x_n)Y_n \qquad (6)$$

각각에 대해 미분을 하면 식 (5)와 비슷한 연립방정식 (6)이 얻어진다.

식 (6)은 M개의 미지수를 가진 선형 연립방정식으로서 Gauss의 소거법이나 Cholesky method, Gauss-Seidel method 등으로 쉽게 풀 수가 있다[4,5].

또한, 최소자승법의 유용한 성질을 고찰하기 위하여 이제까지 와는 약간 다른 방법으로 미분을 해보자.

식 (4)의 좌변을 계수 a_1에 대해 미분을 한 후 0으로 놓으면 식 (7)과 같이 된다.

$$\frac{\partial}{\partial a_1}\sum_{n=1}^{N}\{Y_n-Y(x_n)\}^2=-2\sum_{n=1}^{N}\{Y_n-Y(x_n)\}\frac{\partial}{\partial a_1}Y(x_n)=0 \qquad (7)$$

식 (1)에서

$$\frac{\partial}{\partial a_1}Y(x_n)=f_1(x_n) \qquad (8)$$

이므로, 비슷한 결과가 임의의 다른 계수에 대한 미분에도 성립하여 다음과 같은 형태의 M개의 식이 얻어진다.

$$\sum_{n=1}^{N}\{Y_n-Y(x_n)\}f_m(x_n)=0,\quad m=1,2,\cdots,M \qquad (9)$$

이 식(9)는 실험자료와 적합함수와의 잔차 $Y_n-Y(x_n)$에 임의의 $f_m(x_n)$함수가 곱해진 값의 합은 0이 되도록 조정될 것이라는 것을 의미하며, 이러한 형태의 성질을 직교성(orthogonality)이라고 한다.

2.3 선형 연립방정식 해의 불안정성

2.2절에서 얻어진 표준최소자승법의 식 (6)과 같은 선형 연립방정식의 해(solution)를 구하기 위하여 보통 computer를 사용하고 있으며, 여기서는 computer의 유효자릿수로 인해 발생되는 절단오차의 영향을 Hamming[5]의 예제를 통해 간략하게 설명하고자 한다.

식(10)과 같은 2원1차 연립방정식을 생각해 보자.

$$\text{(1)}\quad a_1+\left(\frac{1}{2}\right)a_2=12$$
$$\left(\frac{1}{2}\right)a_1+\left(\frac{1}{3}\right)a_2=18 \qquad (10)$$

일반적인 방법으로 이를 풀면 $a_2=144$가 얻어지나 유효자릿수가 두자리인 Poorman's computer를 사용하면 $a_2'=150$이 되어 a_2'값은 a_2와 4%의 오차가 있다. 즉, 입력(input)에서 1%의 오차는 출력(output)에서 4%의 오차를 발생시켰으며, 같은 방법으로 a_1에 대해 검토하면 a_1'에는 5%의 오차가 발생한다.

Wampler[6]는 최소자승법에 관한 22개의 program을 사용하여 4개의 각기 다른 computer상에서 검토를 하였다. 즉, 다음과 같은 5차 다항식을 사용하였으며,

표 1. Coefficients Obtained from BMDO2R and G2/TC/MRAF Program

Power of x	Real Coeff.	BMDO2R Value (Wampler)	G2/TC/MRAF Value (This Study)
1.0	1.00000	−17.13281	0.91433
x^1	1.00000	39.34436	1.00007
x^2	1.00000	−13.26675	0.99810
x^3	1.00000	2.92344	1.01808
x^4	1.00000	0.89241	0.92260
x^5	1.00000	1.00212	1.14314

$$Y_n=1.0+1.0x_n+1.0x_n^2+1.0x_n^3+1.0x_n^4+1.0x_n^5$$

x_n을 1부터 20까지의 정수를 대입하여 Y_n을 얻고 최소자승해를 구하였다.

또한 본 연구에서는 G2/TC/MRAF program[7]을 이용하여 Apple Ⅱ plus 개인용 콤퓨터상에서 동일한 자료를 시험하였으며, 그 결과와 Wampler의 결과 일부를 표 1에 수록하였다.

표 1에서 Wampler의 계산은 7-digit UNIVAC 1108 computer 상에서 얻어진 결과이고 본 연구에서 시도한 G2/TC/MRAF program의 계산은 Apple Ⅱ plus 개인용 콤퓨터상에서 얻어진 결과로 실제의 계수와 커다란 오차를 보이고 있음을 알 수 있다.

3. 직교함수를 이용한 최소자승법

선형 연립방정식을 풀므로서 최소자승해를 구하는 방법의 결점을 보완하기 위한 방법에는 Gram-Schmidt의 orthonormalization method (직교정규화 방법)[8]와 Lawson-Hanson의 orthogonal householder transformation method[9]가 있는데 본 연구에서는 Gram-Schmidt의 직교정규화 방법을 다루기로 한다.

3.1 Gram-Schmidt의 직교정규화 방법

식(4)에서 $\sum_{n=1}^{N} f_1(x_n)f_2(x_n)=0$이 되면 계수의 교차항 a_1a_2는 존재하지 않게 된다. 식(5)가 얻어지도록 과정을 반복하면

$$a_1=\sum_{n=1}^{N} f_1(x_n)\,Y_n \Big/ \sum_{n=1}^{N}\{f_1(x_n)\}^2$$

$$a_2=\sum_{n=1}^{N} f_2(x_n)\,Y_n \Big/ \sum_{n=1}^{N}\{f_2(x_n)\}^2 \qquad (11)$$

식(6)의 경우에는 만일 $\sum_{n=1}^{N} f_m(x_n)f_{m'}(x_n)$ $(m\neq m')$ 형태의 항 모두가 0이 되면 식 (11)과 같은 형태로 간단하게 줄일 수가 있다. 즉,

$$a_i=\sum_{n=1}^{N} f_i(x_n)\,Y_n \Big/ \sum_{n=1}^{N}\{f_i(x_n)\}^2 \qquad (12)$$

Gram-Schmidt 방법에서는 원시함수(original function)를 약간 변형시킨 후 재집단화(regrouping)함으로서 교차항을 소거한다. 예로서 다음의 적합 방정식(fitting equation)을 보면

$$Y(x_n)=a_1f_1(x_n)+a_2f_2(x_n)$$

이 식에 $ra_2f_1(x_n)$항을 더하고 빼주면서 다음과 같이 재집단화 한다.

$$Y(x_n)=(a_1+ra_2)\{f_1(x_n)\}+a_2\{f_2(x_n)-rf_1(x_n)\} \qquad (13)$$

다음의 정의를 사용하여

$$C_1\equiv a_1+ra_2$$

$$C_2\equiv a_2$$

$$F_1(x_n)\equiv f_1(x_n)$$

$$F_2(x_n)\equiv f_2(x_n)-rf_1(x_n)$$

식(13)을 다시 쓰면 식(14)가 된다.

$$Y(x_n)=C_1F_1(x_n)+C_2F_2(x_n) \qquad (14)$$

식(14)를 사용하여 잔차제곱의 합을 전개하였을 때 새로운 계수에서의 교차항은 $C_1C_2\sum_{n=1}^{N}F_1(x_n)F_2(x_n)$이다. 이제까지 사용한 매개변수 r은 규정되지 않은 자유 매개변수(free parameter)였지만 아래와 같이 구해진다. $\sum_{n=1}^{N}F_1(x_n)F_2(x_n)=0$이 되도록 위의 정의를 사용하면

$$\sum_{n=1}^{N}F_1(x_n)F_2(x_n)=\sum_{n=1}^{N}F_1(x_n)\{f_2(x_n)-rf_1(x_n)\}$$

$$=\sum_{n=1}^{N}F_1(x_n)f_2(x_n)-r\sum_{n=1}^{N}\{F_1(x_n)\}^2 \quad (15)$$

이 되고, 따라서

$$r=\sum_{n=1}^{N}F_1(x_n)f_2(x_n)\Big/\sum_{n=1}^{N}\{F_1(x_n)\}^2 \quad (16)$$

이 된다. 또한 직교계수(orthogonal coefficient)는 다음 식으로 주어진다.

$$C_1=\sum_{n=1}^{N}F_1(x_n)\,Y_n\Big/\sum_{n=1}^{N}\{F_1(x_n)\}^2$$

$$C_2=\sum_{n=1}^{N}F_2(x_n)\,Y_n\Big/\sum_{n=1}^{N}\{F_2(x_n)\}^2 \quad (17)$$

일단 C_1, C_2가 알려지면 원시계수(original coefficient)는 다음 식에 의해서 바로 주어진다.

$$a_2=C_2$$

$$a_1=C_1-ra_2 \quad (18)$$

식(15)에 보인 성질을 갖으며 함수들의 곱의 합이 0인 함수를 직교(orthogonal)라고 하는데 직교란 용어는 기하학적인 용어로 수선을 의미한다. 식(15)의 대수학 형태와 기하학의 직각간의 관계는 Hamming이 정연하게 표현한 바 있다.

매개변수 r과 직교계수 C_1, C_2를 계산하기 위해서는 직교함수가 정규화(normalization)되었을 경우가 보다 편리하다. 이 정규화는 각각의 직교함수를 축척변수 S_m로 나눔으로서 가능하게 되며 다음과 같은 새로운 함수값이 발생된다.

$$\left.\begin{aligned}&\bar{F}_m(x_n)=\frac{F_m(x_n)}{S_m}\\&\sum_{n=1}^{N}\{\bar{F}_m(x_n)\}^2=1\\&\bar{F}_m(x_n)=F_m(x_n)\Big/\left[\sum_{n=1}^{N}\{F_m(x_n)\}^2\right]^{1/2}\end{aligned}\right\}$$

$m=1, 2, \cdots, M$에 대하여 (19)

위와 같이 정규화된 직교함수를 직교정규함수(orthonormal function)라 하며, 함수값은 모두 (−1, 1) 사이에 놓이도록 조정되었기 때문에 직교계수 $C_1, C_2, \cdots, C_M$의 크기는 총 적합도에 대한 각 직교정규함수의 상대적 중요도를 보여준다. 즉, 몇몇 계수의 값이 다른 것들에 비해 훨씬 작으면 이들이 곱해진 함수는 전체적인 적합함수에 중요하지 않다는 사실을 시사해 준다.

이상의 과정을 3개의 원시함수를 사용하여 일반적으로 설명하면 다음과 같다.

$$Y(x_n)=a_1f_1(x_n)+a_2f_2(x_n)+a_3f_3(x_n) \quad (20)$$

여기에는 식(21)과 같은 3개의 직교성 조건이 만족되어야 한다.

$$\sum_{n=1}^{N}F_1(x_n)F_2(x_n)=\sum_{n=1}^{N}F_1(x_n)F_3(x_n)$$

$$=\sum_{n=1}^{N}F_2(x_n)F_3(x_n)=0 \quad (21)$$

다음의 알고리즘에서 현 단계에 필요한 직교함수들은 이전 단계에서 정규화된 것으로 가정하며, 이후부터 정규화된 함수의 기호도 $F_m(x_n)$으로 사용하기도 한다.

직교함수의 정규화와 직교계수를 구하는 알고리즘은 다음과 같다.

알고리즘

$F_1(x_n)=f_1(x_n)$

정규화

⋮

$$r_{21}=\sum_{n=1}^{N}f_2(x_n)F_1(x_n)$$

$F_2(x_n)=f_2(x_n)-r_{21}F_1(x_n)$

정규화

⋮

$$r_{31}=\sum_{n=1}^{N}f_3(x_n)F_1(x_n)$$

$$r_{32}=\sum_{n=1}^{N}f_3(x_n)\;F_2(x_n)$$

$F_3(x_n)=f_3(x_n)-r_{31}F_1(x_n)-r_{32}F_2(x_n)$

정규화 (22)

이 때 직교계수들은 식 (23)으로 주어진다.

$$C_1=\sum_{n=1}^{N}F_1(x_n)\,Y_n$$

$$C_2=\sum_{n=1}^{N}F_2(x_n)\,Y_n$$

$$C_3=\sum_{n=1}^{N}F_3(x_n)\,Y_n \quad (23)$$

직교계수가 정규화 되었을 때 직교계수를 원시계수로 재치환 시키는 알고리즘은 3.3절에서 설명하기로 한다.

3.2 직교함수의 도해

그림 1과 그림 2는 원시함수가 어떻게 직교함수로 변환되는 가를 보기위해 도시하였다.

그림 1은 3개의 원시함수를 도시한 것으로 $f_1(x_n)$과 $f_3(x_n)$은 다음과 같이 정의된 연속함수이다.

$f_1(x_n)=1.0$

$f_3(x_n)=0.015(x_n)^2$, $x_n=1, 2, \cdots, 20$

에 대하여

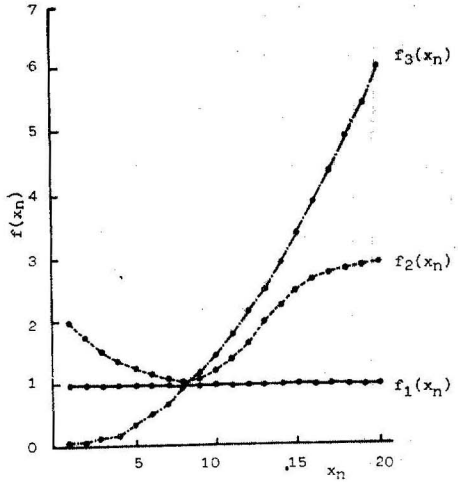

그림 1. Graphs of Three Different Original Function

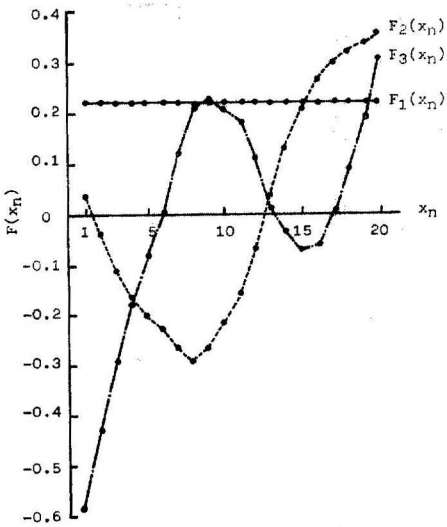

그림 2. Graphs of Three Orthogonal Functions Generated from the Three Original Functions in Fig. 1

또한 $f_2(x_n)$은 $f_2(1)=2.0$에서 시작하여 $f_2(8)=1.0$까지 감소하다가 다시 x_n이 커짐에 따라 3.0을 향하여 증가하는 성질을 갖도록 선택된 경험적 함수(empirical function)이다.

그림 2는 이 세개의 원시함수로부터 발생된 직교함수를 도시한 것이며, 직교함수는 정규화하였다. 식(22)에 따라서 첫번째 직교함수 $F_1(x_n)$은 그것을 정규화하기 위해 축척변수로 나누어진 것을 제외하고는 $f_1(x_n)$과 같고, $F_2(x_n)$은 (+)값에서 시작하여 (−)값으로 감소하다가 다시 (+)값으로 증가하고 있다. 세번째 직교함수 $F_3(x_n)$은 그러한 방향전환을 두번 보여주고 있는데, 이들 원시함수와 직교함수간의 가장 뚜렷한 차이는 직교함수는 (+)와 (−)값을 갖는다는 점으로 (−)값은 식(21)의 관계를 만족시키기 위하여 필요하다.

3.3 직교계수로 부터 원시계수로의 변환

식(22)의 결과를 간단히 재정리하면 다음과 같이 된다.

$$f_1(x_n)=F_1(x_n)$$
$$f_2(x_n)=F_2(x_n)+r_{21}F_1(x_n)$$
$$f_3(x_n)=F_3(x_n)+r_{32}F_2(x_n)+r_{31}F_1(x_n)$$

본 절에서는 명확히 하기 위해서 직교함수를 정규화하지 않기로 하며, 이는 적합함수를 다음과 같이 쓸 수 있음을 뜻한다.

$$\begin{aligned}Y(x_n)&=a_1f_1(x_n)+a_2f_2(x_n)+a_3f_3(x_n)\\&\equiv F_1(x_n)\{a_1+a_2r_{21}+a_3r_{31}\}\\&\quad+F_2(x_n)\{a_2+a_3r_{32}\}\\&\quad+F_3(x_n)\{a_3\}\end{aligned}\tag{24}$$

또한 적합함수는 다음과 같이 쓸 수 있으므로

$$Y(x_n)=C_1F_1(x_n)+C_2F_2(x_n)+C_3F_3(x_n)\tag{25}$$

식(24)와 (25)로 부터 원시계수는 다음과 같이 구해진다.

$$\begin{aligned}a_3&=C_3\\a_2&=C_2-a_3r_{32}\\a_1&=C_1-a_2r_{21}-a_3r_{31}\end{aligned}\tag{26}$$

식(26)의 과정은 마지막(M번째) 원시함수의 계수가 제일 먼저 계산되고 $f_1(x_n)$의 계수는 마지막에 계산되므로 Gram-Schmidt 방법과는 상반되는 의미를 갖고 있다. 이러한 거동은 표 1과 4.1절의 표 2에서 보는 바와 같이 마지막 계수 a_6는 전형적으로 아주 정밀하고 첫번째 계수 a_1은 전체 계수중에서 오차가 가장 큰 사실로 부터도 알 수 있다.

4. 수치검정 및 재직교화 방법에 의한 보정

4.1 수치검정

이제까지의 내용을 돌이켜 보면 중요한 사항은 자릿수를 많이 저장하는 computer 일수록 보

다 정밀한 결과를 준다는 점이다. 이에 대한 수치검정으로 2.3절에서 사용한 것과 동일한 함수를 IBM 360-195와 CDC-6600 그리고 CYBER 170-825 및 개인용 콤퓨터인 Apple Ⅱ plus상에서 실행해 보았다. IBM 360-195는 단정도(single precision) 사용에서 7-digit 정밀도를 갖고 있고 CDC-6600과 CYBER 170-825는 14-digit 정밀도를 갖고 있는 대형 computer이며, Apple Ⅱ plus는 7-digit 정밀도를 갖고 있는 소규모 용량의 개인용 콤퓨터이다.

검정자료 $Y_1, Y_2, \cdots Y_{20}$은 다음의 5차 다항식으로 부터 발생하였으며,

$$Y_n = 1.0 + 1.0x_n + 1.0x_n^2 + 1.0x_n^3 + 1.0x_n^4 + 1.0x_n^5$$

$x_1, x_2, \cdots, x_{20}$은 1부터 20까지의 정수를 택하였다. 선택된 적합함수는

$$Y(x_n) = a_1 + a_2x_n + a_3x_n^2 + a_4x_n^3 + a_5x_n^4 + a_6x_n^5$$

으로 5차 다항식의 적합함수는 발생된 자료로부터 5차 다항식과 부합하는 결과를 주어야만 한다. 즉, 계수 $a_1, a_2, \cdots, a_6$는 모두 1.0이어야만 한다.

표 2는 4가지 computer상에서의 수행결과를 나타낸 것으로 14-digit에서의 직교함수를 이용한 최소자승법의 결과가 매우 정밀함을 보여주고 있으며, 7-digit상에서의 정도는 일부 계수에서는 적당하나 일부에서는 부적당함을 보여주고 있다. 그러나 개인용 콤퓨터인 Apple Ⅱ plus의 배정도(double precision) 결과는 만족할만한 결과를 보여주고 있다.

표 3은 검정자료와 IBM 360-195상에서 단정도로 실행한 결과와 Apple Ⅱ plus상에서 단·배정도로 실행한 결과 및 BMDP 5R program[10]으로 CYBER 170-825상에서 실행한 결과의 종족변수 값을 수록한 것으로 직교함수법의 또다른 특징을 발견할 수 있다. 즉, IBM 360-195 결과와 Apple Ⅱ plus 단정도 결과는 표 2의 계수로 부터 기대되는 결과보다 훨씬 정밀한 종속변수 값을 얻게 된 것으로서, 이는 원시함수를 먼저 직교함수로 변환한 후 직교정규계수를 산정하고, 종속변수 값을 계산하는데 이 직교정규함수와 계수를 사용하기 때문이다. 또한 Apple Ⅱ plus 배정도로 실행한 결과는 검정자료와 거의 같음을 보여주고 있어 개인용 콤퓨터의 활용에 대한 가능성을 시사해주고 있다.

그러나 원시함수의 계수를 구하는 과정은 절단오차의 형태에 따라 종속되는데 이 오차에 대한 부분적인 수정은 4.2절에서 논의하기로 한다.

4.2 재직교화(re-orthogonalization) 방법에 의한 보정

직교정규함수(orthonormal function)의 정밀도가 높아질수록 계산의 정밀도는 높아지게 된다. 3장에서 언급한 직교정규함수를 발생시키는 Gram-Schmidt 방법은 순환방법이었다. 즉, $F_3(x_n)$은 이전의 두개의 직교함수 $F_2(x_n)$, $F_1(x_n)$과 원시함수 $f_3(x_n)$으로 부터 발생된다. 이 과정에서 발생된 함수가 이전에 발생된 함수와 실제로 직교성 조건을 만족하는 지에 대한 일관성 있는 검토가 수립되어 있지 못하고 또 순환방법은 비정밀한 결과를 끊임없이 발생시킬 수 있기 때문에 중요한 문제가 된다. 그러므로 원래의 Gram-Schmidt 방법을 검토하고 이를 보정하는 간단한 기법으로 재직교화(re-orthogonalization)방법[11]을 사용하기로 한다.

발생된 함수 $F_1(x_n)$과 $F_2(x_n)$이 서로 직교한다고 가정하면, 이 가정은 허용정도내에서 다음 관계가 만족되어야 함을 의미한다.

$$\sum_{n=1}^{N} F_1(x_n) F_2(x_n) = 0 \qquad (27)$$

이 때 $F_3(x_n)$은 식(22)로 부터 발생된다. 직교성 검토과정에서 $F_3(x_n)$이 $F_1(x_n)$과 $F_2(x_n)$에 직교가 아님이 판명되었다고 하면, 즉 식(28)의 관계가 성립된다.

$$\sum_{n=1}^{N} F_3(x_n) F_1(x_n) = d_{31}$$
$$\sum_{n=1}^{N} F_3(x_n) F_2(x_n) = d_{32} \qquad (28)$$

여기서, d_{31}, d_{32}는 허용오차 이상의 편차이다. 그러므로 보다 정밀한 재직교화 함수(re-orthogonalization function) $F_3^{(R)}(x_n)$을 다음 식에 의해 발생시킨다.

$$F_3^{(R)}(x_n) = F_3(x_n) - d_{31}F_1(x_n) - d_{32}F_2(x_n) \qquad (29)$$

표 2. Results of Orthogonal Least-Squares Test on Four Different Computers

Coeff.	Real Value	CDC-6600 14-digit	IBM 360-195 7-digit	Apple Ⅱ Plus		CYBER 170-825 BMDP5R
				Double Pr.	Single Pr.	
a_1	1.00000	1.00000	−1.37109	1.06563	0.87679	1.00000
a_2	1.00000	1.00000	1.79687	0.98190	0.90630	1.00000
a_3	1.00000	1.00000	0.87500	1.00801	1.02650	1.00000
a_4	1.00000	1.00000	1.00708	0.99926	0.99660	1.00000
a_5	1.00000	1.00000	0.99992	1.00003	1.00020	1.00000
a_6	1.00000	1.00000	0.99999	1.00000	1.00000	1.00000

표 3. Actual Values of Test Data and Values Obtained from Orthogonal Least-Squares Approach

Data	Real Value	IBM360-195 7-digit	Apple Ⅱ Plus		CYBER 170-825 BMDP5R
			Double Pr.	Single Pr.	
Y_1	6.0	6.1	5.998	6.005	6.0
Y_3	364.0	364.1	363.997	363.983	364.0
Y_5	3906.0	3906.5	3906.000	3906.011	3909.0
Y_7	19608.0	19608.6	19608.010	19607.996	19608.0
Y_9	66430.0	66430.4	66430.000	66429.961	66430.0
Y_{11}	177156.0	177156.4	177155.984	177155.969	177156.0
Y_{13}	402234.0	402234.4	402233.969	402234.031	402234.0
Y_{15}	813616.0	813616.4	813616.000	813616.0	813616.0
Y_{17}	1508598.0	1508598.0	1508598.000	1508598.0	1508598.0
Y_{19}	2613660.0	2613659.0	2613660.000	2613660.0	2613660.0

식(29)를 이용하여 다음 관계가 성립하는지를 검토하고, 동일한 방법으로 $F_3^{(R)}(x_n)$

$$\sum_{n=1}^{N} F_3^{(R)}(x_n) = \sum_{n=1}^{N} F_3(x_n) F_1(x_n) - d_{31} \sum_{n=1}^{N} \{F_1(x_n)\}^2 = d_{31} - d_{31} \simeq 0 \qquad (30)$$

은 $F_2(x_n)$과 직교가 됨을 증명할 수가 있다. 여기서 근사적으로 같다는 기호($\simeq 0$)는 직교성 관계가 허용오차내에서 만족됨을 보이기 위하여 사용하였다.

이와 같은 보정절차는 식(22)에 주어진 과정과 비슷하며, $F_3(x_n)$을 계산한 후에 식(28)의 d_{31}, d_{32}를 구하고 이를 $F_3^{(R)}(x_n)$을 산정하기 위하여 사용한다. 이렇게 재직교화된 보정함수는 그 다음에 정규화하고 보정 알고리즘은 식(28), (29)에서 $F_3(x_n)$이 $f_3(x_n)$으로 대체된 것 외에는 원래의 Gram-Schmidt 방법과 유사하며, 첫번째 함수를 제외하고는 모든 직교함수에 대해 사용한다.

5. 시계열에의 적용

다음과 같은 자기회귀 시계열(auto-regressive time series)을 생각해 보자[12,13].

$$Y_n = a_1 Y_{n-1} + a_2 Y_{n-2} + \cdots + a_M Y_{n-M} + S_n \qquad (31)$$

여기서, 계수 $a_1, a_2, \cdots, a_M$과 M은 미지수이며, 정수 M을 자기회귀 시계열 또는 자기회귀과정(AR process)의 차수(order)라 한다. 확률성분 $\{S_n\}$은 정규분포를 이룬다고 가정하면 확률이론의 최우원리(principle of maximum likelihood)를 사용하여 다음 식을 최소화시키는 계수를 구할 수가 있다.

$$\sum_{n=M+1}^{N} \{Y_n - (a_1 Y_{n-1} + a_2 Y_{n-2} + \cdots + a_M Y_{n-M})\}^2 \qquad (32)$$

미지계수를 구하기 위하여 식(32)를 최소화하는 것은 일반적인 최소 자승문제로서 시계열 해

석에서는 보통 표준최소자승법을 사용하여 왔다. 표준최소자승법으로 계수를 구하기 위해서는 식(6)과 비슷한 선형 연립방정식을 풀어야 하는데, 계수를 곱한 합은 다음의 형태를 이루고 있다.

$$\sum_{n=M+1}^{N} Y_n Y_{n-k} \qquad (33)$$

여기서, k는 1부터 M까지의 정수이다. 이 합의 내용은 시계열 분석가들에게 자료들의 상관에 관한 유용한 정보를 제공해 주고는 있지만, 대형의 시스템 방정식에 대해서 표준최소자승법은 수치적으로 부정확한 계수를 유발시킬 수 있음에 특히 주의해야 한다. 따라서 이와 같은 경우에 직교함수를 이용한 최소자승법을 적절하게 사용할 수 있을 것이다.

그러나 AR 과정에 남아있는 문제는 사계열에 적합한 AR의 차수 M을 합리적으로 구하는 것이다. 이 문제에 대한 연구는 아직도 계속되고 있지만 Akaike[14,15]는 다음과 같은 실용적인 제안을 하였다.

식(32)의 AR 적합함수와 자료치간의 편차제곱의 합을 S_M^2으로 놓으면, M이 증가함에 따라 S_M^2은 일반적으로 감소한다. 그러나 M개의 계수 $a_1, a_2, \cdots, a_M$은 통계학적 분산이나 그에 관련된 불확정성을 갖게 된다. 즉, 아주 긴 시계열 중에서 측정된 구간은 부분적으로 그 시계열을 대표하고 있는 것이므로 자료치에 정확하게 적합되지 않으면 안된다. 그리하여 사용된 계수가 많아질수록 적합도는 높아지나 통계학적으로는 신뢰성이 떨어지게 된다.

Akaike는 적합의 정확성과 통계학적 신뢰성간의 모순을 FPE(final prediction error)로서 해결하였다. 이 기준에 의하면 AR 과정의 차수 M은 다음 식의 값이 최소화되는 M으로서 결정된다.

$$\mathrm{FPE}(M)=\left\{\frac{N+(M+1)}{N-(M+1)}\right\} S_M^2 \qquad (34)$$

여기서 N은 전체자료의 수이다.

본 연구에서는 FPE를 사용하여 평창강 유역의 평창관측소의 월유량(1974~1981 ; 8년간)에 관한 AR 과정의 최적차수를 결정하였다. 이에 대한 결과로 FPE 값의 변화를 도시한 것이 그림 3이며 이 그림에서 보면 2차와 11차가 최소값을 보여주고 있다. 그러나 차수가 높을 경우의 계산의 복잡성으로 인한 불편때문에 이 지점의 월유량은 AR(2)가 적합한 것으로 판단된다.

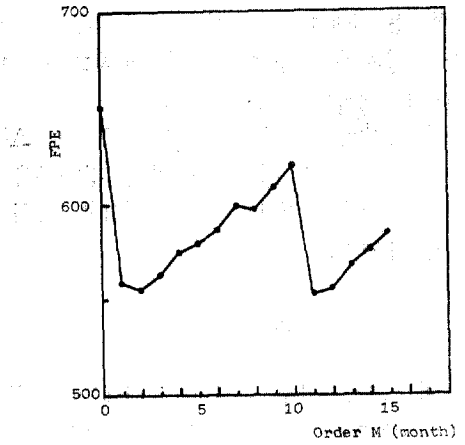

그림 3. The Value of FPE Corresponding to M-th Order AR Process

표 4. Estimated Coefficients of AR(2) for Monthly Runoff at the Pyungchang Station

Term	AR Coefficient
a_1	0.33496
a_2	0.15913

표 4는 평창관측소의 월유량에 대한 AR(2) 과정의 AR 계수를 수록한 것이다.

6. 결 론

본 연구는 각종자료의 해석시 널리 이용되고 있는 최소자승법에 대하여 분석한 것으로서, 표준최소자승법과 직교함수를 이용한 최소자승법을 비교하여 정밀도를 향상시키는 방안에 대하여 연구하였다.

본 연구를 통해 얻어진 성과는 다음과 같다.

1) computer의 유효자릿수는 정밀도에 큰 영향을 미치므로 가능한 한 유효자릿수가 큰 computer를 사용하는 것이 좋은 결과가 얻어진다.

2) 표준최소자승법의 불안정한 결과를 개선시킬 수 있는 직교함수를 이용한 최소자승법의 우

월성을 보였다.

3) 개인용 콤퓨터인 Apple Ⅱ plus 에서도 신뢰할만한 결과를 얻을 수 있도록 재직교화 방법에 의한 보정을 하였다.

4) AR 모형의 적정차수를 결정하기 위한 Akaike 의 FPE 기준을 평창강 유역의 평창관측소의 월유량에 대해 적용한 결과 AR(2) 모형이 적합함을 보였다.

謝 辭

본 연구는 아산사회복지사업재단의 1985 년 연구비 지원에 의하여 수행되었으며 이에 당 재단에 감사를 드린다.

참 고 문 헌

1. 박성현, 「회귀분석」, 대영사, p. 604, 1981.
2. Draper, N.R. and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, Inc., New York, p. 407, 1981.
3. Acton, F.S., *Analysis of Straight-Line Data*, Dover Publ., New York, p. 267, 1966.
4. Carnahan, B., H.A. Luther, and J.D. Wilkes, *Applied Numerical Methods*, John Wiley & Sons, Inc., New York, pp. 269-340, 1969.
5. Hamming, R.W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill Book Co., New York, pp. 229-239, 1962.
6. Wampler, R.H., "An Evaluation of Linear Least Squares Computer Programs", *Jour. of Research of the National Bureau of Standards-B. Mathematical Sciences*, Vol. 73B(2), pp. 59-90, 1969.
7. 東京大學大型計算機センタライブラリープログラム 第 1 集, G2/TC/MRAF, 東京大學 出版會, 1967.
8. Abdelmalek, N.N., Round-off Error Analysis for Gram-Schmidt Method in Solution of Linear Least Squares Problems, *Nordisk Tidskrift for Information Behandling*, Vol. 11, pp. 345-368, 1971.
9. Lawson, C.L. and R.J. Hason, *Solving Least Squares Problems*,, Prentice-Hall, Englewood Cliffs, N.J., p. 340, 1974.
10. Dixon, W.J. et al., *BMDP Statistical Software*, Univ. of California Press, Berkeley, p. 733, 1983.
11. Dixon, R. and E.A. Spackman, The Three-Dimensional Analysis of Meteorological Data, *Meteorological Office Scientistic Paper* No. 31, Her Majesty's Stationary Office, p. 28, 1970.
12. Box, G.E.P. and G.M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, p. 575, 1976.
13. Anderson, O.D., Some Methods of Time Series Analysis, *Math. Scientist*, Vol. 1, pp. 27-41, 1976.
14. Akaike, H., Fitting Autoregressive Models for Prediction, *Ann. Inst. Statist. Math.*, Vol. 21, pp. 243-247, 1969.
15. Akaike, H., Statistical Predictor Identification, *Ann, Inst. Statist. Math.*, Vol. 22, pp. 203-217, 1969.

(接受 : 1986. 8. 26)