연세경제연구 제 V 권, 제 1 호
Yonsei Economic Studies, Vol. V, No. 1 (March 1998)

FAN방법을 이용한 공분산분석모형의 추정

안윤기 · 송돈수

본 연구는 비모수회귀분석에 사용되는 후진방법을 부분선형모형의 확장인 공분산분석모형에서 추정하는 데 적용하여 기존의 추정방법인 GJS방법과 비교를 해 보았다. 또한, 이 두 가지 방법들에서 비모수부분을 추정하는 방법으로는 커널추정법과 1992년 FAN이 제시한 가중국소선형회귀방법을 사용함으로써 네 가지 방법들을 모의실험에 의한 비교를 해 보았다.

일반화된 교차타당성(GCV)을 이용한 최적의 평활계수 h를 선택하였으며, 커널방법 또는 FAN방법을 사용해도 기존의 GJS방법보다 후진방법이 편기와 평균추정오차값의 측면에서 더 나은 결과를 주었다. 또한, 비모수부분을 FAN방법과 후진방법을 사용하는 것이 MASE면에서 제일 좋은 결과를 얻었다.

I. 서 론

반모수모형(semiparametric model) 중에서 가법모형을 가정한 부분선형모형(partial linear model)은 다음과 같이 나타낼 수 있다.

$$Y_i = \beta^T Z_i + m(x_i) + \varepsilon_i, \qquad i = 1, \cdots, n \qquad (1\text{-}1)$$

$$\beta^T Z = \text{선형모수부분}, \qquad m(x) = \text{비모수부분}$$

위의 모형에서 평균반응변수는 선형형태의 일부 변수와 모수화할 수 없는 비모수 함수형태의 다른 변수들의 가법형태이다. 여기서 독립변수 Z_i는 자료로 주어진 고정 p벡터이고 β는 미지의 모수벡터로 $\beta^T Z$는 선형모수부분이며 x_i는 한계영역 $D \subset R^k$의 또 다른 독립변수이다. 잘 알려진 공분산분석(ANACOVA)모형은 Z_i가 질적 변수

연세대학교 응용통계학과. 서울시 서대문구 신촌동 134, 120-749 및 (주)미디어 리서치. 서울시 서초구 서초동 1625-1.

(indicator variable)이고 x_i가 공변량변수(covariate variable)이며 $m(x)$가 선형모형인 부분선형모형의 특수형태이다. 지금까지 β와 m을 동시에 추정하는 몇 가지 방법들이 제시되어 왔다. 그 중 한 가지 방법은 Engle [7], Green [9] 및 Wabha [10]에 의해 벌칙최소자승(penalized least square)의 개념에 기초한 Green-Jennison- Seheult(GJS) 방법이고 나머지 하나는 Denby [5], [6]가 제시한 부분오차분석법(partial residual analysis method)이다. 앞에서 제시한 모든 방법들은 $m(x)$를 추정하는 단계에서 커널 평활(Kernel smoothing)방법들을 사용하였으나 본 논문에서는 커널방법 대신 FAN이 제시한 가중국소선형회귀(local weighted linear regression)방법을 사용하고 Buja [3]가 일반적 가법모형(generalized additive model)에 적용한 후진추정방법을 사용하는 방법을 제시하고 기존방법들과 모의실험(simulation)을 통해 비교분석하였다.

Ⅱ. 후진방법

식 (1-1)에 제시된 모형에서 β와 $m(x)$를 동시에 추정하는 방법은 먼저 모형 (1-1)에 표현된 비모수부분인 $m(x)$를 아래와 같이 모수화한다.

$$(m(x_1), \cdots, m(x_n))^T = Wr \qquad (2\text{-}1)$$

W는 완전랭크의 $n \times q$, r은 부가적인 모수

식 (2-1)을 모형 (1-1)에 대입하면 행렬형태의 모형으로 다음과 같이 나타낼 수 있다.

$$Y = Z\beta + Wr + \epsilon \qquad (2\text{-}2)$$

일반선형모형에서 모수를 추정하는 OLS추정치 β와 r은 아래의 정규공식을 만족하여야 한다.

$$\begin{aligned} &Z'Z\beta = Z'(Y-r) \\ &Wr = P_W(Y-x\beta) \\ &P_W = W(W^TW)^{-1}W \text{ 인 투영행렬} \end{aligned} \qquad (2\text{-}3)$$

Green-Jennison-Seheult [9]는 최소자승식에서 얻어지는 투영행렬 P_W를 비모수회귀분석에서 얻어지는 평활행렬(smoothing matrix) S에 의해 대체할 수 있다고 하였고 그 경우에 추정값은 다음 식을 만족하게 된다.

$$Z^TZ\beta = Z^T(Y-Wr)$$
$$Wr = S(Y-Z\beta) \qquad (2\text{-}4)$$

Green-Jennison-Seheult [9]는 Wr의 추정치는 $\hat{m}$으로 대체하여 다음의 식 (2-5)와 같은 추정치를 제시하였다.

$$\hat{\beta}_{GJS} = (Z^T(I-S)Z)^{-1}Z^T(I-S)Y$$
$$\hat{m}_{GJS} = S(Y-Z\hat{\beta}_{GJS}) \qquad (2\text{-}5)$$

S는 $n\times n$행렬로 식 (2-1)의 모형에서 $\beta=0$으로 단순화시킨 $Y_i = m(x_i)+\varepsilon_i$ 모형으로부터 비모수회귀분석을 통해 $y=(y_1, \cdots, y_n)'$ 관찰치를 추정값인 $\hat{y}$으로 변환시켜 주는 평활행렬로서 $\hat{y} = Sy$이다. 따라서, S행렬은 평활화 방법에 따라 다른 형태를 취하게 되며 다음 장에서 여러 가지 S에 대해 설명하려 한다.

Buja [3]는 다음과 같은 가법모형에서 모수를 추정하고자 할 때 후진추정법(backfitting method)을 제시하였다.

$$Y = \sum_{j=1}^{p} f_j(x_j)+\varepsilon$$
$$E(\varepsilon) = 0, \qquad Cov(\varepsilon) = \sigma^2 I \qquad (2\text{-}6)$$

위의 모형에서 후진추정법의 과정은 다음과 같다.

초기화 : $f_i = f_i^0, \qquad i=1, \cdots, p$

순 환 : $j=1, \cdots, p, 1, 2, \cdots, p, \cdots$

$$f_j \leftarrow S_j(y-\sum f_k), \qquad k \neq j$$

반 복 : 각 함수 추정값의 변화량이 기준치보다 작을 때까지 반복한다.

수치해석(numerical analysis)에서는 Gauss-Seidel 알고리즘이라고 알려져 있다. 위의 순환과정에서 수렴은 일반적으로 즉시 이루어지지 않으나 위의 모형에 대해 Briedman-Friedman [2]은 알고리즘의 수렴을 증명한 바 있고 Bickel [1]은 완화된 가정하에서 위 알고리즘의 수렴을 증명하였다. 그러면 위의 알고리즘을 본 논문에서 주관심 모형인 공분산모형에 적용하여 보자. 공분산분석모형이므로 $f_1 = m(x)$이고 $f_2 = Z\beta$라고 할 수 있다. 우선 초기값을 $f_1^0 = 0$, $f_2^0 = 0$으로 두자. 그러면 앞의 순환과정(iteration process)은 다음과 같다.

$$f_1^{(1)} = S_1 y$$
$$f_2^{(1)} = P_1(y-f_1^{(1)})$$

$$f_1^{(2)} = S_2(y - f_2^{(1)})$$

$$\vdots$$

$$f_1^{(n)} = S_n(y - f_2^{(n-1)})$$

$$f_2^{(n)} = P_n(y - f_1^{(n)})$$

반 복 : 각 함수 추정값의 변화량이 기준치보다 작을 때까지 반복한다.

여기서 n번째 순환과정에서 $S_{(n)}$은 비모수회귀분석을 통해 얻어진 평활행렬을 나타내고 $P_{(n)}$은 일반적인 회귀분석에서 투영행렬을 나타낸다.

Ⅲ. FAN방법의 평활화 방법

모수화된 비모수부분의 추정은 아래와 같이 평활화행렬 S에 의해 얻어진다.

$$(\hat{m}(x_1), \cdots, \hat{m}(x_n))^T = Sy \tag{3-1}$$

$$S = \begin{bmatrix} w_1(x_1) & \cdots & w_n(x_1) \\ w_1(x_2) & \cdots & w_n(x_2) \\ \vdots & \cdots & \vdots \\ w_1(x_n) & \cdots & w_n(x_1) \end{bmatrix}$$

따라서, 비모수함수의 추정식 $\hat{m}(x)$는 다음 식으로 표현된다.

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) y_i \tag{3-2}$$

커널함수 $K(\cdot)$를 이용한 비모수회귀분석은 식 (3-1)에서 어떠한 $w_i(x)$를 사용하느냐에 따라 여러 가지의 추정방법들을 얻을 수 있다. 몇 가지 평활계수 h를 지닌 $w_i(x)$ 형태의 예를 들면 아래와 같은 추정량들을 만들 수 있다.

Nadaraya-Watson 추정량 $w_i(x) = \dfrac{K\left(\dfrac{x - x_i}{h}\right)}{\sum K\left(\dfrac{x - x_i}{h}\right)}$

Gässer-Müller 추정량 $w_i(x) = \int_{s_{i-1}}^{s_i} \dfrac{1}{h} K\left(\dfrac{x-u}{h}\right) du$

$$x_i \le s_i \le x_i - 1$$

Priestly-Chao 추정량 $w_i(x) = \frac{n}{h}(x_i - x_{i-1})\ K\left(\frac{x-x_i}{h}\right)$

FAN [8]은 기존의 커널함수를 이용한 회귀분석 대신 가중국소선형회귀방법을 제시하고 이 방법이 다른 전통적인 비모수회귀방법보다 Minimax등 높은 효율성을 갖는다는 것을 보였다. 본 논문에서는 부분선형모형에서 비모수부분의 추정치를 구하는데 앞에서 제시한 커널함수를 사용한 방법과 가중국소선형회귀방법을 모의실험을 통해 비교하고자 한다.

가중국소선형회귀방법은 아래 식을 최소화하는 a와 b를 찾는 것이다.

$$\sum_{i=1}^{n}(Y_i - a - b(x_i - x))^2\ K\left(\frac{x-x_i}{h}\right) \tag{3-3}$$

$\hat{a}$과 $\hat{b}$이 가중최소제곱문제의 해라고 할 때,

$$\hat{a} = \hat{m}(x) = \frac{\sum_{i=1}^{n} w_i(x)Y_i}{\sum_{i=1}^{n} w_i(x)} \tag{3-4}$$

여기서 $w_i(x) = K\left(\frac{x-x_i}{h}\right)[S_{n,2} - (x-x_i)S_{n,1}]$이며, $S_{n,l}$은 다음과 같다.

$$S_{n,l} = \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)(x-x_i)^l, \qquad l = 1,\ 2$$

모든 평활화 방법을 사용하는 데 평활계수 h의 선택은 매우 중요하다. 우리는 Craven-Wabha [4]가 제시한 일반화된 교차타당성 방법은 식 (3-5)의 일반화된 교차타당성(generalized cross-validation, GCV(h))을 최소화하는 것이다.

$$GCV(h) = \frac{RSS(h)}{(1 - n^{-1}tr(S))^2} \tag{3-5}$$

$$RSS(h) = n^{-1}\,\|y - S(h)y\|^2$$

Ⅳ. 모의실험

다음과 같은 평활함수에 의한 모의실험자료가 있다고 하자.

$$m(x) = \frac{x}{(1+x^{2/3})}$$

첫 번째 20개의 x_i값은 $N(0.25,\ 1)$에서 발생시키고 $Y_i = m(x_i) + 0.2N(0,\ 1)$로 하자. 두 번째 x_i값은 $N(-0.25,\ 1)$에서 20개를 발생시키고 $Y_i = m(x_i) + 0.2N(0,\ 1)$로 하자. 위와 같은 자료를 50번 발생시키고 커널함수는 식 (4-1)의 2차 커널함수(Epanechnikov Kernel Function)를 사용하였고 GCV에 의해 평활모수를 선택하였다. 각 방법에 의해 모수를 추정하고 후진추정법의 순환과정은 반복의 기준에 따라 각각 10번 정도 반복하였다.

$$\text{Epanechnikov} : 3/4\ (1-u^2)I(|u| \le 1) \tag{4-1}$$

$$ASE = n^{-1} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{4-2}$$

각 방법들을 비교하는 기준값은 식 (4-2)에 나타난 각 모의실험 자료의 평균추정오차(average square loss, ASE)를 평균한 값(mean ASE, MASE)을 사용하였다. 또한, 위의 모의실험에서는 실제 β값을 알고 있으므로 $\hat{\beta}$에 대한 분산과 편기의 추정값을 비교할 수 있으며 이는 〈표 1〉에서 주어졌다.

〈표 1〉 모의실험 결과

비교 기준 / 방법	GJS		Backfitting	
	$\hat{\beta}$의 Bias	MASE	$\hat{\beta}$의 Bias	MASE
Kernel	0.0282515	0.0373755	0.000689452	0.0324883
FAN	0.0269139	0.0293824	0.000745218	0.0272608

커널방법 또는 FAN방법을 사용하여도 기존의 GJS방법보다 후진방법이 편기와 평균추정오차값의 측면에서 더 나은 결과를 준다. 또한, 비모수부분을 FAN방법과 후진방법을 사용하는 것이 MASE면에서 제일 좋은 결과를 얻었다.

Ⅴ. 결론 및 요약

반모수모형들 중에서 부분선형모형은 선형모형에서 질적 변수를 포함하는 공분산분석을 일반화시킨 모형으로 고려될 수 있다. 부분선형모형을 추정하는 방법들 중

OLS개념을 도입한 기존의 GJS방법과 비교하기 위하여 본 논문에서는 비모수회귀분석에 사용되는 후진방법을 부분선형모형에 적용시키는 방법을 제시하였다. 또한, 위의 두 가지 방법에서 비모수부분에 관한 추정에서 커널추정법과 FAN이 제시한 가중국소선형회귀방법들을 사용함으로써 네 가지 방법들을 모의실험에 의한 비교를 해보았다.

모의실험의 결과는 FAN방법과 후진방법을 결합하여 사용하는 것이 추정오차 측면에서 가장 좋은 결과를 제시하였다. 위의 결과는 평활계수선택방법에 따라 다른 결과를 보여 줄 수 있으므로 여러 가지 평활계수에 관한 비교분석이 앞으로 연구해야 할 과제이다.

❖ 참 고 문 헌 ❖

1. Bickel, P., Liaassen, C., Ritov, Y. and J. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models,* To appear, 1989.
2. Briedman, L. and J. Friedman, "Estimating Optimal Transformation and Correlation," *Journal of the American Statistical Association*, 1985.
3. Buja, A., Hastie, T. and R. Tibshirani, "Linear Smoothers and Assitive Models," *Ann. Statist.* 17, 1989, pp. 453～555.
4. Craven, P. and H. Wabha, "Smoothing Noisy Data with Spline Functions," *Numer. Math*, 31, 1989, pp. 377～403.
5. Denby, L., "Smooth Regression Functions," PhD Thesis, Department of Statistics, University of Michigan, Ann Arbor, 1984.
6. _____, "Smooth Regression Functions," *Statistical Research Report*, 26, Murray Hill : At & T Bell Laboratory, 1986.
7. Engle, R., Granger, C., Rice, J. and A. Weiss, "Nonparametric Estimates of the Relation between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 1986, pp. 310～320.
8. Fan, J., "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 1992, pp. 998～1004.
9. Green, P., Jennison, C. and A. Seheult, "Analysis of Field Experiments by Least Square Smoothing," *J. R. Statistical Society*, b, 47, 1985, pp. 299～315.
10. Wabha, G., "Cross Validated Spline Methods for the Estimation of Multivariate

Functions from Data on Functionals," *Statistics : an Appraisal, Procceedings the 50th Anniversary Conference*, (eds H. A. David and H. T. David). Ames : Iowa State University Press, 1984.