Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 2, pp. 467 ~ 474

Weighted Support Vector Machines for Heteroscedastic Regression

Hye Jung Park[1] · Changha Hwang[2]

Abstract

In this paper we present a weighted support vector machine(SVM) and a weighted least squares support vector machine(LS-SVM) for the prediction in the heteroscedastic regression model. By adding weights to standard SVM and LS-SVM the better fitting ability can be achieved when errors are heteroscedastic. In the numerical studies, we illustrate the prediction performance of the proposed procedure by comparing with the procedure which combines standard SVM and LS-SVM and wild bootstrap for the prediction.

Keywords : Heteroscedasticity, Support Vector Machine, Wild Bootstrap

1. 이분산 회귀모형

선형 회귀모형에서는 일반적으로 모든 오차항이 등분산을 가진다고 가정한다. 그러나 많은 경우 독립변수들의 값에 따라 오차의 분산이 달라질 수 있다. 예를 들어 개별가구의 연간소득과 연간 소비지출에 관한 자료를 가지고 소득의 함수로서 소비지출을 설명하는 모형을 설정한다고 가정할 때 등분산의 가정이 적절하지 않다. 왜냐하면 고소득 가구보다는 저소득 가구의 소비변동량이 일반적으로 작기 때문이다. 이분산 회귀모형은 일반적으로 다음과 같이 정의된다.

$$y_i = f(\boldsymbol{x}_i) + e_i,\ E(e_i) = 0,\ E(e_i^2) = \sigma_i^2$$

여기서, y_i는 종속변수이며, $\boldsymbol{x}_i$는 입력벡터, $f(\cdot)$는 비선형 함수, e_i는 오차항, σ_i^2는 오차항의 분산이다. 오차의 분산을 알고 있는 경우 전통적인 회귀분석에서는 이

1) 경북 경산시 유곡동 290번지 대구한의대학교 모바일콘텐츠학부 객원교수
E-mail : hyjpark@dhu.ac.kr
2) 교신저자 : 서울특별시 용산구 한남동 산147번지 단국대학교 정보컴퓨터학부 교수
E-mail : chwang@dankook.ac.kr

분산성 문제를 해결하기 위해 일반적으로 가중최소제곱(WLS, Weighted Least Squares)) 또는 일반화최소제곱(GLS, Generalized Least Squares) 기법을 사용한다. 본 논문에서는 오차의 분산을 모르는 경우 이분산 비선형 회귀 함수추정 문제를 다룬다.

오차항이 등분산을 가지고 분포를 모를 경우에 붓스트랩 기반 회귀분석 기법이 함수추정을 잘 하는 것으로 알려져 있다. Freedman(1981)은 Pairwise 붓스트랩 기반 회귀분석 기법을 제안하였다. 그리고 Davidson과 Flachaire(2001)는 이분산 회귀 함수추정을 위해 Wild 붓스트랩을 제안하였다. Horowitz(1997, 2000)는 모의실험을 통해 Pairwise 붓스트랩이 이분산 회귀 함수추정에 문제가 있음을 밝혔다. Flachaire(2003)는 이분산 회귀 함수추정에서 Wild 붓스트랩이 Pairwise 붓스트랩에 비해 성능이 월등함을 모의실험을 통해 보여 주었다.

본 논문에서는 오차항의 분산과 분포를 모르는 경우 이분산 비선형 회귀 함수추정을 위해 잔차를 가중치로 사용하는 가중 SVM(Support Vector Machine)과 가중 LS-SVM(Least Squares Support Vector Machine)을 제안하고 Wild 붓스트랩 기반 이분산 회귀 함수추정 방법과 성능을 비교하고자 한다.

2. 가중 SVM 및 가중 LS-SVM

비모수 회귀함수 기법 중 SVM과 LS-SVM은 비선형 회귀를 위해 커널함수를 사용하고 있다. SVM은 원래 분류(classification)를 위해 Vapnik(1995)과 공동연구자들에 의해 개발되어 문자인식, 얼굴인식 등의 응용분야에서 좋은 결과를 보여주고 있다. 최근 SVM 이론이 회귀 함수추정으로 확장되어 많이 활용되고 있다. SVM은 투사지향(projection pursuit) 알고리즘 및 신경망과 함께 독립변수가 두 개 이상일 때 비선형 함수추정을 위해 사용되는 방법이다. SVM은 QP(quadratic programming) 문제의 어려움은 있으나, 볼록함수(convex function)를 최소화하여 학습이 진행되기 때문에 신경망과는 달리 유일한 해를 구할 수 있는 장점이 있다. SVM은 많은 응용분야에서 좋은 실험결과를 보여주어 점점 인기가 높아가고 있다. 또한, SVM의 QP문제를 해결하며 SVM과 유사한 성능을 발휘하는 LS-SVM 역시 함수추정을 위해 많이 사용되는 방법이다.

회귀분석에서 이분산성의 문제는 가중최소제곱 방법으로 해결 될 수 있다. 한편, 선형 이분산 회귀함수 추정의 경우에 WLS 방법이 OLS 방법 보다 회귀계수를 더 정확하게 추정하여 회귀계수의 표준오차가 과소 추정되는 것을 막을 수 있다는 장점을 가진다. 따라서 본 절에서는 SVM 및 LS-SVM 기법에 가중치를 부여하여 가중 SVM 및 가중 LS-SVM을 제안한다. 일반적으로 오차의 분산을 모르는 경우에 가중치로 $1/y_i^2$ 또는 $1/|r_i|$ 를 많이 사용한다. 여기서 r_i는 잔차를 나타낸다. 본 논문에서는 후자를 사용한다.

2.1 가중 SVM

가중치로 $1/|r_i|$ 를 사용할 때 가중 SVM의 최적화 문제는 다음 식과 같으며

$$\min \frac{1}{2} \| \boldsymbol{w} \|^2 + \gamma \sum_{i=1}^{n} (\theta_i \xi_i + \theta_i \xi_i^*)$$

제약조건은

$$\begin{cases} y_i - (\boldsymbol{w}' \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \le \varepsilon + \xi_i \\ (\boldsymbol{w}' \boldsymbol{\phi}(\boldsymbol{x}_i) + b) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

이다. 여기서 가중치 θ_i는 다음과 같이 정의된다.

$$\theta_i = \frac{1}{|y_i - \widehat{y}_i|}$$

한편 $\widehat{y}_i$는 SVM에 의해 추정된 y_i의 추정치이다. 위의 최적화 문제를 쌍대 문제(dual problem)로 변환하면 다음의 식과 같으며

$$L = \frac{1}{2} \| \boldsymbol{w} \|^2 + \gamma \sum_{i=1}^{n} (\theta_i \xi_i + \theta_i^* \xi_i^*) - \sum_{i=1}^{n} \alpha_i (\varepsilon + \xi_i - y_i + \boldsymbol{w}' \boldsymbol{\phi}(\boldsymbol{x}_i) + b)$$

$$- \sum_{i=1}^{n} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \boldsymbol{w}' \boldsymbol{\phi}(\boldsymbol{x}_i) - b) - \sum_{i=1}^{n} (\eta_i \xi_i + \eta_i^* \xi_i^*) \quad (1)$$

제약조건은

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0 \quad (2)$$

이 된다. 최적화를 위해 L을 변수들($\boldsymbol{w}, b, \xi_i, \xi_i^*$)로 편미분 하면 다음과 같다.

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0, \quad \frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \boldsymbol{\phi}(\boldsymbol{x}_i) = 0, \quad (3)$$

$$\frac{\partial L}{\partial \xi_i} = \gamma \theta_i - \alpha_i - \eta_i = 0, \quad \frac{\partial L}{\partial \xi_i^*} = \gamma \theta_i - \alpha_i^* - \eta_i^* = 0. \quad (4)$$

위의 식(3)~(4)을 식(1)에 대입하면 SVM의 비선형 함수추정의 해는 다음과 같이 구해진다.

$$\text{maximize} \quad -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*)$$

$$\text{subject to} \quad \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]$$

그리고 식(4)에 식(2)의 제약조건을 대입하면

$$0 \le \eta_i^* = \gamma\theta_i - \alpha_i^* \ , \ 0 \le \alpha_i^* \le \gamma\theta_i$$
$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \ , \ \alpha_i, \alpha_i^* \in [0, \gamma\theta_i]$$

따라서, 다음의 식이 유도된다.

$$\boldsymbol{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\boldsymbol{\phi}(\boldsymbol{x}_i) \ ,$$
$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

여기서 K는 커널함수를 나타내며 논문에서는 다음과 같은 RBF 커널함수를 사용한다.

$$K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}}$$

한편, 상수항 b는 다음의 KKT 조건에 의해 구해진다.

$$\alpha_i(\varepsilon + \xi_i - y_i + \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i) + b) = 0,$$
$$\alpha_i^*(\varepsilon + \xi_i^* + y_i - \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i) - b) = 0,$$
$$(\gamma - \alpha_i)\xi_i = 0, \quad (\gamma - \alpha_i^*)\xi_i^* = 0$$

2.2 가중 LS-SVM

SVM에서의 QP 문제를 극복하기 위해 일반적인 LS-SVM에 가중 SVM에서와 동일하게 가중치 $\theta_i = 1/|y_i - \widehat{y}_i|$ 를 부여한다. 이때 $\widehat{y}_i$은 LS-SVM에 의해 추정된 y_i의 추정치이다. 따라서, 최적화문제는 다음과 같이 정의되며

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{2}\sum_{i=1}^{n}\theta_i e_i^2 \ ,$$

제약조건은

$$y_i = \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i) + b + e_i, \ i = 1, \cdots, n$$

이 된다. 라그랑즈 함수는 다음과 같다.

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{2}\sum_{i=1}^{n}\theta_i e_i^2 - \sum_{i=1}^{n}\alpha_i(\boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i) + b + e_i - y_i)$$

여기서 α_i는 라그랑즈 배수를 나타낸다. 그러므로 최적화를 위한 조건식은 편미분을 통하여 간단하게 구해지며 다음과 같은 선형방정식으로 정리된다.

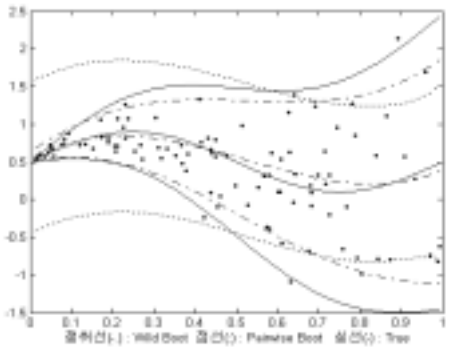
$$\begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & \boldsymbol{K}+\boldsymbol{\theta}_{\gamma} \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix}$$

여기서, $\boldsymbol{y}=(y_1,\cdots,y_n)'$, $\mathbf{1}=(1,\cdots,1)'$, $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_n)$이며, $\boldsymbol{K}=\{K_{kl}\}$는 대표 원소 K_{kl}를 갖는 행렬을 의미한다. 여기서, $K_{kl}=\boldsymbol{\phi}(\boldsymbol{x}_k)'\boldsymbol{\phi}(\boldsymbol{x}_l)=K(\boldsymbol{x}_k,\boldsymbol{x}_l)$, $k,l=1,\cdots,n$ 이고 K는 커널함수를 나타낸다. 한편 가중치 행렬 $\boldsymbol{\theta}_{\gamma}$는 다음과 같은 대각행렬로 정의된다.

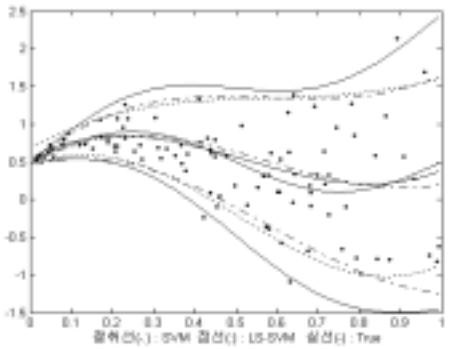
$$\boldsymbol{\theta}_{\gamma}=diag\left\{\frac{1}{\gamma\theta_1},\cdots,\frac{1}{\gamma\theta_n}\right\}$$

3. 실험 및 결과

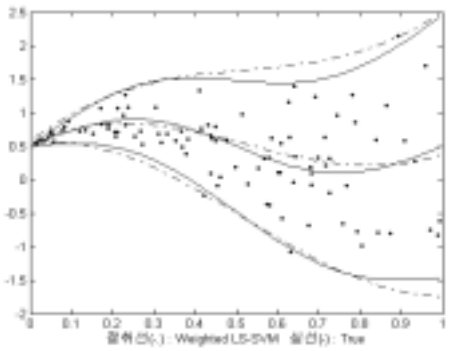
모의자료와 실제자료의 실험을 통해 논문에서 제안된 이분산 비선형 회귀함수 추정 기법들의 성능을 확인하고자 한다. 먼저 모의실험의 결과에 대해 설명한다. 모의실험에 사용된 100개 자료의 입력값 x는 균일분포 $U(0,1)$로부터 생성되었으며 출력값 y는 정규분포 $N(0.5+0.4\times\sin(2\pi x),\, x^2)$로부터 생성되었다. 모수 $\boldsymbol{\theta}=(b,\alpha_1,\cdots,\alpha_{100})'$를 추정하기 위해 자료 $\{(x_i,y_i)\}_{i=1}^{100}$과 RBF 커널이 사용되었으며, 벌칙상수 γ와 커널모수 σ의 값은 자료에 10-중 교차타당성(10-fold cross validation)을 적용하여 각각 500과 1로 결정하였다. 95% 신뢰구간 추정시 Shim과 Hwang (2003)의 방법을 이용하였으며, 성능 비교를 위해서 Flachaire(2001)가 제안한 Wild 붓스트랩 방법도 이용하였다.



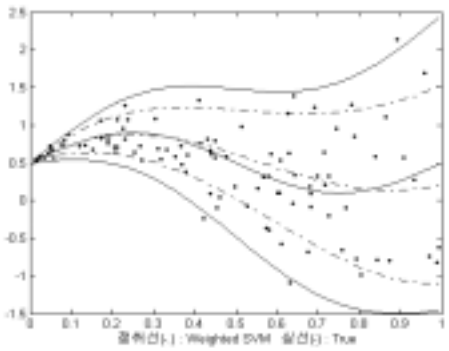
<그림 1> Wild 붓스트랩과 Pairwise 붓스트랩 비교



<그림 2> (LS-)SVM에 적용한 Wild 붓스트랩의 결과



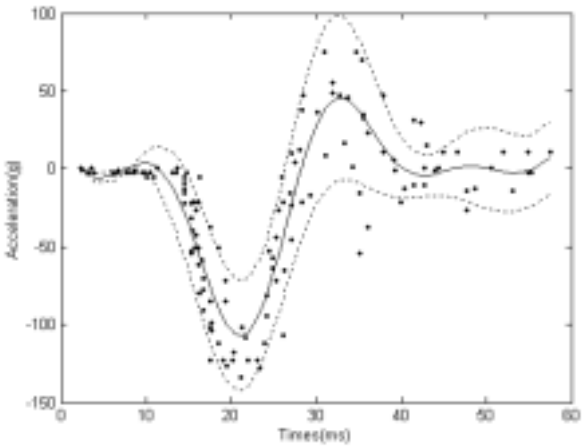
<그림 3> 가중 LS-SVM의 회귀함수 및 신뢰구간 추정



<그림 4> 가중 SVM의 회귀함수 및 신뢰구간 추정

실험에서는 먼저 Flachaire(2003)와 같이 Wild 붓스트랩과 Pairwise 붓스트랩의 성능을 비교하였다. LS-SVM에 Wild 붓스트랩 방법과 Pairwise 붓스트랩 방법을 적용하여 회귀함수를 추정하고 95% 신뢰구간을 구하였으며 실험 결과는 <그림 1>에 설명되었다. Wild 붓스트랩 방법이 Pairwise 붓스트랩 방법보다 회귀함수와 95% 신뢰구간을 더 정확하게 추정하는 것을 확인 할 수 있다. <그림 2>는 SVM과 LS-SVM에 Wild 붓스트랩 기법을 적용한 함수추정과 신뢰구간의 결과를 보여준다. SVM과 LS-SVM이 일반적으로 회귀함수를 잘 추정하는 것으로 알려져 있는데, <그림 2>는 이분산 자료에 대해서 Wild 붓스트랩 기법을 적용한 SVM과 LS-SVM이 비교적 회귀함수를 잘 추정함을 보여주며, 또한 95% 신뢰구간 추정에서도 자료의 성격에 맞게 그림 앞부분에 자료들이 조밀하게 분포하는 부분에 대해서는 신뢰구간을 좁게 추정하고 그리고 자료가 넓게 분포된 부분에 대해서는 신뢰구간을 넓게 추정하여 비교적 제대로 추정함을 보여준다.

한편 <그림 3>과 <그림 4>는 각각 가중 SVM과 가중 LS-SVM을 이용한 이분산 회귀함수 추정과 95% 신뢰구간 추정의 결과를 보여준다. 일반적으로 SVM의 단점인 QP문제를 해결하기 위해 LS-SVM를 많이 사용하는데, <그림 3>과 <그림 4>를 통해 볼 때 논문에서 제안된 가중 SVM과 가중 LS-SVM이 회귀함수와 신뢰구간의 추정에 효과가 있다. 그러나 특히 논문의 모의실험 자료에 대해서는 가중 LS-SVM이 가중 SVM 보다 회귀함수와 신뢰구간를 더 잘 추정하는 것을 확인할 수 있다. 그 이유는 논문의 모의실험 자료와 같이 출력값 y가 정규분포를 따를 때는 최소제곱법을 사용하는 LS-SVM이 최소절대편차법을 사용하는 SVM 보다 회귀함수를 더 잘 추정하기 때문인 것으로 생각된다.



<그림 5> 모터사이클 자료에 대한 가중 SVM의 결과

실제자료 실험을 위해 여러 논문에서 많이 사용되고 있는 벤치마크 자료인 모터사이클 자료(Härdle, 1990)를 사용하고 가중 SVM을 적용하였다. 모터사이클 자료는 안전 보호용 헬멧의 유효성을 연구하기 위해 시간 x 에서 발생한 가상의 모터사이클 충돌사고 후에 측정된 가속도 y와 관련된 자료이다. 10-중 교차타당성 기법을 사용해서 벌칙상수 γ와 커널모수 σ의 값을 각각 500과 1로 결정하였다. 실험결과는 <그림 5>와 같다. 가중 SVM이 전반적으로 회귀함수를 잘 추정하고, 95% 신뢰구간 추정에서도 자료의 성격에 맞게 그림 앞부분에 자료들이 조밀하게 분포하는 부분에 대해서는 신뢰구간을 좁게 추정하고 그리고 자료가 넓게 분포된 부분에 대해서는 신뢰구간을 넓게 추정하여 비교적 제대로 추정함을 보여준다.

4. 결론

본 논문에서는 오차항이 미지의 이분산을 가지는 경우 비선형 회귀함수 추정과 신뢰구간 추정을 위해 가중 SVM과 가중 LS-SVM을 제안하였다. 모의실험을 통해 제안된 가중 SVM과 가중 LS-SVM을 Wild 붓스트랩 기반 SVM 및 LS-SVM과 회귀함수 추정과 신뢰구간 추정 측면에서 비교하였다. 가중 SVM, Wild 붓스트랩 기반 SVM 및 LS-SVM가 비교적 좋은 성능을 보여주지만 특히 논문의 모의실험 자료에 대해서는 가중 LS-SVM이 전반적으로 더 좋은 성능을 보여준다. 한편, 실제자료 실험에는 가중 SVM 만을 적용하였다. 왜냐하면 등분산 또는 이분산의 경우 모두 SVM이 LS-SVM 보다 전반적으로 회귀함수를 더 잘 추정하는 것으로 알려져 있기 때문이다. 실제자료 실험에서 가중 SVM이 회귀함수와 신뢰구간을 대체로 잘 추정하는 것을 알 수 있다. 그리고 사실 가중 SVM과 가중 LS-SVM이 Wild 붓스트랩 기반 SVM 및 LS-SVM 보다 사용하기 훨씬 간단하다. 따라서 성능과 수월성을 고려할 때, 오차항이 미지의 이분산을 가지는 경우 비선형 회귀함수 추정과 신뢰구간 추정을 위해 가중 SVM과 가중 LS-SVM을 사용하는 것을 권고한다.

참고문헌

1. Davidson, R. and Flachaire, E. (2001). The wild bootstrap, tamed at last, Working paper STICERD, London School of Economics, Darp58.
2. Flachaire, E. (2003). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairwise bootstrap, EUREQUA, University Paris I Pantheon-Sorbonne.
3. Flachaire, E. (2001). Bootstrapping heteroskedasticity consistent covariance matrix estimator, Working paper 2001-80, EUREQUA, University Paris I Pantheon-Sorbonne.
4. Freedman, D. A. (1981). Bootstrapping regression models, *Annals of Statistics 9*, 1218-1228.
5. , W. (1990). *Applied Nonparametric Regression*, Econometric Society Monographs No. 19, Cambridge University Press, New York.
6. Horowitz, J. L. (1997). Bootstrap methods in econometrics: theory and numerical performance, *In advances in Economics and Econometrics: Theory and Application*, Vol 3, 188-222.
7. Horowitz, J. L. (2000). The bootstrap, *In Handbook of Econometrics*, Vol 5.
8. Shim, J. and Hwang, C. (2003). Prediction intervals for LS-SVM regression using the bootstrap, *Journal of Korean Data & Information Science Society*, Vol 14, No. 2, 337-343.
9. Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.

[2006년 3월 접수, 2006년 5월 채택]