The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.1.027
29(1), 27–39

Adaptive lasso in sparse vector autoregressive models

Sl Gi Lee[a] · Changryong Baek[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received October 27, 2015; Revised November 26, 2015; Accepted November 30, 2015)

Abstract

This paper considers variable selection in the sparse vector autoregressive (sVAR) model where sparsity comes from setting small coefficients to exact zeros. In the estimation perspective, Davis *et al.* (2015) showed that the lasso type of regularization method is successful because it provides a simultaneous variable selection and parameter estimation even for time series data. However, their simulations study reports that the regular lasso overestimates the number of non-zero coefficients, hence its finite sample performance needs improvements. In this article, we show that the adaptive lasso significantly improves the performance where the adaptive lasso finds the sparsity patterns superior to the regular lasso. Some tuning parameter selections in the adaptive lasso are also discussed from the simulations study.

Keywords: sparse vector autoregressive model, adaptive lasso, high dimensional time series

1. 서론

현대의 급격한 과학 기술의 발전은 기존에는 상상할 수조차 없는 다양하고도 대용량의 데이터를 생산해 내었다. 본 연구에서는 시간에 따라 관측된 고차원의 대용량 시계열 자료를 매우 효과적으로 분석할 수 있는 벡터자기상관회귀 모형(vector autoregressive model; VAR)의 추정을 다룬다. VAR 모형은 변수들 사이의 종속관계(interdependence)를 고려하여 시간에 따른 종속 관계(temporal dependence)를 선형 종속관계로 나타내는 모형이다. 보다 구체적으로 먼저 차원이 K인 다변량 시계열 자료 $Y_1, \ldots, Y_T$에 대해 차수 p를 갖는 VAR(p) 모형은

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + Z_t, \quad t = 1, \ldots, T \tag{1.1}$$

으로 주어진다. 여기에서 이노베이션(innovations) $\{Z_t, t = 1, \ldots, T\}$는 평균이 0이고 분산-공분산 행렬 Σ_Z를 갖는 K 차원의 i.i.d. 확률변수이다. 행렬 $A_1, \ldots, A_p$는 크기가 $K \times K$인 실수 행렬들로 AR계수를 나타낸다.

VAR 모형은 Sims (1980)를 비롯한 계량경제분야를 필두로 기상학, 환경, 금융 등에서 매우 높은 예측력을 가지는 모델임이 밝혀졌다. 하지만, 차원에 따라 모수의 숫자는 제곱함수로 증가하는 차원의 저주를 가지고 있어서 고차원 자료의 경우 추정의 어려움 뿐만 아니라 예측력의 저하와 해석의 어려움을 동

This research was supported by the Basic Science Research Program from the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1006025).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: crbaek@skku.edu

반하는 등 많은 문제를 가지고 있다. 이에 대한 한 가지 해결책으로 VAR 모형의 계수들이 0에 가까운 값을 정확하게 0으로 둠으로써 추정하여야하는 계수의 숫자를 줄이는 소위 희박벡터자기상관회귀모형(sparse VAR models; sVAR)이 높은 차원에서의 VAR 모형의 결점을 보완할 수 있는 모형으로 제안되었다.

이와 더불어 기계학습분야에서 제안된 회귀모형에서 축소방법(shrinkage method)으로 소개된 lasso는 추정계수의 크기에 제곱 벌점을 부과함으로써 변수 선택 및 모수 추정을 동시에 하는 방법으로 Tibshirani (1996)에 의해 제안되었으며 추후 후속 연구를 통해서 고차원 데이터에서도 모형 선택을 잘함이 보고되었다. 예를들어 Hsu 등 (2008), Huang 등 (2008), Hastie 등 (2015)이 있다.

따라서 이러한 흐름 속에서 Davis 등 (2015)은 시계열 모형인 sVAR에 lasso를 적용하여 모형을 추정하는 것에 대해서 연구하였다. 하지만 Davis 등 (2015)의 모의시험에 따르면 lasso 방법이 대체적으로 sVAR 모형의 계수추정에는 적합하나 0이아닌 계수의 숫자가 참값보다 훨씬 크게되는 단점이 있음을 보고하였다. 이에 따라 본 논문에서는 adaptive lasso를 사용할 경우 모형의 추정에 있어서 매우 드라마틱한 성능향상을 기대할 수 있음을 보인다. 즉 매우 큰 노이즈가 있는 sVAR 모형이나 차수가 높은 모형에서도 영이 아닌 계수를 매우 정확하게 선택함을 보인다. 또한 adaptive lasso에 필요한 튜닝 모수의 선택에 대해서도 심도 있는 논의를 한다.

이 논문은 다음과 같이 구성되어 있다. 2장에서는 벡터자기상관회귀 모형과 희박벡터자기상관회귀 모형에 대해 살펴본 후 벌점에 기반을 둔 변수 선택 방법인 lasso와 이를 계승하여 발전시킨 adaptive lasso에 대해서 살펴본다. 3장에서는 모의실험을 통해 adaptive lasso가 lasso보다 영이 아닌 계수들을 매우 정확하게 찾음을 보이고 4장에서는 튜닝 모수의 선택에 대해서 논의하며 마지막으로 5장에서는 결론을 다루었다.

2. 벡터자기회귀 모형의 계수 추정 방법

2.1. 최소자승법(OLS), 최대우도추정량(MLE) 및 릿지 추정량(Ridge estimator)

본 장에서는 차원이 K이고 차수가 p인 VAR(p) 모형 (1.1)의 AR계수 $A_1, \ldots, A_p$의 추정에 대해서 간략하게 소개한다. 우선 다변량 시계열 자료 $Y_1, \ldots, Y_T$에 대해서 $\{Y_t\}$는 인과과정(causal process)임을 가정하며 $\{Z_t\}$는 $\{Y_s, s < t\}$와 독립임을 가정한다. VAR(p) 모형의 추정을 위하여 Lütkepohl (2005)에 따라 다음과 같이 모형을 다시 쓸 수 있다.

$$Y = AL + Z. \tag{2.1}$$

여기에서

$$\begin{aligned} &Y := (Y_1, Y_2, \ldots, Y_T), \quad A := (A_1, A_2, \ldots, A_p), \\ &L_t := \mathrm{vec}(Y_t, Y_{t-1}, \ldots, Y_{t-p+1}), \quad L := (L_0, L_1, \ldots, L_{T-1}), \quad Z := (Z_1, Z_2, \ldots, Z_T) \end{aligned}$$

이며 $Y_{-p+1}, \ldots, Y_0$는 0이다. 수식 (2.1)을 다시 벡터형식으로 적으면 다음과 같다.

$$y := \mathrm{vec}(Y) = \left(L' \otimes I_K\right) \alpha + \mathrm{vec}(Z), \quad \alpha := \mathrm{vec}(A) = \mathrm{vec}\left(A_1, A_2, \ldots, A_p\right). \tag{2.2}$$

따라서 최소자승법에 기반을 둔 추정량은(OLS)

$$\hat{\alpha}^{OLS} = \underset{\alpha}{\mathrm{argmin}} \|y - (L' \otimes I_K)\alpha\|^2 = \left((LL')^{-1} L \otimes I_K\right) y$$

이고 $\|x\| := \sqrt{x_1^2 + \cdots + x_n^2}$으로 정의된 노음이다. 이노베이션 $\{Z_t\}$에 대한 분산-공분산 행렬의 추정량은 다음과 같다.

$$\hat{\Sigma}_Z^{OLS} = \frac{1}{T-p}\sum_{t=p+1}^{T}\left(Y_t - \widehat{Y}_t\right)\left(Y_t - \widehat{Y}_t\right)', \qquad \widehat{Y}_t = \hat{A}_1^{OLS}Y_{t-1} + \cdots + \hat{A}_p^{OLS}Y_{t-p}. \tag{2.3}$$

이노베이션 $\{Z_t\}$에 대해서 다변량 정규분포를 가정하면 최대우도추정량(MLE)를 찾을 수 있다. 가능도함수는

$$-\frac{1}{2}\log|2\pi(I_K \otimes \Sigma_Z)| - \frac{1}{2}\left(y - (L' \otimes I_K)\alpha\right)'\left(I_T \otimes \Sigma_Z^{-1}\right)\left(y - (L' \otimes I_K)\alpha\right)$$

이며 최대우도추정량은

$$\hat{\alpha}^{MLE} = \left(LL' \otimes \Sigma_Z^{-1}\right)^{-1}\left(L \otimes \Sigma_Z^{-1}\right)y = \left(\left(LL'\right)^{-1}L \otimes I_K\right)y$$

으로 OLS 추정량과 같아지지만 분산-공분산 추정량은 다음과 같다.

$$\hat{\Sigma}_Z^{MLE} = \frac{1}{T}\sum_{t=p+1}^{T}\left(Y_t - \widehat{Y}_t\right)\left(Y_t - \widehat{Y}_t\right)', \qquad \widehat{Y}_t = \hat{A}_1^{MLE}Y_{t-1} + \cdots + \hat{A}_p^{MLE}Y_{t-p}. \tag{2.4}$$

회귀분석에서 쓰이는 대표적인 축소방법(shrinkage method)인 릿지 추정량(Ridge estimator)이용한 벡터자기회귀 모형의 계수 추정은

$$\hat{\alpha}^{ridge} = \underset{\alpha}{\operatorname{argmin}}\left\{\|y - \left(L' \otimes I_K\right)\alpha\|^2 + \lambda\|\alpha\|^2\right\} = (LL' \otimes I_K + \lambda I_{K^2p})^{-1}(L \otimes I_K)y$$

으로 튜닝모수 λ가 증가함에 따라서 축소의 정도가 심해진다. 이노베이션에 대한 분산-공분산 추정량은

$$\hat{\Sigma}_Z^{rigde} = \frac{1}{T-p}\sum_{t=p+1}^{T}\left(Y_t - \widehat{Y}_t\right)\left(Y_t - \widehat{Y}_t\right)', \qquad \widehat{Y}_t = \hat{A}_1^{ridge}Y_{t-1} + \cdots + \hat{A}_p^{ridge}Y_{t-p} \tag{2.5}$$

이다.

2.2. 벌점화방식의 lasso 및 adaptive lasso

벌점화 방식인 lasso는 Tibshirani (1996)에 의해 제안된 방법으로 릿지 추정량이 계수의 축소만을 고려한 것에 비해서 추정 모수에 ℓ_1 벌점함수, 즉 $\|x\|_1 = |x_1| + \cdots + |x_n|$를 고려하여 변수의 선택과 축소를 동시에 수행하는 고차원 자료를 다루는데 있어서 매우 획기적이고도 중요한 방법이다. Davis 등 (2015)는 노이즈에 대한 의존성을 고려한 벡터자기회귀 모형에서의 lasso 추정량을 다음과 같이 정의하였다.

$$\begin{aligned}\widehat{\alpha}^{lasso} &:= \underset{\alpha}{\operatorname{argmin}}\, Q_\lambda(\alpha, \Sigma_Z) \\ &= \underset{\alpha}{\operatorname{argmin}}\left\{T\log|\Sigma_Z| + \left\|\left(I_T \otimes \Sigma_Z^{-\frac{1}{2}}\right)y - \left(L' \otimes \Sigma_Z^{-\frac{1}{2}}\right)\alpha\right\|^2 + \lambda\|\alpha\|_1\right\}.\end{aligned} \tag{2.6}$$

모수의 추정은 극좌표 하강 알고리즘(coordinate descent algorithm)에 기반하여 10-fold CV(cross-validation)로 튜닝모수 λ를 추정하고 분산 공분산 행렬 Σ_Z과 모수 α를 반복적으로 업데이트 하는 다음의 알고리즘을 사용한다.

Lasso를 이용한 반복적 VAR 모형 추정법

1. 분산 공분산 행렬의 초기값 $\Sigma_Z^{(0)}$을 설정.
2. 모수 α와 분산 공분산 행렬 Σ_Z을 수렴할 때까지 아래와 같은 방법으로 반복.
   2.1. 극좌표 하강 알고리즘 및 10-fold CV를 통한 튜닝모수 λ 선택을 통해 추정량 계산.

$$\alpha^{(k+1)} = \underset{\alpha}{\text{argmin}} Q_\lambda \left(\alpha, \Sigma_Z^{(k)}\right).$$

   2.2. $\Sigma_Z^{(k+1)} = (1/T)(Y - A^{(k+1)}L)(Y - A^{(k+1)}L)'$, $\alpha^{(k+1)} = \text{vec}(A^{(k+1)})$.

VAR 모형에서의 lasso 방법론에 대한 연구로는 대표적으로 Hsu 등 (2008), Song과 Bickel (2011)이 있으며 최근 활발한 연구가 진행되고 있다. 하지만, 모의실험을 통해 Davis 등 (2015)은 lasso를 이용한 희박 벡터자기상관회귀 계수 추정이 참값보다 훨씬 더 많은 영이 아닌 계수를 추정하는 단점을 지적하였다. 이는 과거 Arnold 등 (2008), Lozano 등 (2009)이 lasso를 이용한 AR 모형 추정이 실제보다 과도한 차수를 추정하는 경향이 있다고 밝힌 것과 그 맥락을 같이한다. 보다 근본적으로 i.i.d. 회귀 모형 가정 하에서 Zou (2006)는 lasso 추정법이 변수 선택 일치성과 점근적 정규성을 보장 할 수 없음을 밝혔고 이러한 단점을 보완하기 위해서 adaptive lasso 추정량을 소개하였다. 그 아이디어는 작은 추정값을 가지는 계수에 대해서 더 많은 가중벌점을 주어서 변수가 선택되지 못하게 하는 것이다. 따라서 본 논문은 adaptive lasso를 이용하여 희박벡터모형을 추정하였을 경우에 어떠한 성능향상을 기대할 수 있는지 모의실험을 통해서 밝히고자 한다. 구체적으로 adaptive lasso는 다음과 같이 정의된다.

$$\begin{aligned}\widehat{\alpha}^{al} &:= \underset{\alpha}{\text{argmin}}\ Q_\lambda^{al}(\alpha, \Sigma_Z) \\ &= \underset{\alpha}{\text{argmin}} \left\{ T \log|\Sigma_Z| + \left\| \left(I_T \otimes \Sigma_Z^{-\frac{1}{2}} \right) y - \left(L' \otimes \Sigma_Z^{-\frac{1}{2}} \right) \alpha \right\|^2 + \lambda \|w'\alpha\|_1 \right\}, \end{aligned} \tag{2.7}$$

여기에서 w는 초기 추정량의 역수로 이루어진 가중치 벡터로 j번째 가중치는

$$w_j = \frac{1}{|\widehat{\alpha}_j|^\gamma}, \quad \gamma > 0 \tag{2.8}$$

으로 주어진다. 따라서 작은 계수값에 많은 벌점을 부과하여 lasso 모형보다 더 희박한 모형을 선택하게 된다. 다음은 adaptive lasso 추정량을 계산하는 반복알고리즘을 정리한 것이다.

adaptive lasso를 이용한 반복적 VAR 모형 추정법

1. 주어진 γ에 대해서 분산 공분산 행렬의 초기값 $\Sigma_Z^{(0)}$ 및 가중치 $w^{(0)}$을 설정.
2. 모수 α와 분산 공분산 행렬 Σ_Z을 수렴할 때까지 아래와 같은 방법으로 반복.
   2.1. 극좌표 하강 알고리즘 및 10-fold CV를 통한 튜닝모수 λ 선택을 통해 추정량 계산.

$$\alpha^{(k+1)} = \underset{\alpha}{\text{argmin}}\ Q_\lambda^{al}(\alpha, \Sigma_Z).$$

   2.2. $\Sigma_Z^{(k+1)} = (1/T)(Y - A^{(k+1)}L)(Y - A^{(k+1)}L)'$, $\alpha^{(k+1)} = \text{vec}(A^{(k+1)})$.
   2.3. $w^{(k+1)} = 1/|\alpha^{(k+1)}|^\gamma$.

3. 모의실험

본 장에서는 adaptive lasso 방법을 이용하여 희박벡터상관회귀 모형을 추정하였을 때 어떠한 성능을 보

이는지에 대한 모의실험 결과를 보고한다. 본 모의실험에는 다음의 두 가지 자료생성과정(Data generating process)을 사용하였다. 첫 번째 DGP(DGP1)는 VAR(1) 모형에 여섯 개의 영이 아닌 계수를 가지는 sVAR(1; 6) 모형으로 모형식은

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \\ X_{t,4} \\ X_{t,5} \\ X_{t,6} \end{pmatrix} = \begin{pmatrix} .8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -.3 & 0 \\ .6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .8 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \\ X_{t-1,4} \\ X_{t-1,5} \\ X_{t-1,6} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \\ Z_{t,4} \\ Z_{t,5} \\ Z_{t,6} \end{pmatrix} \tag{3.1}$$

이며, 두 번째 DGP(DGP2)는 VAR(2) 모형에 영이 아닌 계수가 12개인 sVAR(2; 12) 모형으로

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \\ X_{t,4} \\ X_{t,5} \\ X_{t,6} \end{pmatrix} = \begin{pmatrix} .8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -.3 & 0 \\ .6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .8 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \\ X_{t-1,4} \\ X_{t-1,5} \\ X_{t-1,6} \end{pmatrix} + \begin{pmatrix} .2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .3 \\ -.3 & 0 & 0 & 0 & 0 & 0 \\ .6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .6 & 0 & 0 & 0 \\ 0 & 0 & .4 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_{t-2,1} \\ X_{t-2,2} \\ X_{t-2,3} \\ X_{t-2,4} \\ X_{t-2,5} \\ X_{t-2,6} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \\ Z_{t,4} \\ Z_{t,5} \\ Z_{t,6} \end{pmatrix} \tag{3.2}$$

이다. 여기에서 이노베이션 벡터인 $(Z_{t,1}, \ldots, Z_{t,6})'$는 평균이 $(0,0,0,0,0,0)'$이고 분산 공분산 행렬 Σ_Z은

$$\Sigma_Z = \begin{pmatrix} \delta^2 & \delta/4 & \delta/6 & \delta/8 & \delta/10 & \delta/12 \\ \delta/4 & 1 & 0 & 0 & 0 & 0 \\ \delta/6 & 0 & 1 & 0 & 0 & 0 \\ \delta/8 & 0 & 0 & 1 & 0 & 0 \\ \delta/10 & 0 & 0 & 0 & 1 & 0 \\ \delta/12 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.3}$$

으로 주어진 다변량 정규분포를 따른다고 가정하였다.

모의실험 결과 각 측도의 변수선택 성능을 요약하기 위한 통계량으로는 RMSE(root mean square error), 영이 아닌 계수의 수, MSP(mean squared proportion)를 고려하였다. 우선, RMSE는 추정량의 불일치도를 나타내기 위한 통계량으로 다음과 같이 정의된다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \text{tr}\left(\left(A - \widehat{A}^{(i)}\right)'\left(A - \widehat{A}^{(i)}\right)\right)},$$

여기에서 n은 반복수이고 $\widehat{A}^{(i)}$은 i번째 반복에 대한 sVAR 모형 추정 계수이다. 다음으로 영이 아닌 계수의 수는 전체 모의실험 중 추정된 계수 행렬에 영이 아닌 계수의 수를 평균값으로 나타낸 값이다. 마지막으로, MSP 정의에 앞서 영이 아닌 계수에 대한 지시함수 $M_k(i,j)$을 정의하면 다음과 같다.

$$M_k(i,j) := \begin{cases} 1, & \text{if } A_k(i,j) \text{ is non-zero,} \\ 0, & \text{otherwise.} \end{cases}$$

축도성능 요약 통계량인 MSP은 다음과 같이 정의된다.

$$\text{MSP} = \frac{1}{|\mathcal{I}|} \sum_{i,j,k \in \mathcal{I}} \left(M_k(i,j) - \widehat{M}_k(i,j) \right)^2,$$

여기에서 $\widehat{M}_k(i,j)$는 k번째 AR 계수의 i,j번째 원소인 $A_k(i,j)$를 영이 아닌 것으로 추정하는 상대도수를 의미하며 인덱스집합 $\mathcal{I} = \{(k,i,j)|k = 1,\ldots,p, i,j \in \{1,\ldots,K\}\}$이다. MSP 값이 0에 가까울수록 좋은 성능을 나타내며 값이 클수록 추정 성능이 좋지 않음을 나타낸다.

이번 모의실험에서는 다음의 7가지 방법에 대해서 비교를 하였다. 먼저 i.i.d. 가정에서 출발한 lasso 및 adaptive lasso(al) 방법으로 각각 수식 (2.6)과 (2.7)에서 Σ_Z를 I_K로 대체한 방법이다. 하지만 시계열 모형에서는 i.i.d. 가정을 하지 않으므로 Davis 등 (2015)에서 제안한 분산 공분산 행렬 업데이트 방법을 적용한 lasso 방법을 토대로한 (2.6) 방법을 적용하였다. 이는 adaptive lasso에 의한 추정 성능의 향상인지 혹은 노이즈 벡터의 분산-공분산을 고려하였기때문에 얻어지는 성능 향상인지를 구별하기 위해 고안한 실험이다.

앞서 2.2장에서 설명하였듯이 adaptive lasso를 반복적 알고리즘을 통해 추정하기 위해서는 분산-공분산 행렬의 초기값과 가중치가 벡터 (2.8)이 필요하다. 본 실험에서는 최소자승추정값(al-OSL), 최대우도추정량(al-MLE), i.i.d. 가정하의 lasso 추정량(al-Lasso), 릿지 추정량(al-Ridge) 네 가지 방법을 통해 얻어진 초기 추정값에 대해서 얻어진 분산-공분산 행렬 추정량 (2.3)–(2.5)을 사용하였다. 릿지 추정량에서의 튜닝 모수 λ의 추정은 Cule과 De Iorio (2013)의 방법을 따랐다.

제 3장의 구성은 adaptive lasso 방법에 대한 추정 성능은 3.1절에서 살펴보며 표본크기에 대한 효과는 3.2절에서 다루며 adaptive lasso에 필요한 튜닝모수 γ에 대한 효과는 3.3절에서 살펴본다. 모든 모의실험 결과는 총 500번의 반복을 통해 산출하였다.

3.1. adaptive lasso의 성능

본 절에서는 adaptive lasso의 추정 성능을 알아보기 위해서 튜닝 모수 $\gamma = 1$ 및 표본 크기는 $T = 1000$에 대해서 위에서 제시한 7가지 방법을 두 가지 DGP 모형에 적용한 결과를 보고한다. 또한 이노베이션의 분산-공분산의 노이즈 정도에 따른 성능 차이를 보기 위해서 수식 (3.3)에서 모수 $\delta = 1, 5, 10$ 세 가지 경우에 대해서 결과를 산출하였다.

Table 3.1은 첫 번째 DGP 모형인 sVAR(1; 6)에 대한 결과이다. 첫 네 열은 분산-공분산 행렬을 업데이트하는 알고리듬을 사용한 adaptive lasso 방법에서 초기값(OLS, MLE, Lasso, Ridge)에 따라 그 결과를 정리한 것이고, 다섯 번 째 열의 Lasso는 Davis 등 (2015)에서 사용한 분산-공분산 행렬을 업데이트하는 lasso 방법을 나타낸다. 마지막 두개 열은 이노베이션 공분산에 대해서 i.i.d. 가정, 즉 $\Sigma_Z = I_K$로 가정한 Zou (2006)의 adaptive lasso(al) 및 lasso 방법론(Lasso)을 의미한다. 먼저 adaptive lasso 방법이 lasso 방법에 비해서 작은 RMSE, 영이 아닌 계수의 참값인 0에 훨씬 더 가까운 값을 주며 MSP가 급격하게 작아짐을 볼 수 있다. 분산-공분산 행렬을 고려하지 않다 할지라도 adaptive lasso 방법은 lasso 방법보다 훨씬 더 좋은 성능을 보임을 알 수 있어, 본 실험을 통해서 adaptive lasso가 희박벡터상관회귀 모형의 추정에 있어서 매우 좋은 성능을 보임을 알 수 있다. 하지만 노이즈 정도인 δ가 커지면 분산-공분산 행렬을 고려한 방법이 그렇지 않은 adaptive lasso보다 더 좋은 성능을 보임을 알 수 있다.

초기값에 대한 효과는 릿지 추정량을 제외하고는 대부분 비슷한 성능을 보이고 있다. 릿지 추정량이 다중공선성을 가지는 공변량에 대한 좋은 추정량이기에 VAR 모형에서 좀 더 자연스러운 추정량이라고 생각하였고 또한 Zhang 등 (2008) 등에서는 릿지 추정량이 다차원 시계열의 추정에 있어서는 좋은 이론

Table 3.1. Performance of adaptive lasso with initial estimator from regular lasso for DGP1

δ		Sigma update					$\Sigma_Z = I_K$	
		al(OLS)	al(MLE)	al(Lasso)	al(Ridge)	Lasso	al	Lasso
1	RMSE	0.071	0.070	0.069	0.267	0.120	0.069	0.121
	Non-zero coef	6.602	6.644	6.116	6.550	15.240	6.124	15.758
	MSP∗100	0.038	0.043	0.002	0.038	8.420	0.002	8.898
5	RMSE	0.055	0.055	0.055	0.178	0.089	0.136	0.207
	Non-zero coef	6.050	6.050	6.012	6.060	17.152	6.784	12.076
	MSP∗100	0.001	0.001	0.000	0.005	16.643	0.347	6.213
10	RMSE	0.055	0.055	0.073	0.203	0.074	0.315	0.394
	Non-zero coef	6.000	6.000	5.976	5.966	18.900	6.434	10.394
	MSP∗100	0.000	0.000	0.002	0.006	22.857	1.569	5.386

DGP = data generating process, al = adaptive lasso, OLS = ordinary least squares, MLE = maximum likelihood estimation, RMSE = root mean square error, MSP = mean squared proportion.

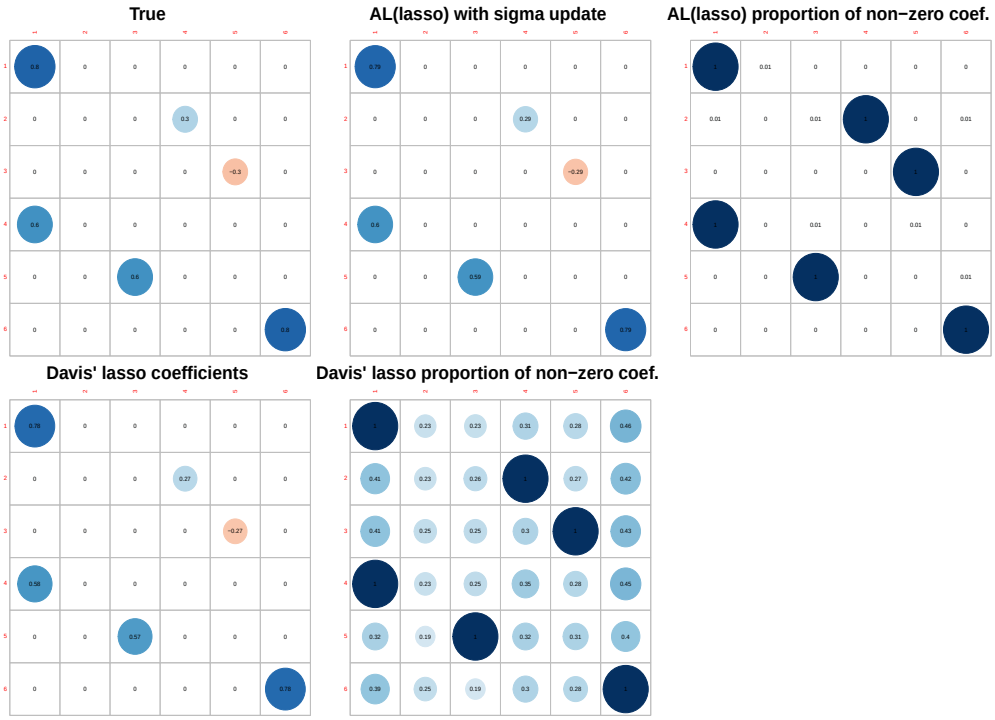

Figure 3.1. The comparison betwen adaptive lasso with initial estimator from regular lasso and Davis' method for DGP1.

적인 성질을 가지고 있음을 보였지만, 이번 모의실험에서는 릿지 추정량이 예상만큼 좋은 성능을 보이지는 못해 추가 연구가 필요할 것으로 본다.

이러한 adaptive lasso의 좋은 성능은 Figure 3.1에서 더 쉽게 볼 수 있다. 표본 크기 $T = 1000$ 그리고 $\delta = 1$에 대해서 i.i.d. 가정하의 lasso 추정량을 초기값으로 사용한 adaptive lasso 추정방법과 Davis

Table 3.2. Performance of adaptive lasso with initial estimator from regular lasso for DGP2

δ		Sigma update					$\Sigma_Z = I_K$	
		al(OLS)	al(MLE)	al(Lasso)	al(Ridge)	Lasso	al	Lasso
1	RMSE	0.069	0.069	0.073	0.511	0.183	0.078	0.155
	Non-zero coef	10.996	10.992	10.736	13.904	29.992	10.626	33.930
	MSP∗100	0.020	0.020	0.047	0.565	11.533	0.107	13.914
5	RMSE	0.265	0.270	0.373	0.474	1.003	0.347	0.180
	Non-zero coef	8.252	8.232	8.644	14.430	22.906	8.976	25.518
	MSP∗100	2.890	2.922	5.483	3.495	14.604	4.578	13.913
10	RMSE	0.301	0.354	0.367	0.594	2.165	0.423	0.276
	Non-zero coef	8.200	8.074	8.040	16.514	14.838	8.548	24.886
	MSP∗100	3.207	3.354	7.179	6.502	14.728	6.548	17.701

DGP = data generating process, al = adaptive lasso, OLS = ordinary least squares, MLE = maximum likelihood estimation, RMSE = root mean square error, MSP = mean squared proportion.

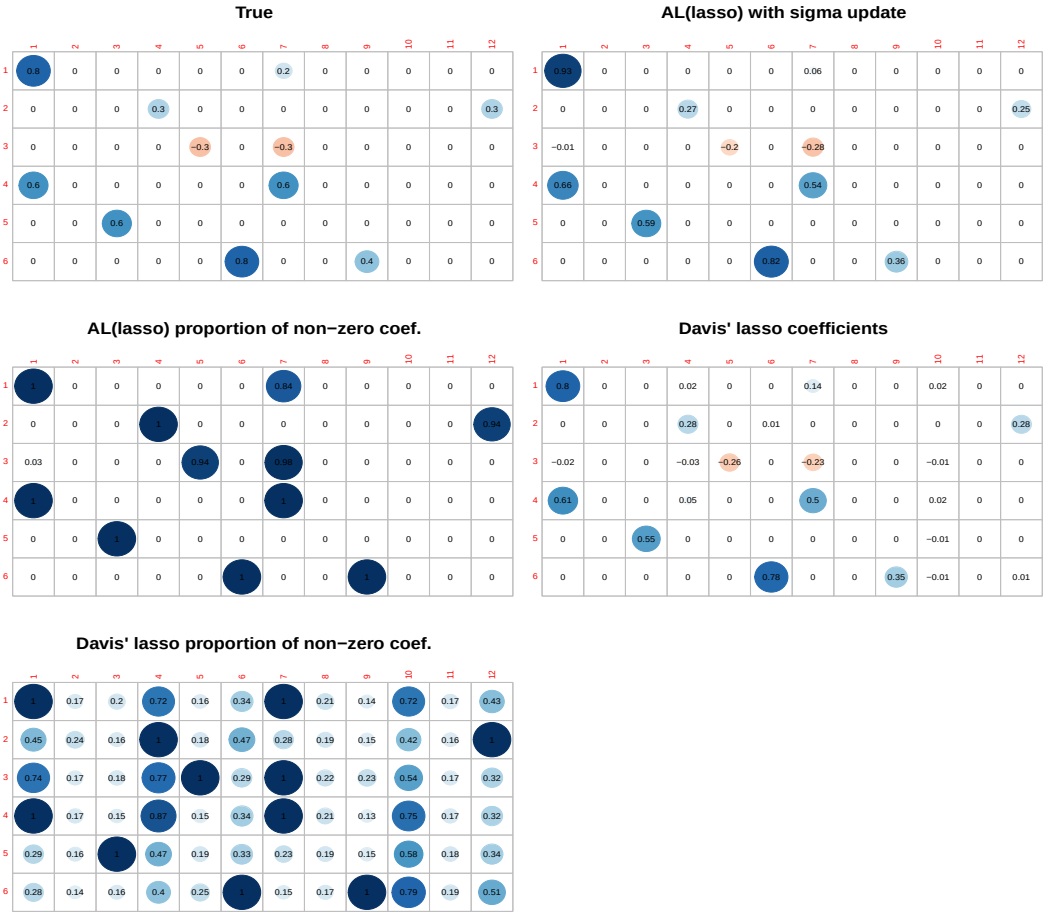

Figure 3.2. The comparison between adaptive lasso with initial estimator from regular lasso and Davis' method for DGP2.

Table 3.3. The effect of sample size for DGP1 with $\delta = 5$.

T		Sigma update					$\Sigma_Z = I_K$	
		al(OLS)	al(MLE)	al(Lasso)	al(Ridge)	Lasso	al	Lasso
200	RMSE	0.224	0.231	0.188	0.456	0.230	0.485	0.539
	Non-zero coef	9.180	9.256	6.666	8.578	16.976	6.854	12.504
	MSP∗100	0.967	1.012	0.091	0.787	16.150	1.966	6.917
500	RMSE	0.082	0.082	0.077	0.197	0.120	0.244	0.313
	Non-zero coef	6.626	6.654	6.134	6.656	17.152	6.962	12.088
	MSP∗100	0.047	0.050	0.004	0.063	16.690	0.786	6.260
1000	RMSE	0.055	0.055	0.055	0.178	0.089	0.136	0.207
	Non-zero coef	6.050	6.050	6.012	6.060	17.152	6.784	12.076
	MSP∗100	0.001	0.001	0.000	0.005	16.643	0.347	6.213

DGP = data generating process, al = adaptive lasso, OLS = ordinary least squares, MLE = maximum likelihood estimation, RMSE = root mean square error, MSP = mean squared proportion.

등 (2015)에서 사용한 분산-공분산 행렬을 고려한 lasso 추정량에 대해서 추정값에 대한 결과를 요약하였다. 상단 왼쪽은 참값으로 AR계수의 값과 영이 아닌 계수의 위치를 나타낸다. 그리고 상단 중간은 adaptive lasso에 의해서 얻어진 추정값들의 평균을 나타내며, 상단 오른쪽 패널은 영이 아닌 계수들의 빈도수를 나타낸다. 아래 위치한 그림들은 Davis 등 (2015)에서 사용한 분산-공분산 행렬을 고려한 lasso 추정량에 대해서 추정값들의 평균과 영이 아닌 계수들의 빈도수를 나타낸다. 하단의 그림에서 볼 수 있듯이 lasso 방법의 경우 0으로 추정해야 할 위치에도 영이 아닌 값으로 추정하여 과대 추정하는 경향이 있음을 확인할 수 있다. 하지만 adaptive lasso를 사용할 경우 추정량 및 영이 아닌 계수의 위치 모두 참값과 매우 가깝게 추정함을 알 수 있다.

두 번째 DGP에 대한 결과는 Table 3.2와 Figure 3.2에서 찾아볼 수 있다. 첫 번째 실험 결과와 비슷하게 sVAR(2; 12)으로 AR의 차수가 높은 복잡한 모형에서도 adaptive lasso가 lasso 방법과 비교하여 훨씬 더 좋은 성능을 보임을 확인할 수 있다. 다만 복잡한 모형의 경우 또한 노이즈 정도인 δ의 값이 높아질수록 adaptive lasso 뿐만 아니라 lasso 방법이 좀 더 희박한 모형을 찾는 것은 흥미로운 사실로 이 부분에 대한 추후 연구가 필요하다고 판단된다.

3.2. 표본크기에 따른 성능 비교

본 논문에서 고려한 adaptive lasso의 성능이 표본 크기에 따라 어떻게 변화하는지에 대해서 알아보기 위해서 본 절에서는 adaptive lasso의 튜닝 모수 $\gamma = 1$에 대해서 이노베이션의 분산-공분산 모수 δ값을 5로 고정하고, 표본수가 200, 500, 1000으로 증가함에 따라 adaptive lasso 추정의 성능을 비교하였다.

Table 3.3는 DGP1에 대한 결과이다. 먼저 작은 표본수인 $T = 200$을 비롯한 본 실험에서 고려한 모든 경우에 대해서 adaptive lasso가 lasso 방법을 개선시키며 그 성능 또한 만족스러움을 볼 수 있다. 또한, 표본의 크기가 증가함에따라 RMSE를 비롯한 성능측도가 감소하는 추세를 볼 수 있다. 또한 초기값의 추정의 경우 릿지 추정량을 제외하고서는 그 우열을 가리기 힘드나 i.i.d.을 가정한 lasso 추정량이 모든 경우에서 근소하나마 가장 좋은 성능을 보였다. DGP2에 대한 결과는 Table 3.4에 요약되어 있다. DGP1과 같이 adaptive lasso가 lasso 방법보다 더 좋은 결과를 주었으며 OLS를 이용한 초기 추정값이 가장 좋은 결과를 주었다. DGP1과 비교하여 모형이 복잡해짐에 따라 RMSE를 비롯한 성능측도들이 감소하는 추세를 보여주지는 못하였지만 표본이 증가할수록 더 희박한 모형을 찾는 경향이 있었다. 이는 lasso 및 adaptive lasso 모두 가지고 있는 성질로 추가 연구가 필요한 흥미로운 점으로 보인다.

Table 3.4. The effect of sample size for DGP2 with $\delta = 5$.

T		sigma update					$\Sigma_Z = I_K$	
		al(OLS)	al(MLE)	al(Lasso)	al(Ridge)	Lasso	al	Lasso
200	RMSE	0.155	0.169	0.199	0.759	0.397	0.382	0.437
	Non-zero coef	12.680	12.268	10.456	17.898	28.764	11.918	26.520
	MSP∗100	1.012	1.059	1.265	3.171	13.289	1.180	9.380
500	RMSE	0.197	0.205	0.305	0.579	0.823	0.293	0.212
	Non-zero coef	9.288	9.158	9.436	14.818	22.348	10.270	25.036
	MSP∗100	1.579	1.705	3.452	2.609	12.029	2.308	10.219
1000	RMSE	0.265	0.270	0.373	0.474	1.003	0.347	0.180
	Non-zero coef	8.252	8.232	8.644	14.430	22.906	8.976	25.518
	MSP∗100	2.890	2.922	5.483	3.495	14.604	4.578	13.913

DGP = data generating process, al = adaptive lasso, OLS = ordinary least squares, MLE = maximum likelihood estimation, RMSE = root mean square error, MSP = mean squared proportion.

Table 3.5. The effect of tuning parameter γ in adaptive lasso for DGP1 with δ=1 and T=500.

γ		sigma update				$\Sigma_Z = I_K$
		al(OLS)	al(MLE)	al(Lasso)	al(Ridge)	al
0.5	RMSE	0.124	0.124	0.115	0.264	0.118
	Non-zero coef	10.414	10.436	7.980	10.074	7.866
	MSP∗100	1.889	1.925	0.382	1.664	0.342
1	RMSE	0.105	0.104	0.099	0.391	0.098
	Non-zero coef	8.162	7.982	6.620	7.610	6.530
	MSP∗100	0.452	0.381	0.038	0.300	0.029
1.5	RMSE	0.097	0.097	0.097	0.370	0.097
	Non-zero coef	6.221	6.222	6.025	6.025	6.022
	MSP∗100	0.006	0.006	0.000	0.017	0.000
2	RMSE	0.100	0.101	0.104	0.568	0.105
	Non-zero coef	6.003	6.003	6.000	5.697	6.000
	MSP∗100	0.000	0.000	0.000	0.051	0.000

DGP = data generating process, al = adaptive lasso, OLS = ordinary least squares, MLE = maximum likelihood estimation, RMSE = root mean square error, MSP = mean squared proportion.

3.3. 튜닝모수 γ에 대한 성능 비교

앞의 두 모의실험 결과를 통해 adaptive lasso가 lasso 보다 더 나은 성능을 보임을 알 수 있었다. 하지만 adaptive lasso 역시 튜닝 모수인 γ 값에 의존하므로 적절한 튜닝 모수를 선택하는 게 중요하다. 따라서 분산 공분산 행렬의 의존구도를 결정하는 모수인 δ 값을 1로, 표본크기 T를 500으로 고정 한 뒤, adaptive lasso 가중치 항에 적용되는 튜닝 모수인 γ의 선택에 대한 성능 차이를 비교해 보았다. 사용되는 튜닝 모수는 $\gamma = 0.5, 1, 1.5, 2$이다.

Table 3.5는 DGP1에 대한 결과이다. 먼저 분산-공분산 행렬을 업데이트하지 않은 i.i.d. 를 가정한 adaptive lasso의 경우 γ 값에 크게 상관없이 RMSE나 영이 아닌 계수의 개수를 추정함을 알 수 있다. 하지만 adaptive lasso의 경우 RMSE에 대해서는 로버스트한 값을 주었으나 영이 아닌 모수의 개수는 γ의 값이 증가할수록 영이 아닌 계수의 수가 감소하여 더 희박한 모형을 추정함을 알 수 있다. 특히, γ 값이 1 이상 값을 가질 때 영이 아닌 계수의 평균도 실제값인 6과 가깝고 RMSE와 MSP값도 낮아 높은 성능을 보임을 알 수 있으며 튜닝 모수 $\gamma = 1$, $\gamma = 1.5$, $\gamma = 2$로 증가하더라도 그 성능의 차이

Table 3.6. The effect of tuning parameter γ in adaptive lasso for DGP2 with δ=1 and T=500.

γ		sigma update				$\Sigma_Z = I_K$
		al(OLS)	al(MLE)	al(Lasso)	al(Ridge)	al
0.5	RMSE	0.137	0.136	0.111	0.283	0.115
	Non-zero coef	18.928	18.754	14.606	20.940	14.644
	MSP∗100	1.615	1.544	0.344	2.752	0.352
1	RMSE	0.265	0.270	0.373	0.474	1.003
	Non-zero coef	8.252	8.232	8.644	14.430	22.906
	MSP∗100	2.890	2.922	5.483	3.495	14.604
1.5	RMSE	0.292	0.296	0.452	0.941	1.056
	Non-zero coef	7.764	7.752	7.828	12.460	22.586
	MSP∗100	3.764	3.787	6.408	6.255	14.827
2	RMSE	0.289	0.291	0.416	1.971	1.014
	Non-zero coef	7.584	7.566	7.680	10.566	22.576
	MSP∗100	4.148	4.188	6.678	7.124	14.474

DGP = data generating process, al = adaptive lasso, OLS = ordinary least squares, MLE = maximum likelihood estimation, RMSE = root mean square error, MSP = mean squared proportion.

가 크지 않았다.

DGP2에 대한 결과는 Table 3.6에 보고되었다. 여기에서는 튜닝 모수 γ에 대한 효과가 좀 더 극명하게 나타난다. 먼저 γ값이 증가할수록 가중치가 더 커지므로 좀 더 희박한 모형을 선택하지만 반면 RMSE는 증가하는 경향을 보인다. 하지만, 영이 아닌 계수의 개수에 비해서 RMSE의 변화는 그리 크지 않아 희박벡터자기상관 모형의 추정에서 adaptive lasso의 튜닝 모수 γ의 영향은 우려만큼 크지 않으며 대략 γ값이 0.5∼1.5 사이의 값이면 실증 자료 분석에서 충분히 좋은 결과를 제공할 것으로 보인다.

4. 결론

희박벡터자기회귀모형은 매우 큰 다차원의 시계열 벡터들 간의 선형 종속관계를 연구할 때 효율적인 변수 선택 방법으로 잘 알려진 모형이다. 본 논문에서는 희박자기회귀모형의 계수 추정방법으로서의 adaptive lasso 벌점화에 대해 알아보고, 기존에 계수 추정방법으로 알려진 lasso와의 비교를 통해 adaptive lasso를 이용한 희박자기회귀벡터 모형 추정 성능을 알아보았다. 그 결과 lasso를 이용한 희박자기회귀벡터 모형 추정에서의 단점인 영이 아닌 계수를 과대 추정한다는 점이 adaptive lasso를 이용하면 크게 보완됨을 모의실험을 통해 확인했다. 특히, 분산 공분산 행렬을 업데이트 하며 adaptive lasso를 사용하였을 때 가장 높은 성능을 보임을 모의실험을 통해 밝혔으며 이를 위한 초기 추정값으로는 릿지 추정량의 경우 가장 낮은 성능을 보였으며 최소자승추정값(al-OLS) 혹은 i.i.d. 가정하의 lasso 추정량(al-Lasso)가 표본 크기, 튜닝 모수 등에 대한 효과를 종합적으로 판단했을 때 가장 좋은 성능을 보였다. 또한, adaptive lasso의 튜닝 모수인 γ값이 증가할수록 영에 가까운 작은 계수들에 대해 가중치가 증가하므로 더 희박한 모형을 추정하나 γ값에 따라 매우 민감하게 변하지는 않아 대략 .5에서 1.5사이의 범위에서의 값의 경우 충분히 좋은 성능을 제공할 것이라 본다.

References

Arnold, A., Liu, Y., and Abe, N. (2008). Temporal causal modeling with graphical Granger methods, In *Proceedings of the 13th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*.

Cule, E., De Iorio, M. (2013). Ridge regression in prediction problems: automatic choice of the ridge parameter, *Genetic Epidemiology*, **37**, 704–714.

Identification of synaptic connections in neural ensembles by graphical models, *Journal of Neuroscience Methods*, **77**, 93–107.

Davis, R. A., Zang, P., and Zheng, T. (2015). Sparse vector autoregressive modeling, arXiv:1207.0520. *Econometrica*, **37**, 424–438.

Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC press.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models, *Statistica Sincia*, **18**, 1608–1618.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using lasso, *Computational Statistics & Data Analysis*, **52**, 3645–3657.

Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery, *Bioinformatics*, **25**, 110–118.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions, *arXiv:1106.3915*.

Sims, C. A. (1980). Macroeconomics and reality, *Econometrica: Journal of the Econometric Society*, 1–48.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Zhang, J., Jeng, X. J., and Liu, H. (2008). Some Two-Step Procedures for Variable Selection in High-Dimensional Linear Regression, arXiv:0810.1644.

Zou, H. (2006). Adaptive lasso and its oracle properties, *Journal of American Statistical Association*, **101**, 1418–1429.

Adaptive lasso를 이용한 희박벡터자기회귀모형에서의 변수 선택

이슬기[a] · 백창룡[a,1]

[a]성균관대학교 통계학과

(2015년 10월 27일 접수, 2015년 11월 26일 수정, 2015년 11월 30일 채택)

요 약

본 논문은 다차원의 시계열 자료 분석에서 효율적인 희박벡터자기회귀모형에서의 모수 추정에 대해서 연구한다. 희박벡터자기회귀모형은 영에 가까운 계수를 정확이 영으로 둠으로써 희박성을 확보한다. 따라서 변수 선택과 모수 추정을 한꺼번에 할 수 있는 lasso를 이용한 방법론을 희박벡터자기회귀모형의 추정에 쓸 수 있다. 하지만 Davis 등 (2015)에서는 모의실험을 통해 일반적인 lasso의 경우 영이아닌 계수를 참값보다 훨씬 더 많이 찾아 희박성에 약점이 있음을 보고하였다. 이에 따라 본 연구는 희박벡터자기회귀모형에 adaptive lasso를 이용하면 일반 lasso보다 희박성을 비롯한 전반적인 모수의 추정이 매우 유의하게 개선됨을 보인다. 또한 adaptive lasso에서 쓰이는 튜닝 모수들에 대한 선택도 아울러 논의한다.

주요용어: 희박벡터자기회귀모형, adaptive lasso, 고차원 시계열

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2014R1A1A1006025).

[1]교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: crbaek@skku.edu